

П. ДЖУРС , Т. АЙЗЕНАУЭР

Распознавание образов В ХИМИИ





CHEMICAL
APPLICATIONS
OF PATTERN
RECOGNITION

PETER C. JURIS

Associate Professor of Chemistry
The Pennsylvania State University

THOMAS L. ISENHOUR

Professor of Chemistry
University of North Carolina

A WILEY-INTERSCIENCE PUBLICATION

John Wiley & Sons

NEW YORK / LONDON / SYDNEY / TORONTO

П. ДЖУРС, Т. АЙЗЕНАУЭР

Распознавание образов В ХИМИИ

Перевод с английского
канд. хим. наук **С. В. КРИВЕНКО**

Под редакцией
доктора хим. наук проф. **А. М. ЕВСЕЕВА**
и канд. техн. наук **Г. Г. ВАЙНШТЕЙНА**

ИЗДАТЕЛЬСТВО «МИР»

МОСКВА

1977

Книга посвящена применению кибернетических методов классификации объектов при помощи ЭВМ для анализа данных химического эксперимента. Эта монография — первая по автоматизации обработки данных научных исследований в области химического анализа как в отечественной, так и в переводной литературе. Используются данные распространенных аналитических методов: масс-спектрометрии низкого разрешения, ИК-спектроскопии, спектроскопии ЯМР, полярографии.

Предназначена для широкого круга химиков — научных работников, специализирующихся в аналитической и органической химии, преподавателей и студентов.

Редакция литературы по химии

Copyright © 1975 by John Wiley & Sons, Inc.
All Rights Reserved.

Authorized translation from English language edition
published by John Wiley & Sons, Inc.

© Перевод на русский язык, «Мир», 1977

Д $\frac{20503-093}{041(01)-77}$ 93—77

ПРЕДИСЛОВИЕ К РУССКОМУ ИЗДАНИЮ

Методы распознавания образов приобрели важное значение в практике научной работы благодаря широкому применению электронно-вычислительной техники. Машинный алгоритм распознавания образов, по-видимому, может имитировать процесс распознавания образов мозгом человека. Однако мозг человека значительно уступает машине в способности запомнить огромное число мелких деталей. Только машина может запомнить большой ряд чисел, характеризующий изучаемый объект, упорядочить их, выявить совпадение чисел, соответствующих объектам одного класса, т. е. провести классификацию множества объектов.

В настоящее время ЭВМ является мощным инструментом научных исследований, которому доступна роль, выполнявшаяся ранее лишь человеком. ЭВМ, снабженная алгоритмом распознавания образов, обладает искусственным интеллектом. Машине можно предъявить ряд объектов, назвать их (т. е. отнести к соответствующим классам) и после такого обучения ЭВМ будет способна опознать новый, не предъявлявшийся ранее объект и отнести его к одному из известных ей классов. Такая «обучающаяся» машина может с успехом применяться в любой области научных исследований.

В химии искусственный интеллект развивается по многим направлениям. Среди них наиболее важные — применение методов распознавания образов в аналитических целях и для предсказания возможности синтеза конкретного соединения и его свойств. При помощи этих методов была предсказана возможность синтеза интерметаллических соединений и окислов с определенной структурой. В большинстве случаев наиболее успешные результаты были получены при применении систем признаков, в которых используются численные данные периодической таблицы элементов Д. И. Менделеева. Это обстоятельство указывает на то, что в статистических методах распознавания образов существенную роль должны играть уже установленные закономерности природы. Интеллектуальная деятельность человека при создании абстрактных теорий и классификации объектов, по-видимому, имеет единый характер. Поэтому только сочетание теории и статистического отбора различных схем

распознавания может обеспечить перспективное развитие искусственного интеллекта в химии.

В предлагаемой читателю книге Джурса и Айзенауэра физико-химическому обоснованию выбора системы признаков уделено относительно мало внимания. До некоторой степени это понятно, так как в основном речь идет о методах классификации и отображении объектов в пространстве образов.

Главное достоинство книги — очень простая форма изложения теории распознавания образов в применении к аналитическим задачам. В первую очередь рассмотрены такие проблемы, как масс-спектрометрический анализ органических веществ и установление брутто-формул и структурных формул соединений. Кроме того, обсуждены возможности анализа полярографических кривых и спектров ЯМР. Объем изложенного материала вполне достаточен для того, чтобы химик мог получить исчерпывающее представление о методах распознавания образов и смог работать в этой области. Конечно, при этом необходимо, чтобы химик владел искусством общения с ЭВМ хотя бы на уровне использования стандартных программ, а также был знаком с элементами регрессионного анализа и математической статистики.

Содержание книги достаточно полно охарактеризовано в предисловии авторов к американскому изданию. В дополнительной литературе, приведенной в конце книги, представлены работы в области распознавания образов, выполненные в Советском Союзе, в том числе посвященные химическим проблемам.

А. М. Евсеев

ПРЕДИСЛОВИЕ К АМЕРИКАНСКОМУ ИЗДАНИЮ

Настоящая книга представляет собой вводный курс, посвященный приложениям некоторых методов распознавания образов к решению химических задач. Распознавание образов охватывает чрезвычайно широкую область разнообразных методов и их применений, однако мы сознательно сузили круг рассматриваемых вопросов. Значительная часть книги посвящена обсуждению таких непараметрических распознающих систем, которые называются «обучающимися машинами». Свое название эти машины получили от способности обучаться: давать с накоплением опыта все более правильные ответы на вопросы, относящиеся к классификации. Основным блоком подобных систем служит приспособляющийся бинарный классификатор образов, или адаптивный пороговый логический элемент. Такие устройства обследуют множество помеченных данных на инвариантность, которую можно использовать при классификации. Эта процедура известна как обучение с учителем. Большая часть рассмотренных в книге работ связана с использованием адаптивных бинарных классификаторов образов, которые при обучении с учителем вырабатывают способность к классификации.

В основном обсуждаются данные спектроскопических исследований, такие, как масс-спектры низкого разрешения, ИК- и ЯМР-спектры. Как раз в этих областях химии методы распознавания образов нашли свое применение. В последнее время эти методы стали использоваться в других областях и, по-видимому, именно возможность анализа разнообразных химических задач представляет основу их дальнейшего плодотворного применения.

Книга состоит из семи глав. В первой изложены общие основы распознающих систем с краткой характеристикой их составных частей. Во второй даны основные математические принципы, используемые в бинарных классификаторах образов, описывается способ обучения при построении последних и рассматриваются критерии оценки их полезности. Третья глава посвящена некоторым методам предварительной обработки данных, позволяющим экономно осуществлять классификацию, с примерами практиче-

ского анализа спектроскопических и электрохимических данных. В четвертой рассмотрены построение разделяющей функции и применяемые при этом процедуры; в этой главе приведены многочисленные примеры из практики химических исследований. В пятой главе описан отбор признаков, который проводится по результатам, показываемым классификатором. Изложение подкрепляется примерами обработки масс-спектрометрических данных. В шестой главе обсуждены сложные преобразования при доклассификационной обработке данных, в том числе генерирование членов, учитывающих взаимодействие между первичными дескрипторами, преобразование Фурье, а также факторный анализ. В седьмой рассмотрена возможность использования теории распознавания образов для решения химических задач другого типа, таких, как предсказание масс-спектра соединения на основе только его двумерной молекулярной структуры. Эти методы применяются в химии при решении самых различных задач, например при анализе лекарственных препаратов. В приложении помещен полный текст программы для моделирования обучающейся машины с показательным набором пятимерных данных на выходе машины. Программа работала на вычислительной машине IBM 370/168, принадлежащей вычислительному центру при Пенсильванском университете, однако она почти без всяких переделок может использоваться и в других вычислительных системах. В конце книги приведен список монографий, позволяющий заинтересованному читателю составить представление о богатой литературе по распознаванию образов.

Авторы выражают благодарность студентам и коллегам, которые помогли им при написании настоящей книги и выполнили многие из рассмотренных в ней экспериментальных исследований.

П. Джурс

Т. Айзенаур

Колледж штата Пенсильвания

Чапел-Хилл, Северная Каролина

Февраль 1975 г.

Создание быстродействующих цифровых вычислительных машин в качестве мощных устройств по обработке самой разнообразной информации привело к коренной перестройке многих отраслей науки и техники. Благодаря огромным достижениям в данной области стало возможным решать задачи, считавшиеся ранее крайне сложными и даже практически неразрешимыми. Многие из таких задач можно, как выяснилось, решить только методами «распознавания образов». Именно такие методы позволили разработать представления и способы, использующиеся повседневно при машинном решении задач, которые прежде считались доступными только силе человеческого разума. В настоящее время распознавание образов стало одной из областей «искусственного интеллекта», или «машинного разума».

Распознавание образов включает обнаружение, восприятие и распознавание закономерностей (инвариантных свойств) среди серий результатов измерений, характеризующих объекты или события. Как правило, задача распознавания сводится к отнесению той или иной выборки экспериментальных результатов к надлежащему классу. Такой общий подход нашел применение при решении многих проблем в самых разнообразных областях.

ОБЛАСТИ ПРИМЕНЕНИЯ РАСПОЗНАВАНИЯ ОБРАЗОВ

Методами распознавания образов удалось решить удивительно много практических задач. Составлено несколько хороших обзоров литературы по применению принципов распознавания образов, иллюстрирующих многообразие решаемых проблем [1—5]. Этим вопросам посвящен также целый ряд монографий (см. список литературы в конце книги).

Интенсивное развитие исследований по теории и практике рас-

познавания образов объясняется в той или иной мере следующими причинами:

1. Стремлением заменить людей машинами при выполнении рутинных операций по обработке информации. Машины, способные обрабатывать информацию быстрее, точнее, надежнее и дешевле, оказываются предпочтительнее.

2. Максимальная эффективность взаимодействия человека с машиной достижима лишь на пути создания устройств, использующих принципы распознавания образов. Такие устройства должны понимать зрительные образы и естественный язык, поскольку люди предпочитают именно эти средства общения.

3. Исследования по распознаванию образов имеют и самостоятельную ценность независимо от их практической направленности. Главная задача исследований здесь заключается в следующем: понять в общем, что нужно (например, в виде конструкции некоторого устройства), чтобы получить при восприятии такую же реакцию, как у человека.

Следует отметить, что нередко исследования по распознаванию образов оказываются успешными сразу во всех трех перечисленных направлениях.

Типы данных, анализировавшихся методами распознавания образов, подразделяются на следующие две категории: 1) точно кодируемые структуры; 2) изображения. Примерами точно кодируемых структур служат печатный и рукописный тексты, отпечатки пальцев и иные данные, которые (в идеальном случае) поддаются двоичному кодированию. Изображения охватывают фотографии и иные зрительные данные, создаваемые путем различных уровней почернения. Изучались и такие типы данных, которые трудно отнести только к одной из этих двух категорий. Ниже речь пойдет о некоторых конкретных проблемах, исследовавшихся методами распознавания образов.

Понятие «распознавание знаков» относится к интерпретации печатных и рукописных знаков [6]. Наиболее разработанное направление в этой области — создание автоматов для чтения специальных символов, например магнитных цифр на банковских чеках, получивших широкое распространение. Усиленно изучается также более трудная проблема чтения символов текста с пишущих машинок и других устройств «ударной» печати. Для таких устройств были получены стандартные образцы текстов, что привело к расширению исследований методами распознавания образов. Предпринимались даже попытки решить еще более трудную задачу чтения печатных букв и цифр, написанных от руки. Автоматизация служ-

бы почтовой связи является дополнительным стимулом для поиска эффективных средств взаимодействия человека с машиной и для создания автоматов, читающих рукописные тексты.

Другая область применения распознающих устройств для связи человека с машиной — это распознавание речи. В этой области имеются две проблемы: 1) собственно распознавание сказанного и 2) идентификация диктора. Едва ли стоит напоминать о значении автоматов, способных выносить подобные решения.

Обработка фотографий представляет собой область, в которой возможно широкое применение методов распознавания образов. Именно по этой причине автоматизация процессов обработки фотографических изображений привлекла столь большое внимание исследователей. Методами распознавания образов удалось интерпретировать микрофотографии таких биологических объектов, как кровяные клетки и хромосомы [7]. Изучались также возможности обработки аэрофотоснимков для военных целей и дистанционного получения данных. Аэрофотосъемка позволяет узнавать культуры посевов на полях, выявлять очаги лесных пожаров, районы засухи и другие особенности сфотографированной местности [8]. Методами распознавания образов анализируют процессы в пузырьковых, искровых и паровых камерах (камерах Вильсона) [9], опознают личность по отпечаткам пальцев [10].

Распознающие системы использовались для разведки полезных ископаемых путем обработки геологических и сейсмических данных. Удалось разработать также приемы диагностики в медицине, позволившие давать заключения по электрокардиограммам (ЭКГ) и электроэнцефалограммам (ЭЭГ), проводить анализ крови [11, 12]. Изучались возможности составления прогнозов погоды по результатам измерений силы ветра, атмосферного давления, влажности и температуры воздуха.

ОБЩАЯ СХЕМА РАСПОЗНАЮЩЕЙ СИСТЕМЫ

Система распознавания образов должна в общем случае «обследовать» эталонную выборку данных, осуществлять предварительную обработку и необходимые преобразования последних и затем правильно классифицировать образ. Общая схема распознающей системы приведена на рис. 1.1. Она состоит из трех взаимосвязанных блоков: преобразователя, препроцессора (устройства для выделения признаков) и классификатора. Хотя при любой реализации распознающей системы эти три блока в сильной степени взаимозависимы, их полезно рассматривать отдельно.

Преобразователь переводит информацию, поступающую из реального мира, в пространство образов* распознающей системы. Поскольку распознающие системы обычно реализуются в виде машинных программ, преобразователь превращает исходные данные

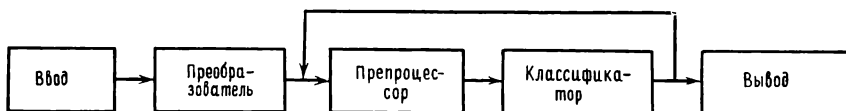


Рис. 1.1. Общая схема распознающей системы.

в форму, пригодную для ввода в вычислительную машину. Обычно такая форма представляет совокупность скалярных результатов измерения (замеров), образующих n -мерный вектор образа: $\mathbf{X} = x_1, x_2, \dots, x_n$. Каждая компонента вектора образа выражает физически измеримую величину.

Задача препроцессора состоит в том, чтобы подготовить поступивший вектор образа к классификации. Затем преобразованные данные подаются на вход классификатора, который их анализирует и выносит классифицирующее решение. О том, как это делается, подробнее говорится ниже.

Преобразователь

Первичные данные от подлежащих классификаций объектов подаются на ввод преобразователя, который преобразует всякий образ, содержащийся в исходных данных, в n -мерный вектор n -мерного евклидова пространства, называемого пространством образов. Совокупность замеров, формирующих вектор образа**, должна содержать существенную часть исходных данных, подаваемых на ввод. Таким образом, фактическая реализация преобразователя полностью определяется природой исходных данных. Когда они образуют временную последовательность сигналов, например ин-

* В литературе употребляются также термины: пространство изображений, пространство объектов. — *Прим. ред.*

** В переводе употреблено выражение «вектор образа» для обозначения одного из членов класса, общим для которого является данный образ. Образ — это обобщенное название (имя) класса объектов или явлений. Образом являются понятия «человек», «дерево», «книга данной библиотеки» и т. д. Вектор может быть сопоставлен одному из объектов множества, соответствующего данному образу. — *Прим. ред.*

терферограммы и электроэнцефалограммы, нужно предусмотреть процедуру взятия дискретных отсчетов по времени. Если же исходные данные представляют функцию частоты, например инфракрасные спектры, то уместно предусмотреть взятие отсчетов по частоте. Если, наконец, исходные данные имеют форму изображений, то поле подлежит исследованию на наличие темных и светлых участков, краев или геометрических конфигураций. С некоторыми способами преобразования изображений в совокупность отсчетов можно познакомиться в обзорной статье Левина [5]. Когда исходные данные вводятся в естественной цифровой форме (например, масс-спектры низкого разрешения), блок преобразования можно исключить. Векторы образов (n -мерные точки), формируемые преобразователем, затем поступают в блок выделения признаков (препроцессор).

Препроцессор (устройство для выделения признаков)

В блок выделения признаков (препроцессор) вводятся формируемые преобразователем векторы образов, которые здесь дополнительно преобразуются для того, чтобы:

1. Устранить или хотя бы уменьшить ту часть содержащейся в исходных данных информации, которая не имеет отношения к решаемой задаче и мешает распознаванию. Так, данные в виде изображений не должны изменяться при переносах, изменении масштаба или повороте; волновые данные должны быть независимыми от сдвигов во времени или по фазе.

2. Сохранить достаточный объем информации, позволяющий отличить один класс образов от другого, т. е. выявить инварианты в образах единого класса.

3. Сохранить информацию в векторе образа в таком виде, который обеспечивает эффективное действие линейного классификатора, если это возможно.

Ниже изложены некоторые особенности применявшихся способов выделения признаков и конструирования препроцессоров.

Простейшая предварительная обработка информации сводится к нормированию и другому масштабированию компонент вектора образа, например к приравниванию суммы этих компонент (или их квадратов) произвольной подходящей константе. Другой способ основан на предположении, что вес того или иного признака (координаты) обратно пропорционален дисперсии этого признака в пространстве образов. Более сложный и тонкий метод — составление матричного уравнения при помощи ковариационной матрицы. Ре-

шая это уравнение, находят собственные векторы и собственные значения, что дает в результате совокупность ортогональных координат, определяющих новое пространство образов, которое получается поворотом прежнего пространства и в котором задача классификации должна упроститься. Этим способом, известным под названием анализа главных компонент (метода главных факторов), или преобразования Карунена—Лоэва, можно также уменьшить размерность векторов образов. При этом оставляют только те из новых преобразованных координат, которые имеют большие собственные значения, а остальные — отбрасывают. Мы ограничимся приведенными примерами линейных преобразований, применявшихся в распознающих системах.

В распознающих системах использовались и некоторые более сложные преобразования. Векторы образов можно, например, подвергать преобразованию Фурье и затем вычислять спектр мощности. Разработаны итерационные способы, предусматривающие итерационную минимизацию критерия ошибки, например разности расстояний между всеми парами точек в исходном пространстве образов и новом, преобразованном пространстве меньшей размерности. Векторы образов можно представлять также в виде полиномиальных разложений.

Для обнаружения важных признаков подлежащие классификации объекты сравнивались с эталонами и прототипами. Такое сравнение проводилось интерактивными методами, иногда с использованием средств графического отображения. Предпринимались попытки определять, например, статистические параметры, такие, как моменты и гистограммы, путем непосредственной обработки образов. Из сказанного выше очевидно, что существует множество алгоритмов выделения признаков в процессе предварительной обработки информации; их число непрерывно и быстро растет, поскольку выбор способов решения конкретной задачи в большой степени обусловлен характером самой задачи. Успех всего исследования по проблеме распознавания образов определяется тем, насколько удачно выполнен этап выделения признаков. Общее признание получила точка зрения, согласно которой новых крупных достижений в этой области следует ожидать как раз на стадии выделения признаков при предварительной обработке информации.

Классификатор

Преобразованные объекты классифицируются третьим блоком распознающей системы. Классификаторы систем распознавания образов создавались на основе различных отраслей прикладной математики: теории статистических решений, теории информации, геометрии и т. п.

Назначение классификатора можно сформулировать в общем виде следующим образом. Решающую функцию $f(\mathbf{X})$ (решающее правило)* определяют при помощи набора преобразованных векторов образов, называемого обучающей выборкой**, так, чтобы эта функция удовлетворяла условиям:

$$f(\mathbf{X}) > 0 \quad \text{для } \mathbf{X} \text{ членов класса 1;}$$

$$f(\mathbf{X}) \leq 0 \quad \text{для } \mathbf{X} \text{ членов класса 2.}$$

Процедуру определения решающей функции $f(\mathbf{X})$ принято называть адаптацией, тренировкой или обучением. Задача сводится к тому, чтобы минимизировать вероятность ошибки.

Определение решающей функции $f(\mathbf{X})$ осуществляют параметрическими методами или непараметрическими методами. Начальной операцией параметрических методов обучения служит оценка статистических параметров образов, составляющих обучающую выборку. Затем такие оценки используют при детализации разделяющих функций. Самой распространенной параметрической разделяющей функцией является правило Байеса, поскольку оно выражает оптимальное решение для задач точно определенного класса.

Чтобы воспользоваться правилом Байеса при решении задачи разбиения объектов на два класса, необходимо знать аналитический вид и параметры функций условной плотности вероятностей для обоих классов. Обычно исходят из предположения о том, что образам присуще нормальное распределение относительно их среднего по классу. Следовательно, нужно знать средние значения векторов и ковариационные матрицы обоих классов, а также конкретизировать функции потерь, учитывающие эффект ошибочной классифи-

*Решающая функция, решающее правило, дискриминантная функция, классифицирующая функция — допустимые синонимы. — *Прим. ред.*

** Допустимы синонимы «тренировочная совокупность», «обучающая совокупность (последовательность)». — *Прим. ред.*

кации. Тогда разделяющую функцию можно записать в виде

$$f(\mathbf{X}) = \frac{P_1 L_1 F_1(\mathbf{X})}{P_2 L_2 F_2(\mathbf{X})} - 1,$$

где P_1 — априорная вероятность принадлежности того или иного образа к классу 1; $P_2 = 1 - P_1$, т. е. априорная вероятность принадлежности этого образа к классу 2; L_1 и L_2 — потери, связанные с ошибочным отношением того или иного образа класса 1 к классу 2, и наоборот; $F_1(\mathbf{X})$ и $F_2(\mathbf{X})$ — плотности распределения вероятностей образов в классах 1 и 2.

Если предположить, что образы имеют гауссово распределение и что средние значения векторов и ковариационные матрицы точно характеризуют свои классы, то байесова разделяющая функция является оптимальной*. По этой причине правило Байеса используют как прототип при сравнении с другими разделяющими функциями.

Когда же образы обучающей выборки не поддаются статистическому описанию, следует применять непараметрическую разделяющую функцию. При выводе непараметрических разделяющих функций приходится использовать единственные имеющиеся в наличии данные — объекты самой обучающей выборки. Такой способ обучения может обеспечить надежность результатов распознавания только тогда, когда обучающая выборка содержит достаточно много объектов, чтобы ее можно считать представительной для той совокупности данных, из объектов которой она составлена. (Объем обучающей выборки представляет необходимое, но не достаточное условие возможности вывода параметрических разделяющих функций путем оценки функций вероятности.)

Из непараметрических бинарных классификаторов широко исследован пороговый логический элемент. Кроме векторов образов обучающей выборки подстраиваемыми параметрами логических элементов являются только их линейные коэффициенты, определяемые в процессе обучения. Адаптивный пороговый логический элемент снабжается устройством для регистрации показаний по отношению к точно известному воздействию (объекту), т. е. способностью изменения своих параметров для того, чтобы реакция была верной. Обычно пороговые логические элементы можно подстраивать только на стадии конструирования. Подробно характеристики пороговых логических элементов рассматриваются в гл. 2. Пороговые логические элементы, соединенные во взаимосвязанные цепи, используют

* То есть обеспечивающей минимальную вероятность ошибки. — *Прим. ред.*

ся в усовершенствованных классификаторах, известных как ку-сочно линейные и послойные.

Еще одним распространенным непараметрическим методом различения является классификация по K ближайшим соседям. При классификации неизвестный объект (изображение) считают принадлежащим к классу, чаще других встречающемуся среди его ближайших соседей. Соседство обычно определяют через расстояние в евклидовом пространстве, хотя это можно сделать и в произвольной метрике. К недостаткам классификации по K ближайшим соседям относятся необходимость запоминания всех векторов образа и большой объем вычислений. Но, как было установлено, вероятность ошибок при классификации по одному ближайшему соседу самое большое в два раза превышает вероятность ошибок, даваемую оптимальным байесовым классификатором, когда известны все исходные вероятности [13]. В этом алгоритме привлекает еще простота его фундаментальных основ и связанных с ним вычислений. Как показали исследования последнего времени, тщательный подбор образов обучающей выборки при классификации этим способом позволяет снизить требования к памяти и сократить объем необходимых выкладок (см., например, [14]).

Если подлежащие классификации данные не поддаются статистическому описанию, а классы векторов отдельных образов неизвестны, то приходится обращаться к классификаторам другого типа. Они называются кластерными, поскольку построены на том, что векторы образов сами должны определить число точечных сгущений (кластеров), поддающихся привязке к классам. В подобных распознающих устройствах речь идет о так называемом «обучении без учителя», поскольку классы образов обучающей выборки неизвестны. Трудную задачу отыскания действенного подхода к решению подобных проблем пробовали преодолевать составлением гистограмм данных с тем, чтобы аппроксимировать функции плотности вероятностей классов и затем строить разделяющую функцию другими методами.

К непараметрическим методам распознавания относится и метод потенциальных функций. В этом случае каждая из известных точек в пространстве образов считается создающей свое собственное потенциальное поле. Чтобы отнести неизвестный образ к одному из двух возможных классов, составляют оценки их полных потенциальных полей и считают этот образ принадлежащим к классу с более сильным полным полем. Этот алгоритм дает исследователям весьма мощное орудие, поскольку форму потенциальных функций, а равно и контролирующие параметры можно широко варьировать

в зависимости от принадлежности к классу, области в пространстве изображений, вероятности наличия разных классов и т. п.

Вот отдельные примеры многочисленных классификаторов, созданных к настоящему времени. Дополнительные сведения о классификаторах можно почерпнуть из приведенных в конце настоящей книги изданий и монографий, а также из хорошей обзорной статьи [15].

СПИСОК ЛИТЕРАТУРЫ

1. *Minsky M.*, Proc. IRE, **49**, 8 (1961); есть русский перевод: *Минский М.*, ТИРИ, **49**, № 1 (1961).
2. *Solomonoff R. J.*, Proc. IEEE, **54**, 1687 (1966).
3. *Rosen C. A.*, Science, **156**, 38 (1967).
4. *Nagy G.*, Proc. IEEE, **56**, 836 (1968); есть русский перевод: *Надь Г.*, ТИИЭР, **56**, № 5, 57 (1968).
5. *Levine M. D.*, Proc. IEEE, **57**, 1391 (1969).
6. *Harmon L. D.*, Proc. IEEE, **60**, 1165 (1972).
7. *Preston K., Jr.*, Proc. IEEE, **60**, 1216 (1972).
8. *Nagy G.*, Proc. IEEE, **60**, 1177 (1972).
9. *Strand R. C.*, Proc. IEEE, **60**, 1122 (1972).
10. *Kanal L. N.*, Proc. IEEE, **60**, 1200 (1972).
11. *Lox J. R., Jr., Nolle F. M., Arthur R. M.*, Proc. IEEE, **60**, 1137 (1972).
12. *Patrick E. A., Stelmack F. P., Shen L. Y. L.*, IEEE Trans., **SMC-4**, 1 (1974).
13. *Cover T. M., Hart P. E.*, IEEE Trans., **IT-13**, 21 (1967).
14. *Gates G. W.*, IEEE Trans., **IT-18**, 431 (1972).
15. *Ho Y.-C., Agrawala A. K.*, Proc. IEEE, **56**, 2101 (1968).

ВВЕДЕНИЕ В ТЕОРИЮ БИНАРНЫХ КЛАССИФИКАТОРОВ ОБРАЗОВ

ВЕКТОРЫ ОБРАЗОВ В ГИПЕРПРОСТРАНСТВЕ

Точками в евклидовом пространстве подходящей размерности можно охарактеризовать самые разнообразные данные. Например, такие свойства, как положение в пространстве или количество движения, можно описать точкой в трехмерном пространстве или — что эквивалентно — трехмерным вектором. Для этого следует выбрать три линейно-независимые оси, при помощи которых можно полностью охарактеризовать интересующее нас свойство в трехмерном пространстве, если заданы координаты (замеры). Этот способ описания легко распространить в случае необходимости и на пространство большей размерности. Так, в механике положение и количество движения частицы принято описывать в шестикординатном пространстве, часто называемом фазовым пространством. Таким образом, любая точка в шестимерном пространстве характеризует и положение, и количество движения частицы. При подобном подходе картина описывается вектором, исходящим из начала системы координат и оканчивающимся в рассматриваемой точке.

Вектором d -мерной размерности

$$\mathbf{X} = x_1, x_2, \dots, x_d \quad (2.1)$$

(где отдельные компоненты x_j — замеры) можно описать разнообразные химические данные. Например, чтобы представить масс-спектр низкого разрешения в виде вектора, можно принять x_j равным интенсивности пика в положении j , выраженном в единицах m/e . В случае табличных данных замером, или координатой, можно считать величины, приведенные в каждом столбце таблицы. Так, таблицу данных о температурах плавления и кипения, атомных весах, показателях преломления и плотностях легко преобразовать в совокупность векторов в пятимерном пространстве, каждый из которых характеризует какое-то определенное соединение.

Преобразование в векторную форму графических данных, носящих на первый взгляд непрерывный характер, — более сложная задача. Пока эти данные имеют дискретный характер, т. е. пока

их можно представить в виде гистограмм, способ привязки к ним размерностей очевиден. Например, в случае масс-спектрометра низкого разрешения, скажем с разрешающей способностью 1 единица массы (е.м.) и разверткой 1—200 е.м., при помощи 200-мерного вектора можно охарактеризовать масс-спектр любого соединения. Многие приборы, однако, дают аналоговые данные непрерывного характера. Приборам, запоминая таким данные, например сканирующим спектрометрам, присуще какое-то предельное разрешение. Обычно электронное и оптическое устройства подобных приборов таковы, что они интегрируют на интервале, чем и определяется их предельная разрешающая способность. В этих случаях проще всего определить разрешение r и затем разбить спектр на R/r компонент или координат, где R — интервал сканирования. Так, если $r=0,1$, то инфракрасный спектр в диапазоне длин волн от 2,0 до 15,0 мкм можно преобразовать в 130 дискретных элементов.

Важно понимать, что гиперпространственное отображение данных — это всего лишь одна из возможностей их описания. В общем случае фактические данные не поддаются корреляции с геометрическими свойствами. Например, даже если ортогональные координаты вектора независимы, *это вовсе не означает, что отдельные части масс-спектра, представленные таким способом, всегда независимы.*

В силу ортогональности компонент вектора некоторые операции над отдельными координатами можно осуществлять обратимым образом. Следовательно, оператор, преобразующий каждую координату, можно использовать без ущерба для исходных данных. Это верно в тех случаях, когда возможна обратная операция, позволяющая воспроизвести исходный вектор. Например, различные нормировочные процедуры, такие, как извлечение квадратного корня или нахождение логарифма каждой компоненты вектора, представляют преобразования, при которых исходный вектор можно восстановить в первоначальном виде, правда с изменением динамического диапазона. Это свойство играет важную роль при некоторых операциях обучения распознаванию.

РЕШАЮЩИЕ ПОВЕРХНОСТИ

Одной из первоочередных задач распознавания образов является правильное разбиение данных на категории или классы. Образом можно считать любую выборку данных, характеризующих то или иное свойство, процесс или совокупность. Например, масс-

спектр какого-либо химического соединения представляет собой образ, описывающий сложные химические и физические процессы. Кроме того, этот образ характеризуется определенными основными свойствами соединения и характеристиками самого процесса измерения. Следовательно, методы распознавания образов можно использовать для решения таких задач, как классификация масс-спектров по типам химических соединений, например кислород-содержащих соединений в отличие от бескислородных.

Представив совокупность масс-спектров в виде точек гиперпространства, мы получим совокупность точек, отображающих образы, которые содержат всю информацию, заключавшуюся в исходных спектрах. Теперь задача сводится к тому, чтобы эту совокупность точек разбить на две подсовокупности, определяемые распознаваемыми классами; иными словами, речь идет о том, как преобразовать пространство образов в классифицирующее пространство.

Один из способов классификации точек образов заключается в нахождении групп, принадлежащих одному и тому же классу, и размещению между ними решающей поверхности (границы). В простом двумерном пространстве такая классификация сводится к проведению линий (не обязательно прямых) между точками разных классов (например, коров можно отделить от лошадей, отгородив их друг от друга забором).

В гиперпространстве точки образов, соответствующие соединениям с родственными характеристиками, должны, по-видимому, образовывать кластеры. Например, точки образов, представляющие масс-спектры спиртов, должны группироваться в какой-то ограниченной области гиперпространства, а точки образов, относящиеся к масс-спектрам алкенов, — в другой области этого пространства. Часто подобное предположение оказывается верным по отношению к совокупностям точек, отображающих химические данные, например масс-спектры. Если образуются кластеры, решающие поверхности удается располагать между ними. В простейшем таком случае решающая поверхность представляет гиперплоскость той же размерности, что и выбранное гиперпространство.

Такая гиперплоскость может и не быть линейной или «плоской» в соответствии с размерностью пространства, однако, когда решающая поверхность линейна и проходит через начало координат, математически задача значительно упрощается. В этом случае гиперплоскость можно охарактеризовать вектором нормали, исходящим из начала координат. Иными словами, формально всякий вектор, исходящий из начала координат, определяет плоскость,

которая является геометрическим местом точек, лежащих на перпендикулярах к этому вектору.

Поскольку очень удобно, когда решающая плоскость проходит через начало координат в пространстве образов, стоит принять меры, чтобы при этом не возникало ущерба для способности к классификации (разделению). Для этого пространство образов расширяют за счет введения еще одной ортогональной координаты, приписывая всем векторам образов $(d+1)$ -ю компоненту, которая может иметь любую величину, но ее принято выбирать равной единице.

Итак, задав вектор нормали к плоской решающей поверхности, мы определяем последнюю. Однако в подобной простой ситуации возникает еще одна важная особенность. Дело в том, что скалярное произведение вектора нормали **W** на вектор образа **X** определяет, по какую сторону от гиперплоскости лежит точка, характеризующая данный образ:

$$\mathbf{W} \cdot \mathbf{X} = |\mathbf{W}| |\mathbf{X}| \cos \theta, \quad (2.2)$$

где θ — угол между двумя векторами, причем

$$\begin{aligned} \cos \theta > 0, \quad \text{если} \quad -90^\circ < \theta < 90^\circ, \\ \cos \theta < 0, \quad \text{если} \quad 90^\circ < \theta < 270^\circ. \end{aligned} \quad (2.3)$$

Вектор нормали перпендикулярен гиперплоскости, поэтому все образы с положительными скалярными произведениями лежат по ту же ее сторону, что и вектор нормали, а образы с отрицательными произведениями — по другую ее сторону. (Точки с нулевыми скалярными произведениями находятся на гиперплоскости, что также можно использовать для определения ее положения.)

Хотя отнюдь не обязательно, чтобы решающие поверхности были линейными (плоскими), их линейность намного упрощает задачу. К тому же, как можно показать, более сложные решающие поверхности можно заменить линейными, если исходные данные должным образом преобразовать в препроцессоре.

Пороговые логические элементы как бинарные классификаторы образов

Название «пороговые логические элементы» связано с тем, что эти системы подобны логическим блок-схемам, находящимся в одном из двух возможных состояний в зависимости от того, выше или ниже входной сигнал определенного уровня — порога.

Когда необходимо принять бинарное решение, т. е. когда образы нужно разбить на две категории, в основном используют пороговые логические элементы. В общем случае нужна функция, дающая один из двух результатов в зависимости от величины входного сигнала.

Хотя наибольшее распространение получили линейные пороговые логические элементы, в качестве бинарных классификаторов образов могут применяться пороговые логические элементы с любой другой функциональной зависимостью от ответа. Необходимо только, чтобы они надежно отличали образы одного интересующего нас класса от образов другого.

Простой биметаллический термостат в бытовом нагревательном приборе представляет хороший пример порогового логического элемента. Комнатная температура используется как входной сигнал, преобразуемый в изгиб биметаллической пластины. До тех пор пока сигнал превосходит определенный пороговый уровень, который можно выразить в градусах, термостат не генерирует напряжение (нулевой сигнал), так что реле питания нагревателя остается в выключенном положении. Когда же температура уменьшается до уровня ниже порогового, термостат начинает генерировать напряжение, включающее печь нагревателя.

Если обратиться снова к понятию гиперплоскости, то для классификации векторов образов пороговый логический элемент можно охарактеризовать как алгоритм, дающий два разных состояния для любого входного вектора. Хотя и не обязательно, но математически удобно выбирать в качестве порога нуль. Именно так чаще всего и поступают. В качестве пороговых логических элементов удобно использовать упоминавшуюся выше линейную разделяющую функцию. Разделяющую гиперплоскость можно, как уже говорилось, охарактеризовать вектором нормали \mathbf{W} . Скалярное произведение этого вектора на вектор образа имеет положительную величину для образов с той стороны от плоскости, на которой расположен вектор нормали, и отрицательную для образов по другую сторону от нее. Следовательно, когда за порог выбран нуль, линейная разделяющая функция определяет классифицирующее пространство (пространство решений) с двумя группами точек, разделенными между собой гиперплоскостью, которая нормальна разделяющему вектору.

Отметим, что требование выбора только одного из двух результатов не накладывает никакого реального ограничения на процесс классификации, поскольку всегда можно довести классификацию до нужной степени, используя последовательно один за другим

несколько пороговых логических элементов. (Примеры приведены ниже.) Например, торгующие автоматы принимают ряд бинарных решений в зависимости от размера монет.

Еще одна интересная возможность использования линейных разделяющих функций в качестве пороговых логических элементов связана с понятием «вектор». Скалярное произведение двух векторов можно эквивалентно определить соотношением

$$\mathbf{W} \cdot \mathbf{X} = |\mathbf{W}| |\mathbf{X}| \cos \theta = w_1 x_1 + w_2 x_2 + \dots + w_d x_d + w_{d+1}, \quad (2.4)$$

означающим, что каждой компоненте вектора \mathbf{X} приводится в соответствие весовая компонента вектора \mathbf{W} . Эта процедура приписывания компонентам вектора \mathbf{X} весовых коэффициентов в пороговом логическом элементе позволяет осуществлять то или иное разделение, для чего весовые коэффициенты задаются с таким расчетом, чтобы соответствующее скалярное произведение было больше или меньше пороговой величины.

ОБУЧЕНИЕ ПОРОГОВЫХ ЛОГИЧЕСКИХ ЭЛЕМЕНТОВ С ИСПРАВЛЕНИЕМ ОШИБКИ ЧЕРЕЗ ОБРАТНУЮ СВЯЗЬ

Как уже отмечалось, пороговый логический элемент можно использовать для разбиения на два класса совокупности данных, представленных в виде точек или векторов в гиперпространстве. Следовательно, задача сводится к отысканию эффективного разделителя, осуществляющего дихотомию для заданного множества классификаций. Именно это и имелось в виду выше, когда речь шла о преобразовании пространства образов в классифицирующее пространство (пространство решений).

Во многих случаях эффективное разбиение данных на два класса удается произвести при помощи линейной разделяющей функции. Однако отыскать подобную функцию для многомерных данных далеко не просто. Невозможно построить многомерный граф, наглядно изображающий точки. Более того, как правило, нецелесообразно вычислять все возможные решающие поверхности для того, чтобы выбрать одну из них, ведущую к нужному решению. Во многих случаях успех приносит эвристический подход.

Такой эвристический подход базируется на выборе исходной классифицирующей, или разделяющей, функции либо произвольным образом, либо при помощи той или иной аппроксимирующей процедуры с последующим «обучением» классификатора путем преобразования этой функции по мере накопления классификатором

опыта по принятию решений. Это обучение проводится на выборке образов с известной классификацией (обучающей выборке). Образы «показывают» обучаемому классификатору по одному. Если классификация произведена неверно, для исправления ошибки решающую поверхность варьируют.

Существует ряд приемов обучения с исправлением ошибки через обратную связь. Один способ, обеспечивающий, как можно показать, сходимость для любой линейно разделимой совокупности данных, исправляет решающую плоскость, заставляя ее как бы отразиться от неправильно классифицированной точки. Как уже говорилось, алгебраический знак скалярного произведения весового вектора на вектор образа указывает, с какой стороны от решающей поверхности расположена точка, представляющая образ:

$$\mathbf{W} \cdot \mathbf{X} = s. \quad (2.5)$$

(Предварительно нужно договориться, какое из двух подмножеств называть положительным, а какое — отрицательным.) Когда образ i из обучающей выборки отнесен не к тому классу, к которому он принадлежит, произведение

$$\mathbf{W} \cdot \mathbf{X}_i = s \quad (2.6)$$

имеет не тот знак, который требуется для правильной классификации \mathbf{X}_i . Теперь задача сводится к тому, чтобы вычислить исправленный весовой вектор \mathbf{W}' , произведение которого на вектор образа

$$\mathbf{W}' \cdot \mathbf{X}_i = s' \quad (2.7)$$

приобрело бы противоположный знак по сравнению с предыдущим случаем. Новый весовой вектор берут в виде суммы исходного вектора \mathbf{W} и некоторой доли вектора \mathbf{X}_i :

$$\mathbf{W}' = \mathbf{W} + c\mathbf{X}_i. \quad (2.8)$$

Объединив соотношения (2.7) и (2.8), получаем уравнение

$$s' = \mathbf{W}' \cdot \mathbf{X}_i = (\mathbf{W} + c\mathbf{X}_i) \cdot \mathbf{X}_i, \quad (2.9)$$

которое после алгебраических преобразований принимает вид

$$c = \frac{s' - s}{\mathbf{X}_i \cdot \mathbf{X}_i}. \quad (2.10)$$

Остается только выбрать подходящее значение s' , чтобы завершить вывод. Хорошие результаты дает следующий выбор: $s' = -s$. Тогда решающая поверхность перемещается так, что после внесения

поправки через обратную связь точка \mathbf{X}_i оказывается на том же расстоянии от решающей поверхности, но с «правильной» стороны, на каком она находилась от нее до этого с «неправильной» стороны. Если в уравнение (2.10) подставить $s' = -s$, то получим новое уравнение:

$$c = \frac{-2s}{\mathbf{x}_i \cdot \mathbf{x}_i}. \quad (2.11)$$

Тогда \mathbf{W}' можно вычислить непосредственно из уравнений (2.8) и (2.11).

Во всех тех случаях, когда допускается ошибка, процедура обучения предполагает итерацию по всем точкам образов обучающей выборки с внесением поправок в весовой вектор до тех пор, пока разделяющая функция не сведется к функции, правильно классифицирующей все точки. Эта процедура аналогична такому процессу обучения, когда тот или иной ряд вопросов задают повторно до тех пор, пока обучаемый прибор не станет отвечать на них правильно во всех случаях. Подобный порядок и привел к введению таких терминов, как «обучение» и «обучающаяся машина». В этом смысле обучение представляет улучшение рабочих характеристик по мере приобретения навыков.

Как отмечалось выше, процедура исправления ошибок позволяет отыскать решение, если таковое существует (это можно доказать). Поэтому исходный весовой вектор можно взять в произвольном виде, хотя, естественно, такой выбор целесообразнее проводить с учетом какой-либо информации.

Здесь следует сказать несколько слов об объеме обучающей выборки N в зависимости от числа координат d точки, представляющей образ. Точно ответить на вопрос о величине отношения N/d , требующегося для надежного обучения, не представляется возможным, однако принято считать, что чем оно больше, тем лучше. Дополнительную трудность создает то обстоятельство, что величина d имеет меньшее значение по сравнению с числом признаков образа, которое необходимо для разбиения на два класса. Заранее это число обычно неизвестно. К настоящему времени в общем установлено, что, если отношение N/d больше порога, приблизительно равного 3, то получаемые результаты не должны вызывать сомнений.

СВОЙСТВА ПОРОГОВЫХ ЛОГИЧЕСКИХ ЭЛЕМЕНТОВ

Ниже рассмотрены четыре характеристики пороговых логических элементов и прочих устройств, предназначенных для классификации образов: *распознающая способность, скорость сходимости, надежность и способность к предвидению (прогнозирующая способность)*.

Распознающую способность определяют как способность разделяющей функции правильно классифицировать те образы, для которых она была выведена, т. е. образы обучающей выборки. Насколько правильно эта функция способна отвечать на вопросы, которые ставили при ее выводе? Здесь напрашивается аналогия с экзаменами, на которых задают вопросы только по пройденному материалу. Однако этот прием может оказаться весьма полезным, позволяя свести процедуру классификации к простой математической операции и отказаться от запоминания большого числа данных и поиска требующейся информации, что дает большую экономию во многих приложениях. В настоящей книге полное (100%-ное) распознавание принимается идеальным, или безошибочным, обучением классификатора.

Скорость сходимости есть скорость, с которой обучающий алгоритм приходит к полному распознаванию. Эта характеристика представляет интерес с точки зрения экономичности разработки эффективных классификаторов образов. Быстрая сходимость важна в случаях ограниченности бюджета машинного времени.

Хотя медленная сходимость еще не означает неразрешимость той или иной задачи, она часто свидетельствует о том, что такую задачу нельзя решить с разумной затратой времени и средств. Поэтому иногда для повышения скорости сходимости приходится поступаться другими рабочими характеристиками классификаторов.

Надежность характеризует способность классификатора правильно классифицировать данные, использованные при его разработке, но претерпевшие те или иные искажения. Всем процессам передачи информации присущ некоторый уровень шума. Например, классификатор можно разработать по масс-спектрам стандартной выборки химических соединений, но последние при повторных исследованиях могут и не давать точно такие же масс-спектры. Степень надежности отражает способность классификатора правильно перерабатывать подобные искаженные, или зашумленные, данные. Более того, она связана с так называемой избыточностью самой процедуры решения. Ясно, что распознающие системы человека

обладают невероятно высокой надежностью классификации образов некоторых видов. Человеку требуется лишь небольшая выборка из огромного множества существующих данных, чтобы узнать, скажем, любимого котенка. Такие фундаментальные свойства, как размеры, окраска, вес и т. п., при этом распознавании не играют почти никакой роли. Скорее здесь более важны общие контуры тела (силуэт), походка и т. д. Ясно, что приложения теории распознавания образов к задачам химии должны предполагать определенную «терпимость» к шуму. Именно в этом и заключается преимущество методов распознавания образов по сравнению с обычным прямым сравнением данных с библиотечными подборками.

Вероятно, самой привлекательной стороной применений теории распознавания образов в химии надо считать способность классификатора *предсказывать*, т. е. классифицировать образы, отсутствовавшие в обучающей выборке, — выявлять неизвестное «новое». Если бы классификатор образов оказался способным уверенно классифицировать такие неизвестные данные, то это означало бы, что при выводе решающей функции удалось выявить какие-то фундаментальные связи между классифицируемыми образами и их классификациями. Подобная способность может не только выявлять неизвестные образы, но и подсказать путь к установлению неизвестных фундаментальных связей, углубляя понимание причинных зависимостей в химии.

ПРЕДВАРИТЕЛЬНАЯ ОБРАБОТКА И ПРЕОБРАЗОВАНИЯ ИСХОДНЫХ ДАННЫХ

Предварительная обработка исходных данных часто определяет успех применения методов распознавания образов и упрощает саму процедуру классификации. Если воспользоваться понятием n -мерного евклидова пространства, называемого пространством образов, то предварительная обработка исходных данных или их преобразование сводятся к следующим задачам:

- 1) раздвинуть кластеры родственных точек, чтобы упростить классификацию;

- 2) уменьшить размерность пространства образов, чтобы добиться экономии времени и средств при классификации.

Нередко одна задача противоречит другой. Однако если сделать расстояние между классами больше, то классификация упрощается. Уменьшение размерности пространства образов снижает расходы на классификацию, потому что время, требующееся для расчета разделяющей функции, как правило, тем больше, чем больше размерность этого пространства. Таким образом, уменьшение размерности может привести к сокращению объема выборки данных, который необходим, чтобы избежать неопределенности классификации.

В связи с предварительной обработкой исходных данных получили распространение следующие термины: препроцессорная обработка, трансгенерирование или преобразование; отбор признаков; выделение признаков. *Препроцессорная обработка и трансгенерирование* относятся к изменению единичных компонент вектора образов. Хотя для преобразования этих компонент могут быть использованы разные функциональные зависимости, размерность пространства изображений остается при этом прежней. Следовательно, здесь речь идет о преобразованиях, которые не изменяют размерности пространства изображений. *Отбор признаков* означает выбор в исходных данных таких признаков, которые считаются более важными. *Выделение признаков* относится к сочетанию отдельных элементарных признаков в признаки более высокого уровня. Эти операции предполагают преобразования, при которых размерная

независимость не соблюдается. К их числу относятся, например, формирование перекрестных членов, преобразование Фурье, факторный анализ и т. п.

Настоящая глава посвящена рассмотрению ряда примеров линейной препроцессорной обработки химических данных, при которых сохраняется размерность пространства изображений. Поскольку данную стадию нельзя рассматривать отдельно от всей системы классификации образов, при сопоставлении различных подходов к препроцессорной обработке приходится обращаться к общим рабочим характеристикам бинарного классификатора. Поэтому в настоящей главе приведены некоторые рабочие характеристики распознающих систем в целом. Однако объем сведений, обстоятельно излагаемых в последующих главах, сведен к минимуму. В начальных разделах данной главы рассмотрена препроцессорная обработка данных, заимствованных из одного источника, т. е. характеризующихся тесной взаимосвязью замеров. В последующих же разделах главное внимание уделено препроцессорной обработке разрозненных данных, т. е. заимствованных из разных источников и полученных в ходе самостоятельных, не связанных между собой измерений.

МАСС-СПЕКТРЫ

Из химических данных мы в первую очередь рассматривали масс-спектры. Ниже изложены результаты сопоставления ряда методов препроцессорной обработки масс-спектров низкого разрешения, проведенного одним из авторов настоящей книги [1].

Это исследование было проведено на выборке из массива масс-спектров в записи на магнитной ленте, представленного отделением масс-спектрометрических данных при Научно-исследовательском центре Управления атомной энергии Англии (Олдермастон, графство Беркшир). На этой же ленте был записан 2261 масс-спектр соединений, исследованных по научно-исследовательскому проекту 44 Американского нефтяного института АНИ. Каждый из 600 образцов выборки таких масс-спектров был преобразован в цифровые сигналы с нормированными интенсивностями в диапазоне от 99,99 до 0,01. Использовались только спектры соединений, содержащих атомы углерода, водорода, кислорода и азота. Все спектры, считанные с магнитной ленты для фактического анализа, имели приблизительно по 50—70 пиков. Число положений m/e (отношение масс ионов к их зарядам) составляло 132; не менее 10 из них соответствовали пикам для всей выборки спектров. До фактических расче-

тов на вычислительной машине проводились обычные операции считывания спектров с ленты, записи на входе, отбора по заранее установленным критериям (например, по числу атомов углерода) и настройки устройства на обработку отобранных таким образом спектров согласно остальной части программы. Как правило, на вход подавали по 600 спектров, из которых 300 составляли обучающую выборку, а остальные 300 — контрольную (экзаменационную).

Большое число исходных данных позволило брать при решении конкретной задачи достаточно однородные выборки спектров. В рассматриваемом случае были отобраны масс-спектры соединений, содержащих от 3 до 10 атомов углерода. (Испытания с использованием выборок спектров соединений, в состав молекул которых входит от 3 до 20 атомов углерода, показали, что полученные в нашем исследовании результаты не были артефактами данных и хорошо зарекомендовали себя также при обучении на менее однородных выборках спектров.) Для каждого машинного просчета было использовано по 600 спектров, соответствующих требованиям в отношении числа атомов углерода и поровну поделенных между обучающей и контрольной выборками. Чаще всего в выборке из 600 спектров имелось 35 500 пиков, распределенных по 132 положениям m/e .

Масс-спектры преобразовывали четырьмя не изменявшими размерность методами: извлечением квадратного корня, извлечением корня четвертой степени, логарифмированием и «возведением в нулевую степень». Последнее преобразование фактически сводится к двоичному кодированию, т. е. пику в спектре соответствует единица, а положению, где пик отсутствует, — нуль. Следовательно, это преобразование можно трактовать как пороговую обработку. Полученные результаты приведены в табл. 3.1.

Для каждого из этих четырех преобразований проводили машинный просчет по программе отбора признаков. Главные моменты обычной процедуры отбора признаков заключались в том, что на каждом ее этапе осуществлялось обучение на двух весовых векторах (ВВ), компоненты которых во всех случаях считались первоначально равными либо $+1$ (ВВ = $+1$), либо -1 (ВВ = -1); затем знаки компонент весовых векторов сопоставлялись. На очередном этапе оставляли только те положения m/e , для которых компоненты весовых векторов имели одинаковый знак. Процедуру обучения проводили до тех пор, пока не достигалась полная ясность в отношении всех положений m/e . На этом программа обучения заканчивалась.

В каждой строке табл. 3.1 приведены результаты обучения по приведенной программе для одних и тех же исходных данных после преобразований, указанных в первой колонке. Динамический диа-

Таблица 3.1

Характеристики бинарных классификаторов в зависимости от характера предварительной обработки данных (классификация по содержанию кислорода; число положений m/e равно 132)

Преобразование	Динамический диапазон	Число коррекций через обратную связь $BB=+1/BB=-1$	Прогнозирующая способность, %		Из 132 положений		Общий прогноз, %
			$BB=+1/BB=-1$	средняя	исключено	осталось	
Извлечение квадратного корня	100	123/114	93,7/96,3	95,0	66	43	94,4
Извлечение корня четвертой степени	10	101/107	94,7/94,3	94,5	49	44	94,6
Логарифмирование	4	141/102	95,3/95,3	95,3	44	57	95,5
Возведение в нулевую степень	1	235/229	93,7/94,0	93,8	23	83	94,0
Обучающая выборка Контрольная выборка			+	—	Всего		
			121	179	300		
			126	174	300		

пазон исходных данных для каждого спектра равен 10^4 (от 0,01 до 99,99). Во второй колонке приведены данные о динамическом диапазоне после соответствующего преобразования. В третьей — данные о числе коррекций, необходимых для обеспечения полного 100%-ного распознавания на первом этапе программы отбора признаков (при сохранении всех 132 положений m/e). В четвертой колонке приведены данные о прогнозирующей способности двух классификаторов образов, а в пятой — средние значения этой способности на рассматриваемом первом этапе. Наилучшие результаты по прогнозирующей способности на первом этапе обеспечивает логарифмическое преобразование. В шестой колонке представлены данные о наличии неясных положений m/e , обнаруженных на этом первом этапе. Чем меньше ширина динамического диапазона, тем меньше неясных положений m/e выявляет классификатор. Программа предусматривала многократное повторение описанного цикла для всех преобразований. В седьмой колонке приведены данные о числе положений m/e , оставшихся после отбрасывания всех неясных положений. И опять-таки оказалось, что чем меньше ширина динамического диапазона, тем меньше число неопределенных положений. В последней колонке приведены данные о прогнозирующей способности каждого классификатора, усредненные по всем этапам программы отбора признаков. И в этом отношении логарифмическое преобразование обеспечивало наилучшую прогнозирующую способность. Такой вывод согласуется с результатом, полученным при применении теории информации к решению проблемы преобразования. На последующих этапах исследования исходные данные во всех случаях подвергались логарифмическому преобразованию.

Целесообразность выбора для преобразования логарифмической зависимости обоснована с позиций теории информации в работе [2]. В этом исследовании, включающем сжатие масс-спектров, логарифмически устанавливался ряд переходных уровней интенсивности, соответствовавших следующим относительным значениям полного ионного тока (%): $1/2$, 1, 2, 4, 8, 16 и 32. Это делалось для того, чтобы приблизительно сравнить число пиков на каждом уровне и тем самым по возможности увеличить объем информации, или так называемую энтропию информации, в независимом канале.

Предположения, связанные с динамическим диапазоном параметров, сильнее всего влияют на сходимость для пороговых логических элементов. Как показано в ряде исследований, приблизительное уравнивание динамических диапазонов для всех признаков часто приводит к достижению лучших рабочих характеристик. Поскольку здесь речь идет о преобразовании, не затрагивающем

размерности пространства изображений, оно не отражается на разделимости бинарных данных выборки. А любой процесс, обеспечивающий более быструю сходимость, наиболее выгоден, так как позволяет экономить машинное время.

Известно [3], что динамический диапазон изменения разных компонент образов может изменяться в широких пределах, однако вне определенных пределов время сходимости резко возрастает. В работе [3] значение $(d + 1)$ -й компоненты любого вектора образа варьировалось в пределах нескольких порядков величины. Как выяснилось, когда эта компонента во много раз превосходит средние значения других компонент, сходимость достигается крайне медленно.

Как показано в ряде работ [1, 4, 5], объем информации в бинарных масс-спектрах весьма значителен, поэтому во многих случаях масс-спектры идеально подходят для исследования методами распознавания образов.

ИНФРАКРАСНЫЕ СПЕКТРЫ

В отношении ИК-спектров остаются в силе все рассуждения, изложенные выше по поводу предварительной обработки масс-спектров. В исследовании [6] ИК-спектры характеризовались всего четырьмя уровнями амплитуды.

В настоящем исследовании были использованы данные фирмы «Садтлер ризёрч лабораториз» (Филадельфия, США). Из 24 142 ИК-спектров стандартных соединений были отобраны попавшиеся первыми 4500 спектров таких соединений, число атомов в которых не превышало 10 для углерода, 4 для кислорода и 3 для азота и которые не содержали других элементов, кроме водорода. Спектры были зарегистрированы в виде полос шириной 0,1 мкм в диапазоне 2,0—14,9 мкм. Это давало 130 координат образа, включая $(d + 1)$ -й член. Амплитуде каждой компоненты образа приписывали то или иное из четырех значений, основываясь на максимальном поглощении в соответствующей полосе шириной 0,1 мкм. Амплитуду максимального пика в спектре считали равной 3,0. Наибольший пик в полосе шириной 1,0 мкм полагали имеющим высоту 2,0. Всем прочим пикам приписывали амплитуду единичной длины. Отсутствие пиков в той или иной полосе шириной 0,1 мкм характеризовали нулевой амплитудой. Поскольку амплитуды могли иметь только ограниченные значения (три ненулевых и одно нулевое) и в большинстве случаев они были равны нулю, представлялось целесообразным сжать информацию об образах, чтобы сэкономить память

машины и сократить время расчетов. Поэтому каждый образ переписывали в виде целочисленного ряда

$$n_1, p_1, p_2, p_3, \dots, p_{n_1}, n_2, q_1, q_2, q_3, \dots, q_{n_2}, n_3, r_1, r_2, r_3, \dots, r_{n_3},$$

где n_1, n_2, n_3 — числа пиков с амплитудами, равными соответственно 1,0, 2,0 и 3,0, а $p_1, \dots, p_{n_1}, q_1, \dots, q_{n_2}$ и r_1, \dots, r_{n_3} — координаты, или позиции, таких пиков. Объем памяти, необходимый для запоминания результирующих образов, при сжатии спектров сократился более чем в три раза. Кроме того, скалярное умножение векторов для определения требующихся при классификации скалярных величин удалось осуществить следующим образом. Если \mathbf{W} — весовой вектор, а \mathbf{Y} — вектор образа, для которого надо получить скалярное произведение, то

$$\mathbf{W} \cdot \mathbf{Y} = w_1 y_1 + w_2 y_2 + w_3 y_3 + \dots + w_{d+1} y_{d+1}. \quad (3.1)$$

Однако, поскольку модуль вектора \mathbf{Y} принимает только одно из значений 0,0, 1,0, 2,0 и 3,0, скалярное произведение можно вычислить в виде

$$\mathbf{W} \cdot \mathbf{Y} = \left(\sum_{j=1}^{n_1} w_{p_j} \right) + 2,0 \left(\sum_{j=1}^{n_2} w_{q_j} \right) + 3,0 \left(\sum_{j=1}^{n_3} w_{r_j} \right) + w_{d+1}. \quad (3.2)$$

При таком описании ИК-спектров для машинного вычисления скалярных произведений требовалось приблизительно в 20 раз меньше времени.

В независимом исследовании [7] по расшифровке ИК-спектров описание исходных данных проводилось совершенно другим способом. Источником данных была все та же фирма «Садтлер ризёрч лабораториз», у которой были приобретены эталонные ИК-спектры. Каждый спектр в диапазоне длин волн 2,0—15,0 мкм разбивали на полосы шириной 0,1 мкм, что в результате давало 131 интервал, и, следовательно, 131 значение x_u . Коэффициент пропускания для каждого интервала переводили в коэффициент поглощения, величину которого выражали ради удобства целыми числами от 0 до 9.

СПЕКТРОСКОПИЧЕСКИЕ ДАННЫЕ ИЗ МНОГИХ ИСТОЧНИКОВ

Еще одна выигрышная сторона распознавания образов заключается в том, что можно использовать данные из разных источников и считать их принадлежащими одному образу. В линейной обучаю-

щейся машине, например, отдельные компоненты спектров обрабатываются как независимые переменные. Поэтому значения спектра, скажем, в 100 положениях по оси абсцисс можно трактовать как результаты 100 независимых опытов. Таким образом, появляется возможность объединить совершенно разные экспериментальные данные (например, ИК-спектры или масс-спектры одного и того же соединения) и рассматривать их как единый образ. В конце концов, оба этих спектрометрических метода используются для уточнения информации об одном соединении. И хотя экспериментальные данные могут быть получены разными путями, можно полагать, что определенные компоненты, составленные из фрагментов образа, и инфракрасное поглощение связаны между собой не в меньшей степени, чем те или иные сочетания самих ИК-спектров.

Работа [5] посвящена исследованию комбинации ИК-спектров, масс-спектров и данных о температурах плавления и кипения как характеристике образа химического соединения. Здесь весьма важное значение приобретают соображения о динамическом диапазоне изменения параметров, поскольку данные из разных источников выражены в произвольных масштабах. Если экспериментальные данные, полученные по одной методике, по величине превосходят результаты опытов, проведенных по другой методике, то первые окажут преобладающее влияние на результат распознавания образов. Если же данные, полученные двумя методами, отнести к одному динамическому диапазону, то признаки образа можно выделить из данных двух источников и в результате достичь высшей прогнозирующей способности.

Авторы этого исследования формировали образы того или иного соединения комбинированием его масс-спектра низкого разрешения с ИК-спектром. Таким способом были составлены образы для 291 соединения. Масс-спектры заимствованы из таблиц Американского нефтяного института, составленных в порядке осуществления уже упоминавшегося научно-исследовательского проекта 44. Пики в этих спектрах, распределяющиеся по 132 возможным положениям m/e , имеют амплитуды в диапазоне 10—100. Их удалось привести к этому диапазону путем извлечения квадратного корня. ИК-спектры были предоставлены во временное пользование фирмой «Садтлер ризёрч лабораториз». В этих спектрах пики ИК-поглощения распределялись по 130 возможным положениям в интервале длин волн 2,0—14,9 мкм. Амплитудам этих пиков приписывали одно из следующих четырех значений: 1) 0,0 при отсутствии пиков; 2) 1,0 пику в полосе шириной 0,1 мкм; 3) 2,0 самому высокому пику в полосе шириной 1,0 мкм и 4) 3,0 самому высокому пику

в спектре. Следовательно, каждый такой ИК-спектр состоял из совокупности 130 упорядоченных чисел. Исследуемые соединения отвечали общей формуле $C_{1-10}H_{1-24}O_{0-4}N_{0-3}$. Каждый образ соединения, составленный из масс-спектра и ИК-спектра, имел по 262 компоненты. Из 291 такого образа 191 произвольно отбирали в обучающую выборку, а остальные 100 — в контрольную.

При объединении данных двух разных источников в единый образ поведение обучающейся машины сильно зависит от относительного вклада данных каждого из двух (или нескольких) типов.

Чтобы оценить эффект комбинирования данных из разных источников, соединения классифицировали по наличию одной или нескольких двойных связей. Попытки осуществить такую классификацию исходя только из ИК-образов или только из масс-спектрометрических образов не принесли заметного успеха, о чем можно судить по данным, приведенным в первых двух разделах табл. 3.2. В каждом из этих двух случаев брали 125 исходных параметров (координат) и путем обычной процедуры отбора признаков отбрасывали те из них, которые из-за относительно малой величины соответствующих компонент весовых векторов считались сравнительно малозначащими. В случае ИК-спектров прогнозирующая способность, составлявшая ~82%, начинала быстро убывать, когда оставалось менее 50 параметров. В случае масс-спектрометрических данных прогнозирующая способность находилась на уровне ~87%; затем она медленно убывала с уменьшением числа параметров, пока оно не доходило до 20; после этого прогнозирующая способность быстро падала до уровня случайного угадывания. Масс-спектрометрические данные тоже не обеспечивали сходимости в пределах отведенного числа коррекций через обратную связь, когда оставалось менее 50 параметров.

В разд. 3—5 табл. 3.2 приведены данные о результатах обучения распознаванию образов, сформированных комбинированием масс-спектра с ИК-спектром. В разд. 3 образы были нормированы таким способом, что интенсивности пиков масс-спектров значительно превосходили ИК-компоненты. Интенсивности пиков в масс-спектрах находились в диапазоне 10—100, тогда как интенсивности линий в ИК-спектрах приравнивались четырем значениям: 0, 1, 0, 2, 0 и 3, 0. Таким образом, вклады ИК-спектров в общую протяженность векторов образов были относительно малыми по сравнению с вкладами масс-спектрометрических компонент данного образа. Как и предполагалось, в этом случае результаты обучения оказались почти такими же, как и при использовании масс-спектрометрических данных; к тому моменту, когда число параметров до-

Таблица 3.2

Данные классификации по наличию двойных связей

Число параметров	1. Инфракрасная спектроскопия (ИК)			2. Масс-спектрометрия (МС)			3. Комбинированные образы			
	число коррекций через обратную связь	число опознанных соединений	прогнозирующая способность, %	число коррекций через обратную связь	число опознанных соединений	прогнозирующая способность, %	число коррекций через обратную связь	число опознанных соединений	прогнозирующая способность, %	МС:ИК
262							1182		86	136:126
162							1024		84	99:6
125	156		79	1105		87	1005		87	92:33
100	144		83	958		88	1056		85	88:12
70	177		82	1529		85	1176		85	69:1
50	261		77	1677		84	1845		85	50:0
30	>5000	168	62	>5000	179	81	>5000	182	79	30:0
20	>5000	136	47	>5000	178	81	>5000	179	82	20:0
10	>5000	129	50	>5000	141	69	>5000	135	61	10:0
5	>5000	129	69	>5000	131	56	>5000	142	61	5:0

Продолжение табл. 3.2

Число параметров	4. Комбинированные образы			МС:ИК	5. Комбинированные образы			МС:ИК	6. Комбинированные образы			МС:ИК: Прочие
	число коррекций через обратную связь	число опознаваемых соединений	прототипизирующая способность, %		число коррекций через обратную связь	число опознаваемых соединений	прототипизирующая способность, %		число коррекций через обратную связь	число опознаваемых соединений	прототипизирующая способность, %	
262	141		78	136:126	105		89	136:126	87		87	136:126:2
162	149		80	52:110	82		88	76:86	83		88	76:76:2
125	145		83	24:101	82		89	61:64	90		89	55:70:2
100	133		80	11:89	87		90	46:54	88		89	44:56:2
70	121		81	1:69	93		89	30:40	121		91	31:39:2
50	128		78	0:50	169		89	23:27	129		90	25:25:2
30	>5000	186	69	0:30	146		92	13:17	468		89	15:15:2
20	>5000	140	52	0:20	>5000	174	88	9:11	1520		92	12:8:2
10	>5000	136	52	0:10	>5000	164	75	6:4	>5000	176	84	8:2:2
5	>5000	139	55	0:5	>5000	137	62	4:1	>5000	129	64	5:1:1

ходило до 50, оставались лишь масс-спектрометрические компоненты образов. Результаты, приведенные в разд. 4 табл. 3.2, иллюстрируют эффект нормировки образов, когда преобладает влияние ИК-спектров. В данном случае интенсивности считались равными 0, 1000, 2000 и 3000, так что большая часть протяженности векторов образов была обусловлена вкладом ИК-спектра. Как и предполагалось, результаты обучения оказались почти такими же, как при обучении только на ИК-спектрах, а когда оставалось 50 параметров, все эти параметры уже принадлежали ИК-образам.

В разд. 5 табл. 3.2 приведены результаты нормировки образов, когда вклады данных обоих источников в полную амплитуду векторов образов множества одинаковы. Интенсивности ИК-полос для всей совокупности данных приравнивались сумме интенсивностей всех пиков в масс-спектрах для той же совокупности. В этом случае исходная прогнозирующая способность составляла ~90% и оставалась весьма высокой, пока число компонент образа не становилось меньше 20. Даже в том случае, когда оставалось всего 10 компонент, прогнозирующая способность все еще была равна 75% при распознающей способности 82%, что намного лучше, чем случайное угадывание. Интересно отметить и то, что как масс-спектрометрические, так и ИК-компоненты сохранялись на протяжении всей процедуры уменьшения числа параметров.

Данные, приведенные в разд. 6 табл. 3.2, свидетельствуют о дальнейшем улучшении классифицирующей способности, достигаемом добавлением к вектору образа температур плавления и кипения. Отметим, что в каждом случае дополнения температурами плавления и кипения общее число параметров было на два больше, чем для образов, комбинированных только из масс-спектрометрических и инфракрасных данных. Сравнение этих результатов с данными, приведенными в разд. 5 табл. 3.2, показывает, что заметного улучшения не наблюдается, пока число параметров остается не меньше 30, поскольку прогнозирующую способность и скорость сходимости в этих двух случаях можно считать приблизительно одинаковыми. Однако добавление к компонентам образов температур кипения и плавления приводит к тому, что обучающаяся машина все еще обнаруживает сходимость даже при 20 параметрах, сохраняя на этом уровне приблизительно 90%-ную прогнозирующую способность. Более того, для 10 параметров она все еще заметно выше, чем в прочих случаях. Процедура принятия решения, которая использовалась для исключения параметров, мало способствующих классификации, оставляет сведения о температурах плавления и кипения почти до самого конца расчетов.

Таблица 3.3

**Данные классификации по наличию двойных связей при использовании
«бинарных спектров»**

Число параметров	1. Инфракрасная спектроскопия				2. Масс-спектрометрия				3. Комбинированные образы		
	число коррекций через обратную связь	число опознанных соедине- ний	прогнози- рующая способ- ность, %	число коррекций через обратную связь	число опознан- ных соединений	прогнози- рующая способ- ность, %	число коррек- ций через обратную связь	число опознан- ных соединений	прогнози- рующая способ- ность, %		
262							124			92	
162							110			89	
125	183		81	454		84	122			90	
100	173		83	486		85	119			89	
70	202		79	373		85	99			88	
50	193		79	494		84	105			89	
30	>5000	172	77	>5000	175	85	234			89	
20	>5000	165	65	>5000	161	81	>5000	178		85	
10	>5000	144	56	>5000	157	75	>5000	161		54	
5	>5000	142	61	>5000	155	69	>5000	158		72	

При помощи аналитического оборудования, например ИК- и масс-спектрометров, зачастую легче получить точные сведения о параметре, по которому делается отсчет, например, о длине волны в первом случае и массе во втором, чем об измеренной интенсивности. Поэтому полезно уметь оценивать степень важности информации об интенсивности. В табл. 3.3 приведены данные подобной оценки результатов классификации по наличию двойных связей, обобщенных в табл. 3.2. При этой оценке все компоненты и масс-спектра, и ИК-спектра считались имеющими единичную интенсивность, когда обнаруживались пики, и нулевую, когда они не обнаруживались. Затем на таких «бинарных спектрах» проводили обучение прежним способом. Сопоставление результатов, приведенных в разд. 1—3 табл. 3.3, с данными разд. 1, 2 и 5 табл. 3.2 показывает, что простые сведения о наличии или отсутствии пиков, по-видимому, позволяют распознавать образы ничуть не хуже соответствующих данных об интенсивности. Иными словами, чтобы получить ответ на вопрос о наличии двойных связей, вполне достаточно иметь информацию о положении пиков как в ИК-спектре, так и в масс-спектре.

Еще об одном исследовании возможностей использования данных из разных источников сообщается в работе [8]. Здесь были использованы данные о масс-спектрах низкого разрешения, результатах измерений методом ядерного магнитного резонанса, показателях преломления и плотностях для чистых углеводов. Авторы интересовали возможности определения типов углеводов и структуры средней молекулы в сложной смеси углеводов (бензин). Векторы образов, составленные по данным разных источников, вводились в алгоритм распознавания образов с использованием процедуры обучения по методу наименьших квадратов.

ЭЛЕКТРОХИМИЧЕСКИЕ СПЕКТРЫ

Авторы работы [9] исследовали методами распознавания образов полярографические вольт-амперные кривые, получаемые при применении стационарных электродов (СЭ). Задача исследования заключалась в изучении возможности отличить по СЭ-полярограммам методами распознавания образов наличие одного вещества в присутствии двух или более веществ. Как и при любом исследовании, проводимом методами распознавания образов, первый шаг заключался в выделении каких-либо особенностей, которые можно было бы использовать при формировании векторов образов.

На основе аналитических зависимостей была составлена совокупность СЭ-полярограмм, введенная в память машины после преобразования в цифровые сигналы. Затем с этими данными оперировали как с экспериментальными результатами. На основе преобразованных в цифровые сигналы СЭ-полярограмм, записанных в виде кривых изменения силы тока во времени, а также вычисленных для них первой и второй производных была определена совокупность 133 признаков (рис. 3.1). В число признаков были включены замеры изменения силы тока в зависимости от потенциала в определенных допиковых и послепиковых положениях, изменения потенциала для токов, составлявших определенные доли пикового тока, данные об асимметричности кривых, величине площадей под разными участками СЭ-полярограмм и их соотношениях, а также разные параметры, связанные со значениями производных в критических точках. Эти признаки были затем подразделены на три общие категории признаков формы, пикового тока и максимального напряжения.

В зависимости от характера процедуры выделения признаков любому такому признаку свойственны свой диапазон изменения и свое распределение значений. Чтобы избежать возможных в связи с этим осложнений, пришлось прибегнуть к предварительной обработке исходных данных.

Такую предварительную обработку осуществляли двумя способами. Первый из них получил название фиксации диапазона. Эта процедура предполагала преобразование исходного значения x_{ij} i -го признака для j -ой полярографической кривой по формуле

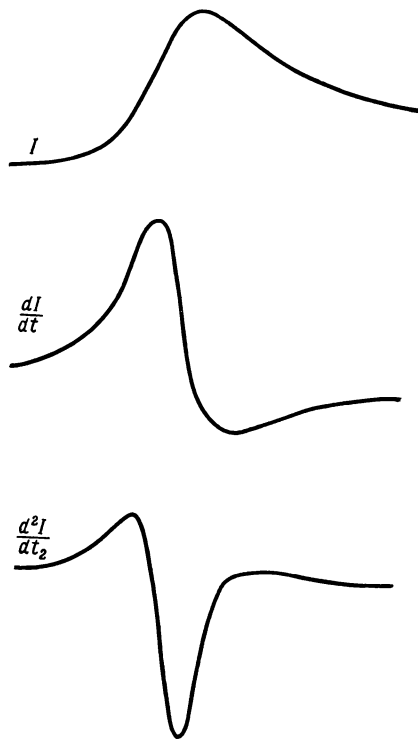


Рис. 3.1. Общий вид полярографических кривых для стационарных электродов.

$$x'_{ij} = a_i x_{ij} + b_i, \quad (3.3)$$

где a_i и b_i — постоянные, связанные с максимальным и минимальным значениями рассматриваемого признака. Эти постоянные можно выбрать так, чтобы зафиксировать для признака i любой диапазон его изменения.

Второй способ преобразования заключался в автоматическом подборе масштаба, рассмотренном в работе [10], и основан на использовании для пересчета соотношения

$$x'_{ij} = \frac{x_{ij} - \bar{x}_i}{\sigma_i}, \quad (3.4)$$

в котором \bar{x}_i — среднее арифметическое, σ_i — стандартное отклонение значений x_{ij} для i -го признака. Автоматический пересчет масштаба означает приравнивание среднего арифметического нулю, а стандартного отклонения — единице. В процессе всего исследования эти авторы использовали для преобразования автокорреляционную функцию. Полученные результаты для полярографических кривых изложены в гл. 4.

СПИСОК ЛИТЕРАТУРЫ

1. Jurs P. C., Anal. Chem., **43**, 22 (1971).
2. Grotch S. L., Anal. Chem., **43**, 1362 (1971).
3. Wangen L. E., Isenhour T. L., Anal. Chem., **42**, 737 (1970).
4. Jurs P. C. et al., Anal. Chem., **41**, 690 (1969).
5. Jurs P. C. et al., Anal. Chem., **41**, 1949 (1969).
6. Kowalski B. R., et al., Anal. Chem., **41**, 1945 (1969).
7. Liddell R. W., III, Jurs P. C., Appl. Spectros., **27**, 371 (1973).
8. Tunnicliff D. D., Wadsworth P. A., Anal. Chem., **45**, 12 (1973).
9. Sybrandt L. B., Perone S. P., Anal. Chem., **44**, 2331 (1972).
10. Kowalski B. R., Bender C. F., J. Am. Chem. Soc., **94**, 5632 (1972).

ПОСТРОЕНИЕ РАЗДЕЛЯЮЩЕЙ ФУНКЦИИ

Среди разнообразных разделяющих функций самой простой в применении и потому наиболее распространенной в химии является линейная разделяющая функция. Как отмечалось выше, линейная разделяющая функция эквивалентна некоторой весовой функции, при умножении которой на вектор образа получается скалярный результат. Несмотря на то что принципиально возможна множественная классификация, самым простым классификатором служит бинарное устройство, дающее один из двух альтернативных ответов. При использовании для бинарной классификации линейной разделяющей функции удобно определять принадлежность образа к одному из двух классов по знаку скаляра.

Большая часть настоящей главы посвящена результатам применения линейных бинарных классификаторов для решения задач расшифровки химических данных. В остальных разделах главы рассмотрены вопросы нелинейной и множественной классификации.

БИНАРНЫЕ КЛАССИФИКАТОРЫ

Простой пороговый логический элемент

Как было показано выше, основной пороговый логический элемент производит классификацию, вычисляя скалярное произведение весового вектора W на вектор X подлежащего классификации образа. Классификацию осуществляют по знаку скалярного произведения:

Если $s > 0$, то X относят к категории 1;

Если $s \leq 0$, то X относят к категории 2.

Весовой вектор находят при помощи процедуры исправления ошибок через обратную связь, выполняемой на обучающей выборке известных векторов образов. Данный раздел посвящен исследова-

нию основных вариантов реализации порогового логического элемента.

О первоначальном исследовании скорости сходимости и прогнозирующей способности пороговых логических элементов сообщается в статье [1]. Данные были заимствованы из таблиц Американского нефтяного института, составленных по проекту 44. Массив данных состоял из 630 масс-спектров низкого разрешения соединений, имеющих молекулярный состав $C_{1-10}H_{2-22}O_{0-4}N_{0-2}$. Интенсивности пиков давались в процентах от интенсивности основного пика каждого спектра. Рассматривались только пики с интенсивностью не менее 1% основного пика; большая часть спектров имела от 15 до 40 таких пиков. Во всем массиве данных пикам соответствовали 155 положений m/e , так что векторы образов имели размерность 156. Интенсивности, вошедшие в векторы образов, были преобразованы в препроцессоре извлечением квадратного корня. В табл. 4.1 приведены результаты «обучения» весовых векторов для обнаруже-

Таблица 4.1

**Скорость сходимости и прогнозирующая способность при обнаружении кислорода
в зависимости от объема обучающей выборки**

Объем обучающей выборки	Число выбранных спектров ^а	Прогнозирующая способность, %	
		фактическая	средняя
50	226	88,3	86,2
	180	84,3	
	255	85,9	
100	654	90,8	88,2
	561	87,7	
	574	86,0	
200	1528	86,6	88,4
	1525	88,6	
	2189	88,1	
300	2590	92,7	90,6
	1959	90,6	
	2986	88,5	

^а Число спектров, необходимых для полного обучения.

ния присутствия кислорода, полученные на случайных выборках из общего массива 630 масс-спектров низкого разрешения. Как правило, скорость сходимости была тем меньше, чем больше объем обучающей выборки. Эту тенденцию показывают данные табл. 4.1, которые относятся к обучению всех весовых векторов до полного распознавания. Интересно отметить, что прогнозирующая способность весовых векторов значительна даже при объемах обучающей выборки в 50 и 100 спектров, т. е. когда число подгоняемых параметров (компонент весового вектора, соответствующих положениям m/e) превосходит число образов в обучающей выборке. Прогнозирующая способность должна быть тем выше, чем больше объем обучающей выборки. Это и подтверждается данными табл. 4.1, несмотря на значительный уровень шума.

Возрастание прогнозирующей способности любого классификатора с увеличением объема обучающей выборки — один из основных моментов распознавания образов во всех его формах. Как уже отмечалось в гл. 1, байесова теория решений позволяет вычислять наиболее вероятный класс того или иного образа, когда известно распределение классифицируемых образов. Однако распределение какого-либо образа в генеральной совокупности данных часто не известно. Например, согласно данным химического структурного анализа, синтезированные химические соединения исчисляются миллионами, и для довольно многих из них определены масс-спектры. Однако наиболее состоятельные выборки масс-спектров составлены лишь для нескольких тысяч соединений. Но даже если бы все спектры известных химических соединений были получены в стандартных условиях, то была бы охвачена лишь небольшая часть всех возможных соединений. Следовательно, любая классификационная схема должна исходить из оценки уместного вероятностного распределения, если не существует надежной теории, способной предсказать, каким должен быть масс-спектр любого возможного химического соединения.

Таким образом, какую бы классификационную схему мы ни выбрали, она сама должна в той или иной форме оценивать распределение неизвестных масс-спектров на основе совокупности известных спектров. Линейный бинарный классификатор делает это при помощи решающей поверхности, эффективно делящей многомерное пространство изображений масс-спектров на два интересующих нас класса. Затем такой классификатор используется для классификации неизвестных спектров. Эту математическую процедуру химики обычно на практике не применяют, но по существу она сводится к предсказанию новых ответов на базе прошлой практики. Поэтому

процедура обучения и предсказания представляет обычный способ, который позволяет ученым интерпретировать данные разных типов. Преимущество такого способа прогнозирования при применении линейных бинарных классификаторов состоит в том, что он допускает проверку. Для этого в качестве обучающей выборки берут случайно выбранную совокупность известных данных, а на остальной части исходного массива данных проверяют прогнозы после завершения обучения. Результаты проверки служат мерой прогнозирующей способности, поскольку данные, взятые для «экзамена», не использовались при обучении. Разумеется, всегда существует вероятность того, что необычно высокая прогнозирующая способность есть следствие артефакта, однако не исключено, что бинарный классификатор образов с высокой прогнозирующей способностью отражает фундаментальные свойства изучаемого явления.

В статье [2] сообщается о попытках проверить, в какой степени бинарные классификации образов могут быть использованы для определения параметров молекулярной структуры на основе данных масс-спектрометрии низкого разрешения. Такую проверку проводили на массиве данных, заимствованном из уже упоминавшихся таблиц Американского нефтяного института. В этом исследовании данные были разделены на две совокупности: 387 спектров углеводородов (CH) и 243 спектра соединений, содержащих кислород и азот (CHON).

Из 387 спектров углеводородов 200 случайно выбранных спектров были включены в обучающую выборку для обучения весовых векторов, а на остальных 187 проверяли их прогнозирующую способность. В табл. 4.2 приведены результаты обучения и проверки 43 весовых векторов, построенных для определения структурных параметров углеводородов. Каждый вектор должен был обеспечивать бинарное решение. Для всех параметров, кроме отношения числа атомов углерода к числу атомов водорода, положительный ответ означал, что данный параметр имеет значение выше порогового, а отрицательный, — что параметр меньше или равен порогу. Например, первый вектор (9 атомов углерода) настраивали с таким расчетом, чтобы он давал положительное скалярное произведение для масс-спектров соединений, содержащих 10 атомов углерода, и отрицательное, — когда в состав молекул входит не больше 9 атомов углерода. Вектор параметра C:H должен был давать положительный ответ, когда это отношение точно соответствует формуле соединений интересующего класса, и отрицательный — во всех других случаях. Например, *n*-гексан дает положительное скалярное произведение для $C:H = 2n + 2$ и отрицательное — для всех

Результаты обучения бинарных классификаторов образов на углеводородах

Таблица 4.2

	Порог	Обучающая выборка			Экзамениционная выборка		
		отрицательная категория	положительная категория	число коррекций через обратную связь	отрицательная категория	положительная категория	прогнозирующая способность, %
Число атомов углерода	9	163	37	227	154	33	89,3
	8	121	79	167	113	74	92,5
	7	80	120	185	77	110	93,6
	6	53	147	99	44	143	94,1
	5	30	170	71	21	166	97,9
	4	15	185	42	7	180	97,3
Число атомов водорода	20	196	4	53	182	5	97,3
	18	168	32	170	154	33	95,7
	16	143	57	202	132	55	97,3
	14	110	90	58	100	87	94,1
	12	72	138	51	55	132	95,2
	10	47	153	59	34	153	96,8
Отношение C:H	8	28	172	31	19	168	96,8
	6	9	191	34	10	177	97,3
	$2n + 2$	156	44	25	143	44	96,8
	$2n$	125	75	28	107	80	96,8
	$2n - 2$	153	47	36	154	33	96,8
	$2n - 4$	191	9	39	180	7	98,9
Метил	$2n - 6$	185	15	13	170	17	98,9
	4	191	9	177	165	22	90,9
	3	160	40	800	136	51	86,1

Продолжение табл. 4.2

	Порог	Обучающая выборка			Экзаменационная выборка		
		отрицательная категория	положительная категория	число коррекций через обратную связь	отрицательная категория	положительная категория	прогнозирующая способность, %
Метил	2	114	86	859	97	90	86,6
	1	62	138	648	47	140	86,6
	0	29	171	328	24	163	89,3
Этил	1	166	34	1518	153	34	80,7
	0	104	96	>2000	97	90	73,3
<i>n</i> -Пропил	1	191	9	211	177	10	90,4
	0	145	55	>2000	148	39	71,7
Наибольшее кольцо	6	199	1	11	184	3	97,9
	5	142	58	141	137	50	89,8
	4	121	79	149	114	73	90,9
	3	120	80	194	112	75	91,4
Число атомов углерода в боковой цепи	2	174	26	356	163	24	90,9
	1	108	92	>2000	90	97	61,5
	0	42	158	1486	36	151	87,2
Число связей —C=C—	2	171	29	11	165	22	98,4
	1	159	41	153	158	29	92,5
	0	102	98	>2000	98	89	77,5
Число атомов углерода, не связанных с водородом	1	167	33	432	156	31	83,4
	0	111	89	1327	97	90	70,6
Бензольное кольцо	0	179	21	37	168	19	96,8
Число связей —C≡C—	0	184	16	163	174	13	93,6
CN ₃ =CH—(винил)	0	166	34	>2000	158	29	80,2

других значений этого отношения. Категории, представленные в табл. 4.2, отвечают следующим определениям: числа метильных, этильных и *n*-пропильных групп указывают, сколько таких групп можно получить при разрыве одинарной связи. Так, 3-метилгексан по такому определению содержит три метильные, две этильные и одну *n*-пропильную группы. Классификация по наибольшему кольцу охватывает разбиение на насыщенные, ненасыщенные или ароматические соединения. Число атомов углерода в боковой цепи дает число атомов, связанных непосредственно не менее чем с тремя другими атомами углерода. Бензол классифицируется как кольцо с тремя двойными связями $C=C$. В этой таблице среди структурных параметров отдельно фигурирует число атомов углерода, не связанных с атомами водорода. Весовые векторы в трех последних случаях выявляют наличие или отсутствие бензольных колец, тройных связей $C\equiv C$ и структурных признаков виниловых соединений соответственно.

В третьем и четвертом столбцах таблицы приведены данные о числе соединений двух категорий в обучающей выборке. В пятом—указано число коррекций через обратную связь, необходимых для обеспечения схожести. Если в этом столбце указано >2000 , то это означает, что 2000 таких коррекций не позволили завершить обучение. В шестом и седьмом столбцах приведены данные о числе соединений двух категорий в экзаменационной выборке. Данные в последнем столбце характеризуют процентную долю верных распознаваний 187 соединений, не вошедших в обучающую выборку. Прогнозирующая способность составляла от 61,5 до 98,9% со средним, равным 90,3%. Процент правильных предсказаний может служить мерой достоверности ответа для неизвестного спектра. При случайном угадывании такая вероятность равна 50%. Таким образом, полностью эмпирический вычислительный метод позволяет получать с высоким уровнем достоверности информацию о структуре углеводов.

Из 243 соединений типа $CHON$ 150 случайно выбранных были включены в обучающую выборку, а на остальных 93 соединениях проверяли прогнозирующую способность весовых векторов. В табл. 4.3 приведены результаты обучения 65 весовых векторов, построенных для распознавания разных структурных особенностей. Остановимся на определении категорий, представленных в табл. 4.3. Категория «максимальная длина углеродной цепи» выражается максимальным числом связанных друг с другом атомов углерода. Категории «метил», «этил», «наибольшее кольцо» и «число двойных связей» имеют тот же смысл, что и в рассмотренном выше случае

Таблица 4.3

Результаты обучения бинарных классификаторов образов на соединениях типа CHON

	Обучающая выборка				Экзаменационная выборка		
	порог	отрицательная категория	положительная категория	число коррекций через обратную связь	отрицательная категория	положительная категория	прогнозирующая способность, %
Число атомов углерода	9	141	9	49	86	7	94,6
	8	131	19	120	81	12	87,1
	7	124	26	220	76	17	90,3
	6	111	39	266	71	22	83,9
	5	80	70	516	52	41	80,0
	4	48	102	159	35	58	82,8
	20	146	4	22	92	1	98,9
Число атомов водорода	19	143	7	73	91	2	98,9
	18	141	9	93	91	2	98,9
	17	138	12	84	90	3	96,8
	16	137	13	74	90	3	96,8
	15	130	20	159	89	4	89,2
	14	127	23	115	89	4	92,5
	13	116	34	279	80	13	86,0
	12	113	37	345	78	15	80,6

Продолжение табл. 4.3

	Обучающая выборка				Экзаменационная выборка		
	порог	отрицательная категория	положительная категория	число коррекций через обратную связь	отрицательная категория	положительная категория	прогнозирующая способность, %
Число атомов кислорода	11	92	58	254	66	27	87,1
	10	82	68	229	63	30	82,8
	9	55	95	237	45	48	80,6
	8	49	101	204	39	54	81,7
	7	33	117	80	27	66	81,7
	6	20	130	63	20	73	83,9
	5	14	136	62	10	83	87,1
	2	145	5	46	85	8	94,6
	1	102	48	963	55	38	76,3
	0	43	107	88	26	67	93,5
Число атомов азота	1	146	4	55	87	6	93,5
	0	99	51	67	63	30	91,4
Отношение C:H	2n+3	138	12	26	4	89	97,8
	2n+2	100	50	213	28	65	87,1
	2n+1	87	63	223	32	61	81,7

Продолжение табл. 4.3

	Обучающая выборка				Экзаменационная выборка		
	порог	отрицательная категория	положительная категория	число коррекций через обратную связь	отрицательная категория	положительная категория	прогнозирующая способность, %
Максимальная длина углеродной цепи (по числу атомов углерода)	$2n$	37	113	112	63	30	91,4
	$2n-1$	31	119	113	65	28	91,4
	$2n-2$	22	128	45	71	22	94,6
	$2n-3$	17	133	62	75	18	92,5
	$2n-4$	15	135	36	79	14	93,5
	$2n-5$	9	141	19	80	13	93,5
	9	146	4	20	90	3	96,8
	8	137	13	96	87	6	88,2
	7	132	18	164	81	12	89,2
	6	125	25	168	78	15	85,0
	5	108	42	123	66	27	86,0
	4	89	61	133	56	37	88,2
	3	53	97	193	37	56	85,0
	2	29	121	75	22	71	90,3
	1	10	140	56	4	89	95,7

Продолжение табл. 4.3

	Обучающая выборка				Экзаменационная выборка		
	порог	отрицательная категория	положительная категория	число коррекций через обратную связь	отрицательная категория	положительная категория	прогнозирующая способность, %
Число таких цепей	2	136	14	187	78	15	90,3
	1	87	63	544	55	38	75,3
Метил	2	119	31	279	80	13	80,6
	1	58	92	935	43	50	78,5
	0	23	127	340	11	82	85,0
Этил	1	124	26	326	80	13	85,0
	0	90	60	852	52	41	68,8
Наибольшее кольцо (по числу атомов углерода)	5	128	22	112	74	19	93,6
	4	114	36	116	66	27	93,6
Число связей —C=C—	2	140	10	23	79	14	94,6
	1	132	18	29	73	20	94,6
	0	123	27	145	66	27	88,2
Карбоильная группа	0	106	44	852	61	32	73,1
Кислородная связь	0	92	58	582	59	34	75,3

Продолжение табл. 4.3

	Обучающая выборка				Экзамнационная выборка		
	порог	отрицательная категория	положительная категория	число коррекций через обратную связь	отрицательная категория	положительная категория	прогнозирующая способность, %
Простой эфир	0	111	39	1005	75	18	76,3
Гетероатом в кольце	0	116	34	288	74	19	85,0
Амины	0	112	38	83	73	20	92,5
Спирты	0	129	21	122	81	12	89,2
Ароматические соединения	0	138	12	16	76	17	94,6
Нечетное число атомов водорода	0	104	46	194	69	24	86,0

углеводородов. Карбонильная группа предполагает наличие в соединении двойной связи углерод—кислород. Кислородная связь означает, что два атома углерода соединены друг с другом кислородным мостиком. «Простой эфир» и остальные категории трактуются в обычном понимании. В каждом случае весовой вектор подбирали таким образом, чтобы он давал положительный ответ для чисел выше порога и отрицательное скалярное произведение — для чисел меньших или равных порогу. Для всех категорий класса CHON* обучение прошло успешно. Прогнозирующая способность изменялась от 68,8 до 98,9% со средним, равным 88,0%. Следовательно, для соединений класса CHON тоже существует возможность успешного предсказания структурных параметров.

Не следует думать, что молекулярную структуру соединений всегда можно однозначно определить эмпирическим анализом масс-спектров низкого разрешения, но выявляемая при этом информация может принести большую пользу. Была проведена качественная оценка такой возможности. Из обучающей и экзаменационной выборки для углеводородов и соединений типа CHON отобрали по 10 случайных соединений. Масс-спектр каждого контрольного соединения классифицировали при помощи обученных весовых векторов соответствующего класса. Затем результаты подобной классификации передали одному из авторов без какой бы то ни было дополнительной информации с заданием вывести по мере возможности молекулярную формулу и определить структуру каждого контрольного соединения. В табл. 4.4 и 4.5 приведены результаты, показанные обучающейся машиной при подобной проверке, вместе с заключениями о сделанных выводах и их точности. В табл. 4.5 под рубрикой «Число неправильных классификаций» указано, сколько весовых векторов для каждого соединения дали неправильные ответы, т. е. число ошибочных прогнозов из 43 возможных для углеводородов. (Эти данные были включены в таблицу дополнительно после вывода формул и определения структур.)

Как видно из табл. 4.4, обучающая выборка для углеводородов позволила получить расчетным путем такую информацию, которая оказалась достаточной для вывода правильной молекулярной формулы во всех случаях. В пяти случаях (соединения 1, 3, 6, 9 и 10) удалось однозначно определить также структуру соединений. В двух случаях (2 и 4) ответы показывали, что эти соединения имеют структуру одного из двух изомеров; еще в двух случаях (7 и 8)

* Надо иметь в виду, что к соединениям типа CHON автор относит и соединения подтипов CHO и CHN. — *Прим. перев.*

Таблица 4.4

Соединения из обучающей выборки для углеводородов

Номер соединения	Молекулярная формула	Отношение $C:H = n:2n+2$	Число радикалов			Наибольшее кольцо	Число атомов углерода в боковой цепи	Число связей $C \equiv C$	Число атомов C, не связанных с H	C_6H_5	Число связей $C \equiv C$	Винильная группа $(CH_2=CH-)$
			Метил	Этил	n-пропил							
1	C_8H_{12}	$2n+2$	3	1	0	Нет	1	1	0	Нет	Нет	Нет
2	C_6H_{12}	$2n$	3	1	0	»	1	1	1	»	»	»
3	C_8H_8	$2n-2$	2	0	0	»	0	2	1	»	»	»
4	C_7H_{12}	$2n-2$	3	1	0	»	1	0	>1	»	Есть	»
5	$C_{10}H_8$	$<2n-6$	0	0	0	>6	2	>2	>1	»	Нет	»
6	C_8H_{18}	$2n+2$	4	1	1	Нет	1	0	1	»	»	»
7	C_8H_{16}	$2n$	2	0	0	6	2	0	0	»	»	»
8	C_8H_8	$<2n-6$	0	0	0	6	2	>2	2	Есть	»	»
9	C_9H_{18}	$2n$	2	0	0	5	2	0	0	Нет	»	»
10	C_9H_{10}	$<2n-6$	0	0	0	6	1	>2	0	Есть	»	Есть

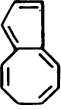
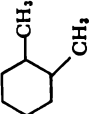


Молекулярная формула		Молекулярная структура	
Номер соединения	Предельная	предсказанная	действительная
	формула		
1	C_4H_{10}	C_4H_{12}	$\begin{array}{c} CH_3 \\ \\ CH_3-CH-CH_2-CH_3 \end{array}$
2	C_6H_{12}	C_6H_{12}	$CH_3-CH=CH-CH_2-CH_3 \text{ или } CH_3-CH=CH-CH_2-CH_3$
3	C_5H_8	C_5H_8	$\begin{array}{c} CH_3 \\ \\ CH_3-CH=CH-CH_3 \end{array}$
4	C_7H_{12}	C_7H_{12}	$CH_3-CH-CH_2-CH_2-CH_3 \text{ или } CH_3-CH-CH_2-CH_2-CH_3$
5	$C_{10}H_8$	$C_{10}H_8$	<p>2 кольца, 5 двойных связей</p> 
6	C_8H_{18}	C_8H_{18}	$\begin{array}{c} CH_3 \\ \\ CH_3-CH_2-CH_2-CH_2-CH_2-CH_2-CH_2-CH_3 \\ \\ CH_3 \end{array}$
7	C_8H_{16}	C_8H_{16}	<p>или изомеры</p> 
8	C_8H_8	C_8H_8	<p>или изомеры</p> 
9	C_6H_{10}	C_6H_{10}	$\begin{array}{c} CH_3 \\ \\ CH_3-CH-CH_2-CH_2-CH_2-CH_3 \\ \\ CH_3 \end{array}$
10	C_8H_{10}	C_8H_{10}	

Таблица 4.5

Соединения из экзаменационной выборки для углеводородов

Номер соединения	Число атомов угле- рода	Число атомов водо- рода	Отношение C:H	Метил	Этил	n-Пропил	Наибольшее кольцо	Число атомов угле- рода в боковой цепи	Число связей —C—C—	Число атомов C, не связанных с H	Бензольное кольцо	Тройная связь	Винильная группа (CH ₂ =CH—)	Число неправильных классификаций
1	5	10	2n	2	0	0	0	2	1	0	Нет	Нет	Нет	3
2	6	14	2n	3	0	0	0	1	1	0	»	»	Есть	2
3	6	12	2n	3	>1	0	0	2	1	<1	»	»	»	6
4	5	8	2n—2	2	0	0	0	1	0,2	0	»	»	Нет	4
5	10	20	2n	4	1	0	5	2	0	1	»	»	»	3
6	8	16	2n—6, 2n—2	2	<1	0	6	2	>2	>1	Есть	»	»	6
7	8	16	2n	1	1	1	5	2	0	0	Нет	»	»	2
8	8	14	2n—2	1	1	0	5	1	0	1	»	»	»	6
9	9	20	2n+2	4	1	0	0	2	0	1	Нет	Нет	Нет	2
10	9	18	2n	2	0	1	6	1	0	0	»	»	»	2

Номер соединения	Молекулярная формула		Молекулярная структура	
	предсказанная	действительная	предсказанная	действительная
1	C_5H_{10}			
2	C_6H_{12}			
3	C_6H_{12}			
4	C_6H_8			
5	$C_{10}H_{20}$	5-членное кольцо, $3C_3H_7$, $1C_2H_5$		
6	C_8H_{10}	$C_{10}H_{14}$ C_8H_5 , ксилол		
7	C_8H_{16}		или изомеры	
8	C_8H_{14}			
9	C_9H_{20}			
10	C_9H_{18}			

прогнозировалась структура одного из трех возможных изомеров и, наконец, лишь в одном случае (5) информация была недостаточной для того, чтобы свести определение структуры к выбору из немногих вариантов, хотя некоторые заключения о структуре соединений она все же позволила сделать. Проиллюстрируем ход рассуждений, позволивших сделать только что изложенные выводы, на примере первого соединения из обучающей выборки для углеводородов. Весовые векторы для чисел атомов углерода и водорода показали, что эти числа равны соответственно 5 и 12, что согласуется с отношением $C:H = 2n+2$. Кроме того, данное соединение нельзя отнести к ненасыщенным, поскольку двойных и тройных связей, винильных групп, бензольных и других колец не имеется. Таким образом, соединение имеет молекулярную формулу C_5H_{12} , т. е. может быть либо *n*-пентаном, либо изопентаном, либо же неопентаном. Наличие трех метильных радикалов означает, что соединение не может быть ни *n*-пентаном, ни неопентаном. Этот вывод подтверждается наличием в составе соединения одной этильной группы и одного атома углерода в боковой цепи. Наконец, отсутствие *n*-пропильного радикала тоже говорит о том, что соединение не может быть *n*-пентаном. Следовательно, исследуемое соединение — изопентан. Отметим большую избыточность наличной информации в данном случае. Подобная избыточность особенно полезна в случае составления прогнозов для соединений из экзаменационной выборки, когда приходится считаться с большей вероятностью ошибочных выводов.

Молекулярную формулу углеводородов из экзаменационной выборки удалось правильно предсказать во всех случаях, кроме одного (см. табл. 4.4, 4.5). Менее успешным, чем в случае обучающей выборки, было прогнозирование структуры соединений из экзаменационной выборки. Правильно предсказать удалось только своеобразную структуру одного соединения (2); предполагаемая структура другого соединения (7) представляла одну из двух возможных структур. (Изомеры, образованные ответвлениями от углеродных атомов кольца, не различались по другим рассматриваемым структурным параметрам и потому отдельно не рассматривались.) В остальных семи случаях прогнозы структуры пяти соединений (1, 3, 4, 8 и 9) были довольно близки к правильным в том смысле, что предсказанные структуры отличались от действительных только по положению двойной связи или по точке разветвления. В одном случае — для соединения 6 — молекулярная формула была предсказана неправильно, но в данном случае не было никаких шансов на составление верного прогноза. Однако полученные результаты

оказались полезными в том отношении, что позволили правильно предсказать каркас структуры (замещенное бензольное кольцо); заместители были определены неверно. Для соединения 5 результаты вычислений были настолько противоречивыми, что из них не удалось извлечь никакого смысла. Поэтому здесь не предпринималось попыток прогнозирования структуры соединения. Однако следует отметить, что даже подобные противоречивые данные полезны, поскольку предостерегают от попыток их использования для прогнозирования.

В оригинальной работе были проведены также аналогичные исследования спектров соединений типа CHON .

Пороговые логические элементы с ненулевым порогом [3]

Рассмотренные выше простые пороговые логические элементы осуществляют классификацию путем сравнения вычисленного значения скаляра s с нулем. Выбор ненулевого порога Z означает дальнейший шаг по пути обобщения. В этом случае скалярное произведение вектора образа \mathbf{X} на весовой вектор \mathbf{W} сравнивается с величиной Z . Если s превосходит Z , то считается, что образ принадлежит к одной категории; если s окажется меньше $-Z$, то образ считается принадлежащим к другой категории. Когда же s попадает в интервал $(-Z, Z)$, образ не классифицируется. Область между $-Z$ и Z называют мертвой (свободной) зоной. Этот способ классификации легко охарактеризовать при помощи двух гиперплоскостей, относя образы к одной категории, если они оказываются по одну сторону от обеих плоскостей, и к другой категории, если они находятся по другую сторону от плоскостей. Когда же образы попадают в промежуток между плоскостями, их не классифицируют.

Обучение с $Z \neq 0$ проводят во многих отношениях так же, как и в случае $Z = 0$. Обратная связь используется только для исправления ошибок, но под ошибками теперь понимают и неправильную классификацию, и неспособность классифицировать. В процессе обучения новый весовой вектор вычисляют из исходного по формуле:

$$\mathbf{W}' = \mathbf{W} \pm c\mathbf{X}, \quad (4.1)$$

где \mathbf{W}' — новый весовой вектор, \mathbf{W} — исходный весовой вектор, \mathbf{X} — вектор классифицируемого образа, c — поправка; знак выбирают в зависимости от характера ошибки. Поправку вычисляют

по соотношению

$$c = \frac{2}{\mathbf{X} \cdot \mathbf{X}} (\pm Z - s), \quad (4.2)$$

где $\mathbf{X} \cdot \mathbf{X}$ — квадрат длины вектора образа, s — величина скаляра, приведшая к неправильной классификации ($s = \mathbf{X} \cdot \mathbf{W}$); знак выбирается в зависимости от характера ошибки. Этот способ коррекции через обратную связь сдвигает решающую поверхность таким образом, что после коррекции точка образа лежит с «правильной» стороны от решающей поверхности и на таком же расстоянии от нее, на каком она была до коррекции, находясь с «неправильной» стороны. Можно ожидать, что при осмысленном выборе Z прогнозирующую способность удастся повысить по сравнению с нулевым порогом, поскольку здесь многомерный объем, в котором может находиться гиперплоскость, оставаясь «хорошей» решающей поверхностью, намного сужается. В случае ненулевого порога возможно также повышение надежности.

В табл. 4.6 приведены результаты обучения бинарных классификаторов образов с порогом и без него. Обучение проводилось на том же массиве данных, что и в случае табл. 3.1. Для каждой проверки из массива случайно выбирали по 300 масс-спектров. В результате трех процедур рандомизации были составлены три варианта обучающей и экзаменационной выборок, состав которых указан в нижней части табл. 4.6. Пороговые логические элементы с $Z = 50$ в каждом случае для достижения 100%-ного распознавания требуют больше коррекций, но каждый раз они обеспечивают повышение прогнозирующей способности. Такие пороговые логические элементы с $Z = 50$ способны правильно классифицировать 95,0, 95,8 и 96,0% полностью неизвестных образов из экзаменационной выборки (остальные данные табл. 4.6 обсуждаются ниже).

Цифры, характеризующие такую распознающую способность, вычисляли следующим образом. После обучения до полного распознавания объекты обучающей выборки вторично предъявляли классификатору, но на этот раз с искажением случайного вектора. Величина каждой компоненты случайного вектора подчинялась гауссову распределению. Данные относятся к указанным в таблице значениям относительного стандартного отклонения гауссова распределения. Например, $\sigma = 5\%$ означает, что приблизительно одна треть компонент каждого вектора образа искажалась более чем на $\pm 5\%$ первоначальной величины. Искажения вводили после логарифмического преобразования, так что частота ошибок была больше, чем в реальной лабораторной ситуации, где ошибки предшествуют

всем преобразованиям. Из табл. 4.6 видно, что при классификации с ненулевым порогом процент правильных распознаваний возрастает.

В табл. 4.7 приведены результаты исследования характеристик бинарных классификаторов образов в зависимости от величины порога. Как и следовало ожидать, число коррекций через обратную связь, необходимых для правильной классификации всех образов из обучающей выборки, тем больше, чем выше порог Z . Одновременно возрастают прогнозирующая способность и надежность. В последнем столбце указано, сколько из исходных 132 положений m/e , использованных при обучении, были исключены как неопределенные. По мере возрастания Z и сокращения объема, в который попадает решающая поверхность, число используемых классификатором образов положений m/e возрастает.

Если обучение проводится с порогом, его можно использовать и для составления прогнозов. Результаты такого исследования отражены в табл. 4.8. Во втором столбце приведены результаты классификации обученными весовыми векторами неизвестных объектов обучающей выборки, когда порог равен нулю. Затем объекты контрольной выборки классифицировали с указанными в таблице значениями порога Z . Таким образом, неклассифицированными оставались те объекты, для которых скалярное произведение попадало в интервал $(-Z, Z)$. Во всех случаях процент правильных предсказаний при ненулевом пороге оказался больше, чем при $Z=0$. Две серии данных о прогнозирующей способности были получены при помощи двух весовых векторов, обучение которых начиналось с разными исходными значениями. В этом случае абсолютное изменение прогнозирующей способности было незначительным, однако каждый раз с увеличением порога Z она становилась выше.

Другое исследование ненулевых порогов привело к несколько иной формулировке проблемы [4]. В процессе обучения новый весовой вектор вычисляют из исходного по формуле

$$\mathbf{W}' = \mathbf{W} + c\mathbf{X}, \quad (4.3)$$

в которой поправку c находят по формуле (4.2):

$$c = \frac{2(\pm Z - s)}{\mathbf{X} \cdot \mathbf{X}}. \quad (4.4)$$

Коррекцию через обратную связь осуществляют с учетом следующего дополнительного ограничения: весовой вектор должен всегда иметь единичную длину ($|\mathbf{W}|=1$). В данном случае поправку мож-

Таблица 4.7

Характеристики бинарных классификаторов образов в зависимости от величины порога

Z	Число коррекций через обратную связь		Прогнозирующая способность, %		Распознающая способность, %		Число исключенных положений m/ε
	фактически	среднее	фактическая	средняя	при $\sigma = 2\%$	при $\sigma = 5\%$	
0	135/118	126	96,3/94,0	95,3	98,5	96,4	31
25	156/150	153	95,3/96,0	95,7	99,3	97,5	32
50	239/184	206	96,0/96,0	96,0	99,6	98,6	27
75	275/203	239	96,0/96,3	96,1	99,7	98,9	22

Таблица 4.8

Повышение прогнозирующей способности бинарных классификаторов образов при использовании порогов

Z	Средняя доля предсказаний, %	Начальное приближение A		Начальное приближение B		Средняя прогнозирующая способность, %
		прогноз с порогом, %	число неклассифицированных спектров	прогноз с порогом, %	число неклассифицированных спектров	
25	95,7	96,3	3	96,3	4	96,3
50	96,0	97,0	5	96,3	7	96,7
75	96,1	97,6	8	96,3	9	96,9

но вычислить из квадратного уравнения:

$$\alpha c^2 + \beta c + \gamma = 0, \quad (4.5)$$

где

$$\begin{aligned} \alpha &= \mathbf{X} \cdot \mathbf{X} (4Z^2 - 4Zs + s^2 - \mathbf{X} \cdot \mathbf{X}), \\ \beta &= 2s (s^2 - 4Zs + 4Z^2 - \mathbf{X} \cdot \mathbf{X}), \\ \gamma &= 4Z^2 - 4Zs. \end{aligned} \quad (4.6)$$

Это квадратное уравнение имеет два решения, из которых в нашем случае только одно имеет смысл. Заметим, что при $Z = 0$ соотношения (4.6) принимают более простую форму.

Ниже излагаются результаты обучения весовых векторов на прежнем массиве из 630 масс-спектров низкого разрешения по программе их настройки в соответствии с соотношениями (4.6).

Введение положительного порога Z должно повышать прогнозирующую способность вектора \mathbf{W} для линейно разделимых ситуаций благодаря более оптимальной решающей поверхности. В табл. 4.9

Таблица 4.9

Сравнение результатов прогноза в случаях $Z = 0$ и $Z > 0$

Положительная категория	$Z_{\text{макс}}$	Доля верных предсказаний, %		
		отрицательная категория	положительная категория	Всего
Наличие кислорода и азота	0,0	77,2	91,8	86,4
	6,0	83,0	93,8	89,7
Наличие кислорода	0,0	93,3	80,7	89,7
	4,7	95,6	87,2	93,9
C:H = 1:2	0,0	82,1	77,9	80,6
	0,27	82,1	79,7	81,2
Больше 6 атомов углерода	0,0	95,7	90,0	92,4
	1,3	95,1	91,5	93,0

приведены результаты обучения классификации положительным порогом для задач, известных как линейно разрешимые. В этой таблице сопоставляются данные о прогнозирующей способности

после максимизации Z с соответствующими показателями при $Z=0$. В каждом случае обучающая выборка состояла из 300 случайно взятых образов, а оставшиеся 330 спектров рассматривались как контрольная выборка. Образы контрольной выборки классифицировали только по знаку скалярного произведения, так что любое изменение прогнозирующей способности было обусловлено иным положением решающей поверхности. Положительные категории перечислены в этой таблице, а отрицательные категории охватывали все те образы, которые не удовлетворяли рассматриваемому критерию. Во всех случаях Z стремилось к максимуму. Однако, как показали наблюдения, если прогнозирующая способность повышалась, то возможности такого повышения почти полностью исчерпывались при $Z=0,95 Z_{\text{макс}}$. Это лишало смысла продолжение итераций для значений $Z>0,95 Z_{\text{макс}}$, проведение которых было сопряжено с большими трудностями при расчетах и не давало почти никакого выигрыша в прогнозирующей способности. В табл. 4.10

Таблица 4.10

Прогнозирующая способность как функция Z

Классификация по кислороду и азоту		Классификация по кислороду	
Z	прогнозирующая способность, %	Z	прогнозирующая способность, %
0,0	86,4	0,0	89,7
4,0	89,1	2,0	91,5
5,0	90,0	3,0	93,6
6,0	89,7	4,0	93,6
6,125	89,1	4,5	93,9
6,25	89,1	4,7	93,9

приведены данные о возрастании прогнозирующей способности в зависимости от величины Z в наиболее выигрышных случаях. Надо отметить, что общая прогнозирующая способность возрастала во всех случаях — в одних больше, а в других меньше.

Способу классификации с ненулевым порогом присуще и другое преимущество. Вспомним о том, что величина скалярного произведения пропорциональна расстоянию до решающей плоскости, на котором находится неизвестный образ. Это дает нам критерий

достоверности в виде такого расстояния при классификации того или иного образа. В табл. 4.11 приведены наглядные примеры зависимости числа правильных предсказаний от величины $Z_{\text{макс}}$,

Таблица 4.11

**Доверительный уровень (доля верных предсказаний)
в зависимости от расстояния до решающей поверхности**

$s = W \cdot X $	Классификация по кислороду		Классификация по кислороду и азоту	
	число верных предсказаний/ число всех предсказаний	доля верных предсказаний, %	число верных предсказаний/ число всех предсказаний	доля верных предсказаний, %
$s > Z$	278/285	98	265/276	96
0,8—1,0	14/17	82	14/17	82
0,6—0,8	4/6	67	6/10	60
0,4—0,6	5/6	83	8/11	73
0,2—0,4	3/7	43	6/13	55
0,0—0,2	6/9	67	2/3	67
Итого для $s < Z$	32/45	71	36/54	67
Всего	310/330	94	301/330	90

т. е. от расстояния до решающей поверхности. Несмотря на значительный уровень шума в случае небольших чисел, общая тенденция очевидна. Когда расстояние образов до решающей поверхности превосходит Z , доверительный уровень намного выше, причем с уменьшением скалярного произведения он все более снижается. Следует обратить внимание на разницу между предсказаниями внутри разделяющей полосы и вне ее. Последние целесообразно сопоставить с результатами, даваемыми пороговыми логическими элементами (см. табл. 4.9).

Третье преимущество решающей поверхности с конечной толщиной выявляется при применении отрицательного порога Z к линейно неразделимым категориям, что позволяет использовать те же алгоритмы, что и для линейно разделимых категорий.

Для иллюстрации подобной методики обратимся к примеру с масс-спектрами низкого разрешения. Категории, которые были линейно разделимыми при использовании всех пиков, превышаю-

ших 0,5%, были преобразованы в неразделимые, причем в каждом спектре оставлялось всего шесть наиболее интенсивных пиков. Иначе говоря, массив образов состоял из 630 спектров, в каждом из которых имелось только по 6 наиболее интенсивных пиков.

Как ни странно, при таком использовании всего шести главных пиков большая часть задач оказывалась линейно разделимой и только в случае двух задач, указанных в табл. 4.12, пришлось

Таблица 4.12

**Сравнительные результаты прогнозирования при помощи
пороговых логических элементов и при $Z < 0$
для линейно неразделимых категорий**

	Доля верных предсказаний, %			
	подкласс $C_n H_{2n+2}$		подкласс $C_n H_{2n}$	
	$Z < 0$	$Z = 0$	$Z < 0$	$Z = 0$
Отрицательная категория	91,2	83,2	93,4	72,7
Положительная категория	72,4	55,3	92,2	84,5
Всего	88,0	76,8	92,8	77,1
Число неклассифицированных спектров	55	0	121	0

использовать отрицательные значения Z . Между прочим, эти задачи убедительно доказывают превосходство Z -метода над пороговыми логическими элементами. Во всех случаях наличия области, в которой решение не принимается, прогнозирующая способность намного возрастала. Особенно показательны в этом отношении результаты для случая $C:H = 1:2$, и прежде всего — разница данных о доле верных предсказаний для отрицательных категорий (93,4% при $Z < 0$ и 72,7% при $Z = 0$).

Система голосований [3]

Во всех рассмотренных до сих пор исследованиях для классификации образов использовался один пороговый логический элемент. Дальнейшим шагом по пути обобщения является переход к использованию двухуровневых пороговых логических элементов. При таком подходе образ одновременно предъявляют трем (пяти,

семи и т. д.) логическим элементам. Результаты, выданные пороговыми элементами первого уровня, поступают на пороговый элемент второго уровня, который действует по принципу «большинства голосов». Таким образом, целая группа классификаторов передает свое решение «собирателю голосов», который относит исходный образ к той категории, за которую «высказалось» большинство.

Простой и полезный метод обучения состоит в том, чтобы на каждом этапе обходиться минимальным необходимым числом коррекций через обратную связь. Если в процессе обучения дается неправильная классификация, то веса пороговых элементов, допускающих небольшие ошибки, подправляют по обычному уравнению обратной связи. Делать это нужно для как можно меньшего числа пороговых элементов, пока не будет обеспечена правильная классификация.

В табл. 4.6 были приведены результаты обучения по такой трехзвеньевой системе голосования на тех же выборках данных, что и при классификации образов другими методами. Эти результаты свидетельствуют о быстрой сходимости системы к высокой (100%) прогнозирующей способности. Однако такая система голосования очень чувствительна к изменениям состава обучающей выборки, что видно по доле верных распознаваний.

Можно построить систему классификации образов, в которой каждое звено будет иметь ненулевой порог. Обучение проводится как и раньше: для регистрации классификации скалярное произведение в каждом «голосующем» звене должно выходить из мертвой зоны. Собиратель голосов фиксирует окончательную классификацию только в том случае, когда с ней согласна большая часть звеньев. Надо полагать, что такая система классификации окажется более надежной, чем простые классифицирующие системы.

В табл. 4.6 были приведены результаты обучения системы голосования, состоящей из трех пороговых логических элементов с $Z = 50$, на тех же трех выборках данных, что и в других случаях. Число коррекций через обратную связь для этой машины больше, но остается на приемлемом уровне. Однако процент правильных классификаций для нее гораздо выше. Эта машина правильно классифицировала 95,7, 96,3 и 98,0% объектов контрольной выборки, составленной из неизвестных ей образов. Процент распознаваний для нее также очень высок: эта машина почти никогда не ошибается на образах рассматриваемой размерности.

Применение пороговых логических элементов для классификации электрохимических данных

Другим и совершенно отличным от прежних массивом химических данных, на котором испытывали простые пороговые логические элементы, является соеокупность СЭ-полярограмм [5, 6]. Сами эти данные и их предварительные преобразования подробно обсуждались в гл. 3. Для каждого образа было выделено по 123 признака. Программа обучения предусматривала следующие операции: построение двух весовых векторов со всеми исходными компонентами $+1$ в одном случае и -1 в другом; отбрасывание признаков, для которых компоненты двух обученных весовых векторов имели разные знаки; повторение процесса обучения.

Была составлена обучающая выборка полярографических кривых со следующими характеристиками: для однокомпонентных образов брали значения n , равные 1,00; 1,02; ... ; 2,98; 3,00 со случайным расположением пиков, что дало $3 \cdot 101 = 303$ возможных образа. Для двухкомпонентных образов брали значения n , равные 1,0; 1,4; 1,8; 2,2; 2,6; 3,0 с разделением пиков на 8, 10 и 12 мВ. Для отношений высот пиков были выбраны значения 20:1, 10:1, 1:1, 1:10 и 1:20. Из 540 возможных двухкомпонентных образов было отобрано 533. При взятии отсчетов точность измерения E_p была равна 0,1 мВ, и все значения E_p оценивали с такой точностью в интервале до 2 мВ. Замеренные высоты пиков нормировали на единицу. Затем для каждой СЭ-полярограммы выделяли по 133 признака, подлежащих включению в обучающую выборку.

Контрольную выборку СЭ-полярограмм составляли следующим образом. Для однокомпонентных образов брали 100 значений n : 1,01; 1,03; ... ; 2,97; 2,99. 200 значений n в диапазоне 1,000—2,999 были выбраны случайным образом. Размещение пиков тоже рандомизировали. Для двухкомпонентных образов значения n случайно выбирали из интервала 1,000—3,000; а промежутки между пиками — из диапазона 8,000—12,000 мВ; для 300 пиков отношение высот случайно выбирали в интервале между 1:1 и 1:20, а для 300 пиков — между 20:1 и 1:1. В результате были построены 300 однокомпонентных и 600 двухкомпонентных образов; их распределяли так же, как и для обучающей выборки. Для каждого образа имелось по 133 признака.

В табл. 4.13 приведены результаты обучения и данные о прогнозирующей способности, полученные при классификации по этим выборкам СЭ-полярограмм. В таблице отражены рассмотренные раньше пять этапов отбора признаков и процесса обучения. Ответ

Таблица 4.13
Распознавание образов обучающей выборки и предсказание образов по СЭ-поляrogramмам

Номер этапа	Число признаков	Число итераций	Обучение				Предсказание			
			одна компонента		две компо- ненты		одна компонента		две компоненты	
			неопределенность, %	точность, %	неопределенность, %	точность, %	неопределенность, %	точность, %	неопределенность, %	точность, %
1	133	50	19,5	98,8	10,1	94,1	25,7	97,8	10,0	89,1
2	75	100	13,2	95,4	8,8	96,1	15,3	94,1	8,8	91,6
3	66	50	14,9	97,7	6,8	95,8	16,7	96,0	6,5	90,5
4	59	50	12,5	98,9	8,1	96,3	13,7	96,9	4,3	89,0
5	57	5	5,6	90,9	3,6	96,9	11,0	92,1	2,2	89,9
		10	6,6	89,7	5,1	87,1				
		15	9,6	94,5	5,6	91,8				
		20	8,9	96,7	5,1	92,5	11,3	97,7	5,2	88,6
		25	7,6	95,0	3,9	93,4	8,3	94,6	3,0	88,1
		30	7,3	96,8	3,0	93,6	5,0	96,1	1,5	88,3
		35	7,6	97,5	3,9	94,5	6,7	98,2	2,3	88,9
		40	9,6	97,4	4,9	95,3	9,0	97,8	2,3	89,2
		45	6,3	96,5	4,7	95,9	10,3	97,4	2,2	89,2
		50	5,0	92,0	3,0	96,3	6,7	92,9	2,0	89,8
		75	6,3	93,0	3,0	97,1	10,7	92,2	1,7	90,0
		100	5,0	91,7	2,1	96,9	6,3	87,2	2,0	90,6
		125	3,3	91,5	1,7	96,4	5,3	88,0	1,7	89,8
		150	7,3	93,2	2,4	96,1	8,0	88,4	1,7	90,0
			7,3	94,3	2,1	96,0	7,0	88,9	2,3	90,4

считается неопределенным в том случае, когда два построенных весовых вектора при обучении относят объект к противоположным категориям. Ошибкой считается неправильная классификация образа обоими весовыми векторами. Точность выражается как доля (%) правильно классифицированных образов по отношению ко всем проведенным классификациям.

Как видно из табл. 4.13, два весовых вектора способны довольно хорошо классифицировать одно- и двухкомпонентные СЭ-полярограммы. Распознавание образов обучающей выборки не страдало от сокращения 133 исходных признаков до 57. При этом точность составляла 96%, а доля неопределенных ответов — 5 — 6%. Показатели прогнозирующей способности хуже, но следуют той же тенденции.

В последующих исследованиях проводилось изучение возможностей дальнейшего сокращения признаков. В связи с этим уместно отметить только то, что пороговые логические элементы способны обнаружить наличие дублирующих СЭ-полярограмм при довольно широко изменяющихся условиях, когда визуальная расшифровка не всегда возможна.

Итерационное обучение по методу наименьших квадратов

Разработан другой полезный для химических приложений способ построения весовых векторов по методу наименьших квадратов [7]. Образы аппроксимируют при помощи невырожденной функции с линейными параметрами; существуют разные способы такой аппроксимации. Авторами была использована линеаризация — разложение в ряд Тейлора [8]. Этот метод предполагает использование результатов линеаризации по методу наименьших квадратов последовательными этапами.

В качестве нелинейной функции был выбран гиперболический тангенс, поскольку он хорошо подходит для разбиения образов на два класса. Свойства этой функции выгодно отличают такой вариант от обычного метода наименьших квадратов для нескольких категорий, характеризующегося соотношением

$$R = \sum_{i=1}^N (s_i - Y_i)^2, \quad (4.7)$$

где Y_i — правильный численный ответ, s_i — скалярное произведение, i — индекс i -го образа и N — число образов в обучающей

выборке. Обычный метод наименьших квадратов для нескольких категорий также пробовали применять для интерпретации химических данных [9].

Величина Y_i считается равной $+1$, если образ принадлежит первой категории, и -1 , если образ принадлежит второй категории. Затем необходимо найти такой весовой вектор \mathbf{W} , чтобы разделяющая функция $g(\mathbf{X}_i)$ приобретала положительное значение, если ответ равен $+1$, и отрицательное, когда $Y_i = -1$. Компоненты вектора \mathbf{W} определяют обычной линейризацией по методу наименьших квадратов.

В качестве функции $F(s_i)$, при помощи которой аппроксимируются образы, был выбран гиперболический тангенс $\text{th } s_i$, где $s_i = \mathbf{W} \cdot \mathbf{X}_i$ скалярное произведение для i -го образа. Выбор гиперболического тангенса обусловлен тем, что он положителен при положительных значениях аргумента и отрицателен при отрицательных значениях. Таким образом, вся процедура эквивалентна минимизации числа несовпадений знаков для $F(s_i)$ и Y_i . Тогда в соответствии с методом наименьших квадратов минимизируемая функция записывается в виде

$$Q = \sum_{i=1}^N [Y_i - F(s_i)]^2. \quad (4.8)$$

Необходимо найти такой весовой вектор \mathbf{W} , при котором функция Q становится минимальной. Как и раньше, N — число образов в обучающей выборке.

Чтобы начать итерационную процедуру с величиной $s_i^0 = \mathbf{W}^0 \cdot \mathbf{X}_i$ в качестве начального приближения, функцию $\text{th } s_i$ разлагают в ряд Тейлора до члена с первой производной. Такое разложение достаточно точно аппроксимирует функцию. Таким образом, имеем

$$\text{th } s_i = \text{th } s_i^0 + \sum_{j=1}^{d+1} \left. \frac{\partial \text{th } s_i}{\partial w_j} \right|_0 dw_j, \quad (4.9)$$

или по правилу дифференцирования сложных функций:

$$\text{th } s_i = \text{th } s_i^0 + \sum_{j=1}^{d+1} \text{sch}^2 s_i^0 x_{ij} dw_j, \quad (4.10)$$

где i соответствует i -му образу, а j — j -й компоненте. Тогда минимизируемая функция принимает форму

$$Q = \sum_{i=1}^N \left(Y_i - \text{th } s_i^0 - \sum_{j=1}^{d+1} \text{sch}^2 s_i^0 x_{ij} d\omega_j \right)^2. \quad (4.11)$$

Минимум на каждом шаге итерации достигается для всех тех весовых векторов $d\mathbf{W} = d\omega_1, \dots, d\omega_{d+1}$, для которых выполняются условия экстремума:

$$\frac{\partial Q}{\partial \omega_k} = 0 \quad \text{для } k = 1, \dots, d+1. \quad (4.12)$$

Отсюда получаем

$$0 = -2 \sum_{i=1}^N \left(Y_i - \text{th } s_i^0 - \sum_{j=1}^{d+1} \text{sch}^2 s_i^0 x_{ij} d\omega_j \right) \text{sch}^2 s_i^0 x_{ik}. \quad (4.13)$$

Последнее равенство можно записать в виде следующей системы линейных уравнений:

$$\sum_{i=1}^N (Y_i - \text{th } s_i^0) \text{sch}^2 s_i^0 x_{ik} = \sum_{j=1}^{d+1} \sum_{i=1}^N \text{sch}^4 s_i^0 x_{ij} x_{ik} d\omega_j. \quad (4.14)$$

В матричных обозначениях эти уравнения принимают вид

$$A d\mathbf{W} = b, \quad (4.15)$$

где

$$a_{jk} = \sum_{i=1}^N \text{sch}^4 s_i^0 x_{ij} x_{ik} \quad (4.16)$$

и

$$b_k = \sum_{i=1}^N (Y_i - \text{th } s_i^0) \text{sch}^2 s_i^0 x_{ik}. \quad (4.17)$$

Вектор $d\mathbf{W}$ определяют как решение системы

$$A d\mathbf{W} = b. \quad (4.18)$$

Матрица \mathbf{A} — действительная, симметричная, положительно определенная, следовательно, она не вырождена, так что система имеет единственное решение. Его можно найти любым методом решения системы линейных уравнений — прямым или итерационным.

Поскольку $F(s_i)$ зависит от значения \mathbf{W} и \mathbf{W} — искомое решение, первый вектор \mathbf{W}' , полученный из \mathbf{W}^0 , не может быть точным. Согласно принципу линеаризации, принимают, что $\mathbf{W}^{(1)} = \mathbf{W}^0 + d\mathbf{W}^{(1)}$, и продолжают итерации. Процедуру повторяют до тех пор, пока величина $\mathbf{W}^{(l+1)} = \mathbf{W}^{(l)} + d\mathbf{W}^{(l+1)}$ не будет удовлетворять условию

$$|d\mathbf{W}^{(l+1)}| / |\mathbf{W}^{(l+1)}| < \varepsilon. \quad (4.19)$$

Иначе говоря, отношение нормы $d\mathbf{W}^{(l+1)}$ к норме $\mathbf{W}^{(l+1)}$ должно быть меньше некоторой произвольно выбранной малой величины ε . Когда это условие выполняется, итерации прекращают, поскольку дальнейшие вычисления не дают заметного выигрыша.

В рассматриваемом исследовании область задания F ограничена замкнутым интервалом $(-2,5, +2,5)$; при этом значения функции F оказывались в интервале $(-0,987, +0,987)$. Начальные значения весового вектора выбирали таким образом, чтобы удовлетворить условию $|s_i| < 2,5$. Когда это требование выполняется, скорость сходимости возрастает, поскольку в этом случае значения компонент мало изменяются и ошибки при вычислении \mathbf{W} незначительны.

Для решения системы линейных уравнений существует много прямых методов. В рассматриваемом исследовании исключение переменных осуществляли по методу Гаусса — Жордана с полной привязкой. Подпрограмма считала обратную матрицу, определитель и вектор решения.

Массив данных для этого исследования составляли 450 масс-спектров низкого разрешения из уже упоминавшихся таблиц Американского нефтяного института. Каждый спектр состоял из набора интенсивностей, перечисленных в порядке возрастания значений m/e . Минимальная учитываемая интенсивность в каждом спектре составляла 0,01% максимальной. По всем 450 спектрам можно было отобрать 132 положения m/e , имеющие не менее чем по 10 пиков. Поэтому 132 положения образует верхний предел размерности векторов образов d . Все спектры принадлежали органическим соединениям с общей молекулярной формулой $C_{1-10}H_{1-22}O_{0-4}N_{0-2}$. Исходные данные пришлось нормировать, поскольку, как показала практика, всякое уменьшение исходного отношения максимальной интенсивности к минимальной, равного 10 000 (100,0/0,01), намного ускоряло сходимость. Поэтому во всех случаях интенсивности нормировали по формуле

$$I' = 10 \lg (I \cdot 1000), \quad (4.20)$$

где I — первоначальная интенсивность и I' — нормированная интенсивность (I' соответствует x_{ij} , где i — номер спектра, а j — ин-

декс положения m/e). Для уменьшения объема необходимой машинной памяти значения I' округляли до ближайшего целого числа. Нормированные интенсивности находились в интервале 10—50, так что отношение максимальной интенсивности к минимальной было равно 5.

Из всего массива данных были составлены две выборки — обучающая и контрольная. Обучающая выборка обычно состояла из 150 выбранных случайно масс-спектров, которые использовались для построения весовых векторов. Контрольную выборку составляли из некоторых или же из всех оставшихся образов. На такой выборке проверяли способность созданного классификатора распознавать неизвестные образы.

Итерационным методом наименьших квадратов определяли присутствие кислорода в составе органических соединений небольшого молекулярного веса. При этом был проведен отбор признаков, в результате чего из 132 исходных признаков остался 31 признак. Доля верных предсказаний составила при полном распознавании 93,9%.

В этой задаче Y_i считали равным $+1$, если i -й спектр свидетельствовал о присутствии кислорода, и -1 в противном случае. Начальные компоненты весовых векторов брали по такому правилу: при $\omega_j = p$ величину ω_{j+1} считали равной $-p$. Значение ω_1 было равно либо $+0,01$, либо $-0,01$. Эти начальные значения обеспечивали попадание скалярного произведения в диапазон $(-2,5, +2,5)$. Как оказалось, минимизация расстояния между кластерами ускоряет сходимость.

Таблица 4.14

Обнаружение присутствия кислорода

Число положений m/e	Состав обучающей выборки		Доля распознаваний, %	Состав контрольной выборки		Доля верных предсказаний, %
	+	—		+	—	
31	42	108	99,3	26	74	98
	39	111	100,0	81	219	96
	43	107	98,7	77	223	98
22	43	107	99,3	77	223	98
18	43	107	99,3	77	223	98

В табл. 4.14 приведены результаты определения присутствия кислорода как по 31 первоначальному пику, так и после дальнейшего отбора признаков. При почти полном распознавании прогнозирующая способность составляла 98% независимо от числа используемых признаков.

Другой пробной задачей было определение присутствия или отсутствия атомов азота в составе органических соединений с небольшим молекулярным весом. Число признаков сократили от 132 до 43. Из этих 43 признаков оставили 31 признак с меньшими значениями отношения m/e . Как и в случае кислорода, величину Y_i считали равной +1, если спектр отражал содержание азота, и —1 в противном случае. Начальные значения весовых векторов выбирали в прежнем порядке.

В табл. 4.15 приведены результаты определения присутствия азота как по 31 исходному пику, так и после дальнейшего отбора признаков. И здесь не наблюдалось снижения распознающей и прогнозирующей способностей. Доля верных предсказаний составила 96,7% для 31 признака и 98% для 21 признака.

Таблица 4.15

Обнаружение присутствия азота

Число положений m/e	Состав обучающей выборки		Доля распознавания, %	Состав контрольной выборки		Доля верных предсказаний, %
	+	—		+	—	
31	12	138	99,3	20	280	95
	8	142	100,0	24	276	97
	9	141	99,7	23	277	94,7
	11	139	99,3	21	279	96,7
21	8	142	98,7	24	276	96
	17	133	98,7	15	285	97
	20	130	98,0	12	288	97
14	12	138	98,7	20	280	97

Кусочно линейные пороговые логические элементы

Другой метод использования нескольких пороговых элементов для принятия бинарного решения изложен в работе [10]. В этом исследовании набор одновременно действующих линейных пороговых элементов образовывал кусочно линейный классификатор.

Ниже изложено математическое описание подобной процедуры.

Выбирают число категорий R в зависимости от характера искомой информации, содержащейся в исходных данных, например

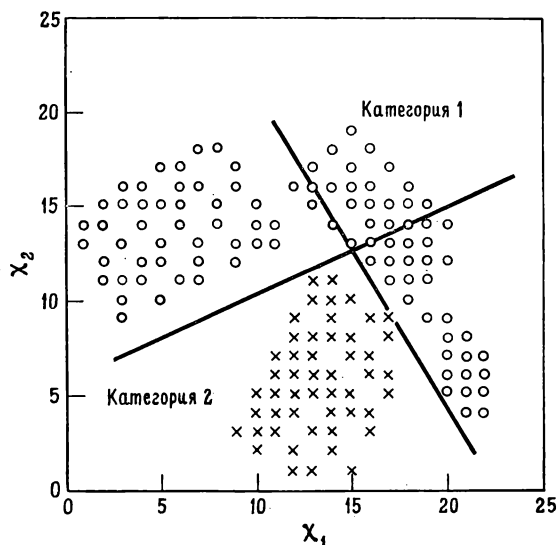


Рис. 4.1. Двумерная классификационная задача ($R=2$, $L_1=3$, $L_2=1$).

информации о параметрах молекулярной структуры, и принимают, что в пространстве образов объекты одной категории располагаются поблизости друг от друга. Фактическое распределение объектов по каждой категории может быть довольно сложным и содержать ряд локальных максимумов, соответствующих подкатегориям L_i , где $i=1, 2, \dots, R$. Простой двумерный пример [без $(d+1)$ -го направления] иллюстрируется на рис. 4.1. Здесь $R=2$, $L_1=3$, $L_2=1$.

Тогда задача классификации сводится к построению совокупности разделяющих (в данном случае кусочно линейных) функций s_i , описывающих d -мерные гиперплоскости, которые отделяют объекты одной категории от объектов всех других категорий. Лучшие всего изучены кусочно линейные разделяющие функции вида

$$s_i = \max_{j=1, \dots, L_i} s_i^{(j)} \quad \text{для } i = 1, \dots, R, \quad (4.21)$$

где $s_i^{(j)}$ — вспомогательные разделяющие функции, определяемые соотношением

$$s_i^{(j)} = f_i^{(j)}(\mathbf{X}) = w_{i1}^{(j)} x_1 + w_{i2}^{(j)} x_2 + \dots + w_{id}^{(j)} x_d + w_{id+1}^{(j)} x_{d+1}. \quad (4.22)$$

При такой системе классификации образ относится к той категории k , для которой s_k превосходит по величине все разделяющие функции s_i ($i=1, \dots, R$).

Итерационная процедура обучения с исправлением ошибок через обратную связь подбирает индивидуальные веса в векторах $\mathbf{W}_i^{(j)}$ до тех пор, пока обучающаяся машина не начнет классифицировать все образы правильно. Допустим, что классификатору предъявлен образ, принадлежащий k -й категории, и что наибольшую величину из всех разделяющих функций имеет s_l . Тогда коррекция через обратную связь сводится к определению положительной поправки c из соотношений

$$\begin{aligned} \mathbf{W}'_k &= \mathbf{W}_k + c\mathbf{X}, \\ \mathbf{W}'_l &= \mathbf{W}_l - c\mathbf{X}. \end{aligned} \quad (4.23)$$

Существует ряд правил выбора c . В рассматриваемом исследовании поправку выбирали по одному из вариантов правила частичной коррекции. Правило частичной коррекции предполагает, что поправку c всегда выбирают с таким расчетом, чтобы решающая поверхность, определяемая весовыми векторами \mathbf{W}_k и \mathbf{W}_l , сместилась на заданную долю λ своего обычного расстояния до точки образа \mathbf{X} . Так, при $\lambda=2$ новая решающая поверхность, определяемая векторами \mathbf{W}'_k и \mathbf{W}'_l , лежит на прежнем расстоянии до точки образа \mathbf{X} , но с другой стороны от нее. В случаях $\lambda>2$ возникают трудности, связанные с вариациями длины весовых векторов. Поскольку от длины весовых векторов зависит величина $s_i^{(j)}$, решение смещается в сторону векторов большей длины. Подобного смещения можно избежать нормированием длины весовых векторов на единицу. Нормирование, естественно, налагает дополнительные ограничения. Таким образом, соотношения (4.23) преобразуются к виду

$$\mathbf{W}'_k = \frac{\mathbf{W}_k + c_k \mathbf{X}}{|\mathbf{W}_k + c_k \mathbf{X}|}, \quad \mathbf{W}'_l = \frac{\mathbf{W}_l - c_l \mathbf{X}}{|\mathbf{W}_l - c_l \mathbf{X}|}. \quad (4.24)$$

Подстановка этих соотношений в исходные уравнения приводит к квадратным уравнениям относительно c_k и c_l следующего вида:

$$c_k^2 [(\mathbf{X}\mathbf{X})^2 - \mathbf{X}\mathbf{X}s_l^2] + c_k [2s_k (\mathbf{X}\mathbf{X} - s_l^2)] + s_k^2 - s_l^2 = 0, \quad (4.25)$$

$$c_k^2 [(\mathbf{X}\mathbf{X})^2 - \mathbf{X}\mathbf{X}s_k^2] + c_l [2s_l (\mathbf{X}\mathbf{X} - s_k^2)] + s_l^2 - s_k^2 = 0.$$

При каждой коррекции через обратную связь значения c_k и c_l следует вычислять из этих уравнений.

Обычно при формулировке задачи классификации известно число категорий R , но, как правило, неизвестно, сколько существует подкатегорий L_i . Удобный метод преодоления этой неопределенности состоит в том, чтобы до начала обучения приписать каждой категории адекватное, но фиксированное заранее число весовых векторов. Хотя такая процедура и позволяет отыскать то или иное решение, она чревата возможностью лишних вычислений как при обучении, так и при последующей классификации, если число векторов превосходит действительно необходимое. Избыточное число весовых векторов может привести также к излишне точной «подгонке» распределения данных в обучающей выборке, что ухудшает классификацию неизвестных образов.

При разработке рассматриваемого кусочно линейного классификатора преследовалась цель создания простого средства для внутреннего генерирования новых весовых векторов по мере необходимости непосредственно в процессе обучения. Авторы исследования надеялись тем самым создать довольно простое классифицирующее устройство, уровень сложности которого обеспечивал бы только получение нужного решения. Любой классификатор заданной сложности, обучение которого производится по методу исправления ошибок через обратную связь, обнаруживает «колеблющееся» поведение, когда перед ним ставят неразрешимую задачу. Например, классификатор, использующий единственную линейную разделяющую функцию, никогда не сможет решить задачу, представленную на рис. 4.1. В этом случае положение единственной решающей поверхности долго испытывало бы значительные колебания. Естественно поэтому искать пути обнаружения подобных колебаний, свидетельствующих о необходимости усложнения решающей поверхности.

В предлагаемом здесь методе используется периодическая оценка функции по следующему критерию:

$$\sum_{m=1}^M \frac{s_k - s_l}{M}, \quad (4.26)$$

где M — число всех образов в обучающей выборке, k — номер категории для m -го образа и s_l — наибольшая разделяющая функция среди множества функций s_i ($i \neq k$). Так, если классифи-

кация образа m произведена правильно, он вносит в сумму положительный вклад, если неправильно — отрицательный. Мы не можем строго обосновать законность использования этого критерия и, в частности, вскрыть, как должна изменяться его абсолютная величина, однако эмпирически установлено, что относительные изменения этой величины отражают колебания решающей поверхности, и поэтому должны быть проанализированы.

Практически оценивают функцию в следующей форме:

$$P_t = \frac{1}{\tau} \left(P_{t-1} + \sum_{m=1}^M \frac{s_k - s_l}{M} \right). \quad (4.27)$$

Поэтому функция P представляет собой „текущий взвешенный интеграл“ прошлых значений суммы с постоянной времени τ . Функции такого общего вида свойственно сглаживающее действие, поскольку вклад начальных колебаний, зависящий от постоянной времени, убывает во времени.

Решение о необходимости введения новой вспомогательной разделяющей функции принимают, руководствуясь следующим простым критерием: такую функцию вводят, если $P_t < \alpha P_{t-1}$, где постоянная α удовлетворяет условию $0 < \alpha \leq 1$. Когда же $P_t \geq \alpha P_{t-1}$, функцию не вводят. Параметры τ и α определяют чувствительность функции P к изменениям положения решающей поверхности; их подбирают экспериментально.

На рис. 4.2 иллюстрируется поведение функции P в процессе обучения в случае двумерной задачи, показанной на рис. 4.1. Задачу классификаций образов начинают с одной решающей гиперплоскостью, но поскольку одной такой поверхности недостаточно, значение P в ходе четвертого опыта снижается (при $t = 4$). Для этой точки $P_4 < P_3$. Поэтому вводят новую вспомогательную решающую функцию, обеспечивающую сходимость процесса обучения. Штриховыми линиями на рис. 4.2 показаны результаты при отказе от введения такой функции.

Рассматриваемое исследование проводилось на массиве данных, заимствованных из уже упоминавшихся таблиц Американского нефтяного института. Это были масс-спектры низкого разрешения для 387 углеводов с общей формулой $C_{1-10}H_{4-22}$. Каждый спектр состоял из последовательной совокупности пар чисел, характеризующих номинальную массу фрагмента и интенсивность пика. Пики с интенсивностью менее 1% не рассматривались. Каждый спектр в среднем имел приблизительно по 35 пиков.

В общем случае 200 случайно отобранных спектров включали в обучающую выборку, а остальные 187 использовали для проверки прогнозирующей способности обученного классификатора.

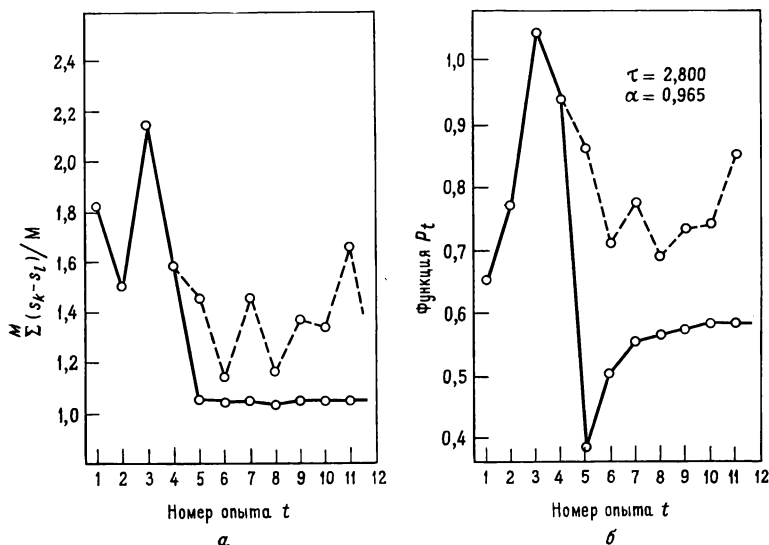


Рис. 4.2. Графики изменения суммы $\sum (s_k - s_l) / M$ (а) и функции P_t (б) в зависимости от порядкового номера опыта в классификационной задаче, показанной на рис. 4.1.

Интересная возможность проверить кусочно линейный классификатор и его способность усложняться по мере необходимости представилась в задаче обнаружения двойной связи $C=C$, которая является важной структурной особенностью. В молекуле того или иного соединения двойная связь может реализоваться несколькими интересными способами. Простой бинарный классификатор не дает хорошей сходимости при обучении в этой задаче даже после нескольких тысяч коррекций через обратную связь, а прогнозирующая способность результирующего весового вектора составляет $\sim 75\%$. Результаты обучения кусочно линейного классификатора приведены в табл. 4.16, где указаны значения τ и α , число классификаций в процессе обучения, число коррекций, число понадобившихся весовых векторов и, наконец, прогнозирующая способность обученного классификатора. Приведенные в этой таблице данные

Таблица 4.16

**Характеристики самогенерирующего кусочно линейного классификатора
в задаче о двойной связи углерод — углерод**

Номер опыта	τ	α	Число классификаций	Число коррекций	Число весовых векторов	Прогнозирующая способность, %			
						диапазон изменения	средняя	категория 1	категория 2
1	2,80	0,985	6925	930	3—4	78,6—86,1	81,9	82,3	81,5
2	3,10	0,985	6680	892	4	78,1—85,0	81,0	82,5	79,6
3	3,50	0,985	7623	1085	3—5	77,5—86,6	81,4	83,1	79,9
4	4,50	0,985	6159	914	5	79,1—86,6	81,7	84,8	79,2
5	2,00	0,990	8898	1206	6—10	78,6—82,9	81,5	84,1	78,9

по каждому опыту являются усредненными по пяти отдельным прогонам, которые отличались друг от друга начальными значениями весов и порядком предъявления классификатору образов обучающей выборки. При каждом прогоне достигалась сходимость, а число классификаций можно считать примерно пропорциональным длительности процесса обучения.

В каждом опыте к классификации приступали лишь с двумя весовыми векторами, пополняя в дальнейшем их число по мере необходимости. Очевидна зависимость конечного числа построенных векторов от постоянной времени τ . Чем больше τ , тем в большей степени функция P подвержена колебаниям и тем больше число весовых векторов. (Кажущееся большим число векторов в пятом опыте объясняется одновременным добавлением пары векторов по одному для каждой категории.) Прогнозирующая способность изменялась приблизительно от 78 до 87%, что на 3—12% лучше, чем для бинарного классификатора.

В оригинальном исследовании в программу обучения были введены некоторые усовершенствования, однако принципиальный механизм автоматического построения новых векторов изменять не понадобилось.

КЛАССИФИКАЦИЯ ОБРАЗОВ НА НЕСКОЛЬКО КАТЕГОРИЙ

До сих пор рассматривалась только бинарная классификация. Теперь остановимся на задаче разбиения на несколько категорий, которая очень важна в химии, поскольку многие ее вопросы имеют

несколько ответов. Таков, например, вопрос о числе функциональных групп в молекуле. Один из способов осуществления множественной классификации основан на использовании набора бинарных классификаторов, расположенных в том или ином порядке. Эти вопросы обсуждаются в последующих разделах настоящей главы.

Дерево бинарных классификаторов образов

На рис. 4.3 изображено дерево (ветвящаяся схема) бинарных классификаторов, построенное для вывода молекулярных формул [11]. Для обеспечения правильности разбиений на два класса ин-

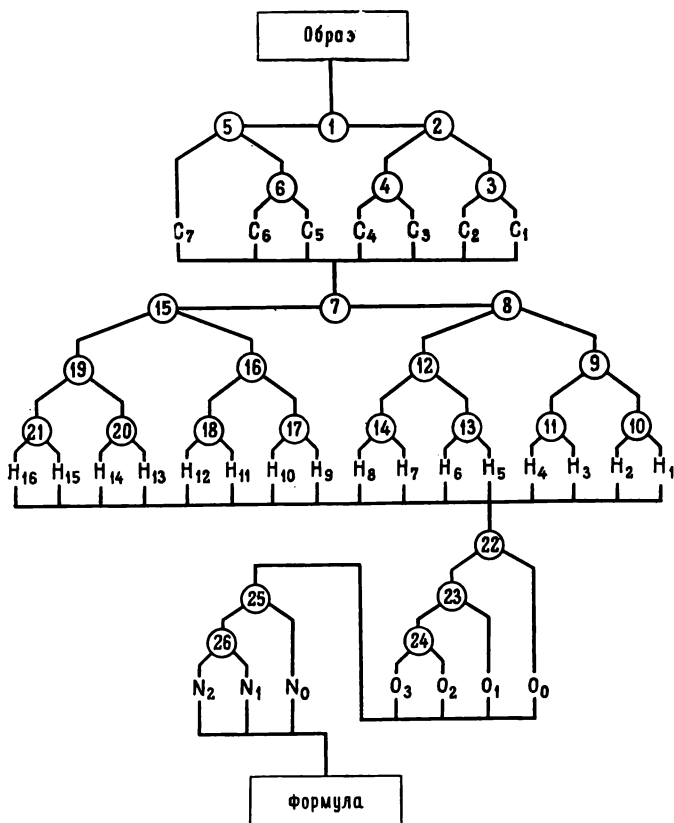


Рис. 4.3. Дерево бинарных классификаторов.

дексов формул для масс-спектров 346 соединений с общей формулой $C_{1-7}H_{1-16}O_{0-3}N_{0-2}$ пришлось построить 26 весовых векторов. При помощи этих весовых векторов (каждый из них той же размерности, что и исходный вектор образа) дерево бинарных классификаций позволило рассчитать молекулярные формулы. Это означает более чем 10-кратную экономию памяти, поскольку вместо 346 масс-спектров требовалось запомнить только 26 векторов. Отсюда вовсе не следует, что весовые векторы охватили всю информацию, содержащуюся в масс-спектрах. Речь идет только об информации, нужной для выведения молекулярных формул. Иначе говоря, задав конкретный вопрос, можно сократить объем хранимой информации. Таким образом, молекулярную формулу выгоднее „рассчитать“ вычислениями скалярных произведений, нежели искать ее в библиотечных данных. (Для расчета требуется ~ 50 мс на машинах второго поколения и 30 мин на настольной клавишной машине.)

Другая работа по изучению возможности классификации на несколько категорий тоже проводилась на масс-спектрах низкого разрешения [12]. Массив исходных данных состоял из 600 масс-спектров низкого разрешения, заимствованных из таблиц Американского нефтяного института в записи на магнитной ленте. Это были спектры соединений с молекулярной формулой $C_{3-10}H_{2-22}O_{0-4}N_{0-2}$. Их случайным образом разделили на 200 спектров обучающей выборки и 400 спектров контрольной выборки. Второй массив данных состоял из 600 спектров только углеводородов. Пять углеводородов с тремя атомами углерода пришлось исключить. Оставшиеся 372 соединения C_4-C_{10} были случайным образом разделены на обучающую выборку из 200 соединений и контрольную из 172 соединений. При всех расчетах использовались одни и те же обучающие и контрольные выборки и поэтому результаты, даваемые разными классификационными схемами, можно было легко сравнивать.

Число существенных положений m/e составляло 132, так что каждый спектр был представлен 132-мерным вектором образа. Исходные интенсивности пиков, нормированные по отношению к максимальному пику спектра, находились в диапазоне 0,01—99,99. Чтобы привести все спектры к единой шкале, интенсивности пришлось нормировать еще раз отношением к полному ионному току или суммарной интенсивности для каждого спектра. Последующее логарифмическое преобразование перевело все интенсивности в диапазон 10—59.

В выборке для углеводородов 3,5% спектров не имели основного пика (интенсивность не превышала 0,001% полного ионного тока).

Еще 8,6% спектров имели основной пик с интенсивностью от 0,001 до 0,1%, а для 12,1% спектров интенсивность основного пика лежала в диапазоне 0,1—0,5%. Поскольку значительная доля спект-

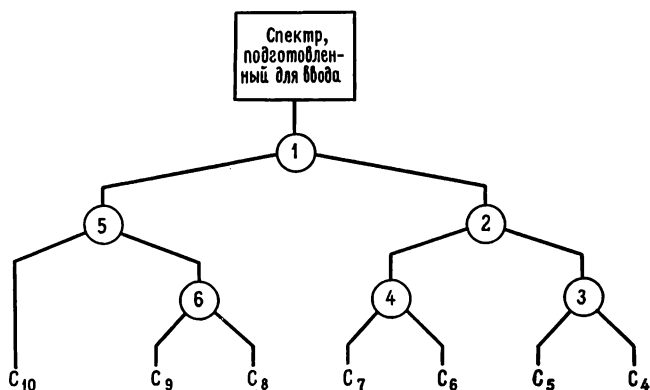


Рис. 4.4. Дерево бинарных классификаторов образов.

ров не имела основного пика, отличимого от шума, определение числа атомов углерода в молекулах углеводов было нелегкой задачей.

Бинарные классификаторы образов были расположены по ветвящейся схеме. Классификацию осуществляли по выборке спектров углеводов. Каждый бинарный классификатор обучали разбиению векторов образов на две категории по схеме, изображенной на рис. 4.4. При построении шести классификаторов были использованы только те спектры, которые относились к соответствующей точке ветвления. Например, весовой вектор для точки 3 строили только для тех спектров полной обучающей выборки из 200 соединений, число атомов углерода в молекулах которых было равно 4 или 5. Прогнозирующую способность каждого весового вектора определяли аналогично на контрольной выборке из 172 векторов. Результаты такого обучения и данные о прогнозирующей способности приведены в табл. 4.17. Затем обученные весовые векторы были использованы для классификации по главной программе ветвящейся классификации. Доля правильных предсказаний составляла при этом 95,4%.

Наряду с этим была проведена ветвящаяся классификация для случая, когда весовые векторы настраивали на полной обучающей

Таблица 4.17

**Бинарные классификаторы образов для ветвящейся классификации
(массив углеводов)**

Номер ветви (см. рис. 4.4)	Обучающая выборка			Контрольная выборка		
	отрица- тельная категория	положи- тельная категория	число коррекций	отрица- тельная категория	положи- тельная категория	прогнозирую- щая способ- ность, %
1	81	119	83	67	105	98,3
2	18	63	23	24	43	100,0
3	7	11	3	6	18	100,0
4	29	34	25	17	26	100,0
5	80	39	67	74	31	97,1
6	42	38	31	35	39	97,3

выборке для каждой точки ветвления, т. е. весовые векторы получали по параллельной схеме (рис. 4.3.) Только в случае первой точки ветвления (с порогом, равным 7) весовые векторы двух типов были идентичны. Общая прогнозирующая способность по второму методу составила 94,2%.

Ветвящуюся классификацию параллельными весовыми векторами проводили также на массиве из 600 спектров. Доля правильных классификаций составила 76,3%. Здесь понадобилось симметричное расположение семи бинарных классификаторов образов, чтобы разделить все углеводороды на 8 классов по числу атомов углерода в молекулах (от 3 до 10 включительно).

Параллельное соединение бинарных классификаторов образов

Ряд бинарных классификаторов можно соединять параллельно с возрастанием порогов. Например, при классификации углеводов, содержащих от 4 до 10 атомов углерода, по схеме, приведенной на рис. 4.4. Сначала соединения, содержащие от 4 до 7 атомов углерода, относили к одной категории, а соединения, с числом атомов от 8 до 10 — к другой. Затем эти категории подразделяли, согласно схеме дерева, изображенной на рис. 4.4. Другой способ состоит в том, чтобы расположить 6 классификаторов так, как это показано

на рис. 4.5. Здесь первый пороговый логический элемент относит соединения с 4 атомами углерода к одному классу, а все остальные — к другому. Второй элемент относит соединения с числом атомов

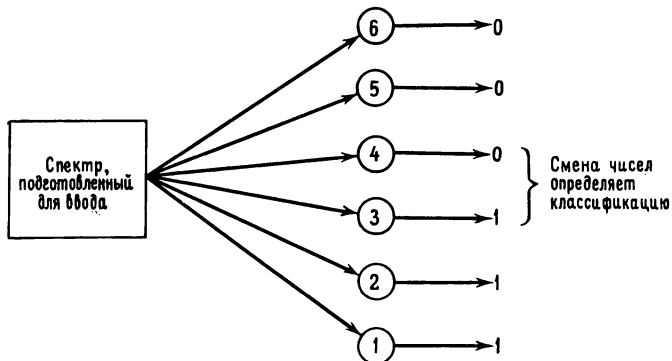


Рис. 4.5. Параллельное соединение бинарных классификаторов образов.

углерода 4 и 5 к одному классу, а остальные соединения (C_6 — C_{10}) — к другому и т. д. Параллельное соединение менее удобно, чем ветвящаяся классификация, поскольку требует больше вычислений. Однако в данном случае классификации свойственна большая избыточность, что обычно приводит к возрастанию общей прогнозирующей способности. Такая процедура обладает также определенной способностью к выявлению неизвестных образов, слабо представленных в обучающей выборке. О них можно догадываться по резким расхождениям показаний классификаторов при попытке отнесения неизвестного образа к должной категории.

На рис. 4.5 иллюстрируется параллельное соединение бинарных классификаторов образов для определения числа атомов углерода на том же массиве данных, что и в предыдущем случае. При таком расположении каждый бинарный классификатор относит векторы к одной из двух категорий, в зависимости от уровня порога. Положительная категория состоит из векторов образов с числом атомов углерода, превосходящим порог, а отрицательная — из векторов с числом атомов углерода не больше порога. Классификатор с порогом 7, например, обучали таким образом, чтобы он давал положительное скалярное произведение для векторов с числами атомов углерода, равными 8, 9 и 10, и отрицательное — для векторов с числом атомов углерода не меньше 7.

Таблица 4.18

Параллельное соединение бинарных классификаторов

Порог	Обучающая выборка			Контрольная выборка		
	отрицательная категория	положительная категория	число коррекций	отрицательная категория	положительная категория	прогнозирующая способность, %
<i>Массив углеводов</i>						
9	161	39	93	141	31	97,7
8	123	77	61	102	70	98,3
7	81	119	83	67	105	98,3
6	47	153	51	41	131	99,4
5	18	182	24	24	148	100,0
4	7	193	3	6	166	99,4
<i>Полный массив исходных данных</i>						
9	179	21	91	335	65	92,8
8	151	49	180	272	128	97,0
7	117	83	115	217	183	97,3
6	85	115	124	171	229	95,5
5	59	141	149	101	299	95,5
4	29	171	101	52	348	95,0
3	8	192	71	21	379	98,3

В табл. 4.18 приведены данные об индивидуальной прогнозирующей способности бинарных классификаторов, обучение каждого из которых проводили со своим порогом. Здесь, как и во всем этом исследовании, весовые векторы строили по каждой категории двумя способами: в первом случае все начальные компоненты считали равными +1, а во втором случае —1. Из двух таких векторов сохраняли один с лучшей прогнозирующей способностью.

Совокупность построенных векторов использовали затем в главной программе для предсказания числа атомов углерода по схеме, изображенной на рис. 4.5. Последовательность n бинарных решений можно записать в виде n -разрядного двоичного числа, состоящего из нулей в местах отрицательных решений и единиц в местах положительных решений. Затем число атомов углерода определяют по

пороговому значению, дающему первую отрицательную классификацию. Например, в случае углеводов с числом атомов углерода от 4 до 10 (включительно) число 000000 соответствует четырем атомам углерода, а число 000111 — семи. Когда цифры изменяются более одного раза, как, например, в числе 000101, это означает, что было принято по крайней мере одно неверное решение, так что соответствующий образ уже нельзя с полной определенностью отнести к должному классу. Для массива углеводов удалось правильно классифицировать 93% неизвестных спектров, а для всего массива данных — 76,0%.

Классификация при помощи бинарного кода

Чтобы отнести вектор образа к одной из m категорий, как при параллельном соединении классификаторов, так и при соединении по схеме дерева, требуется $m-1$ бинарный классификатор. Другой метод состоит в том, чтобы использовать n классификаторов, где $2^{n-1} < m \leq 2^n$, так что при представлении результата каждой классификации нулем или единицей полученное двоичное число означало бы десятичный номер соответствующей категории. До восьми категорий можно описать при помощи трехразрядного двоичного числа $d_3d_2d_1$. Например, 000 \equiv категория 1 (C_3), 001 \equiv категория 2 (C_4), ..., 111 \equiv категория 8 (C_{10}). Для углеводов, где возможны только семь категорий, существуют два способа кодирования — описанный выше (назовем его вариант 1) и еще один (назовем его вариант 2), при котором 000 $\equiv C_4$, ..., 110 $\equiv C_{10}$. Для каждого варианта проводилось обучение трех весовых векторов, разделяющих образы на соответствующие категории по числу атомов углерода (табл. 4.19). Общая прогнозирующая способность при определении числа атомов углерода составила для массива углеводов 83,7 и 80,2% по вариантам 1 и 2 соответственно и 57,8% для полного массива данных.

Точность двоичных чисел можно повысить введением дополнительных разрядов для формирования кода Хэмминга, исправляющего ошибки [13, 14]. Если оцененное двоичное число отличается от истинного только одним разрядом, т. е. если расстояние Хэмминга равно 1, то ошибку можно исправить при помощи k контрольных битов, где $2^k \geq n+k+1$ для данных из n битов. Таким образом, расстояние Хэмминга, равное 1 для исходного n -разрядного числа, увеличивается до 3 и более в $(n+k)$ -разрядном числе. В данном случае для исходных трех битов информации понадобились три контрольных бита. Это соответствует коду Хэмминга типа (6, 3).

Таблица 4.19

Бинарная классификация образов при помощи бинарных кодов

Биты	Число атомов углерода в положительной категории	Обучающая выборка			Контрольная выборка		
		отрицательная категория	положительная категория	число коррекций	отрицательная категория	положительная категория	прогнозирующая способность, %

Массив углеводов

Вариант 1	($C_4 \equiv 001$)							
	d_1	10864	83	117	395	83	89	87,8
	d_2	10965	83	117	157	67	105	95,9
	d_3	10987	47	153	51	41	131	99,4
	c_1	9854	102	98	450	74	98	92,4
	c_2	9764	92	108	535	84	88	91,3
	c_3	8765	84	116	79	76	96	98,3
	c_4	10754	109	91	603	91	81	93,6
Вариант 2	($C_4 \equiv 000$)							
	d_1	975	117	83	395	89	83	87,8
	d_2	1076	98	102	450	98	74	92,4
	d_3	1098	81	119	83	67	105	98,3
	c_1	10965	83	117	157	67	105	95,9
	c_2	10875	74	126	649	62	110	89,5
	c_3	9876	57	143	82	55	117	95,9
	c_4	865	118	82	323	102	70	93,6

Полный массив исходных данных

d_1	10864	98	102	2158	179	221	72,8
d_2	10965	95	105	1369	153	247	80,0
d_3	10987	85	115	124	171	229	95,5
c_1	9854	87	113	909	202	198	76,8
c_2	9764	93	107	2088	190	210	73,8
c_3	8765	78	122	404	180	220	90,3
c_4	10754	96	104	1460	209	191	78,3

Каждый контрольный бит c_i строят так, чтобы при правильности всех исходных битов сумма контрольного бита и двух ассоциированных с ним исходных битов была четной. В этом случае соответствующий бит четности p_i принимается равным нулю. В противном случае бит четности $p_i=1$. Три множества ассоциированных битов записываются в следующем виде:

$$p_1: c_1 + d_1 + d_2,$$

$$p_2: c_2 + d_1 + d_3,$$

$$p_3: c_3 + d_2 + d_3.$$

Ошибка в исходных битах приводит к нечетной сумме в двух группах. В этом случае двоичное число, образованное разрядами четности $p_3p_2p_1$, отлично от нуля, а его двоичный эквивалент соответствует позиции неверного разряда в числе $d_3d_2p_3d_1p_2p_1$. Например, в обозначениях по варианту 1 двоичное число для атомов углерода записывается в виде 100100. Если d_1 окажется ошибочно равным нулю, то два разряда четности, включающие d_1 (p_1 и p_2), будут равны единице, поскольку соответствующие суммы нечетные. В результате получится число 100011. Число $p_3p_2p_1 = 011$ есть двоичный эквивалент 3; это означает, что ошибка допущена в третьем разряде и, следовательно, он должен быть равен не нулю, а единице.

Для исправления ошибок понадобилось обучать еще три бинарных классификатора c_1 , c_2 и c_3 ; их обучение проводили на подходящих подвыборках из обучающей выборки, служивших положительной и отрицательной категориями. Данные о выборках, а также результаты обучения и прогнозирования приведены в табл. 4.19. Прогнозирование числа атомов углерода при помощи кодов (6, 3) повысило долю правильных ответов до 93,0 и 93,6% по вариантам 1 и 2 соответственно для массива углеводов и до 68,5% для полного массива исходных данных.

Контрольные биты имеют точность предсказания менее 100%, поэтому разумно прибегать к исправлению ошибок только тогда, когда исходные биты действительно неверны. В противном случае ошибка в битах четности может привести к неправильной классификации. Можно воспользоваться дополнительным битом четности для обнаружения ошибок в исходных битах проверкой их полной четности. Так, если сумма (по модулю 2) $d_3+d_2+d_1+c_4$ есть четное число, то $p_4 = 0$, а если нечетное, то $p_4 = 1$. Четная сумма означает, что число неверных битов равно либо нулю, либо двум; в случае ее нечетности это число равно либо единице, либо трем. Предполагается, что вероятность ошибки сразу в трех разря-

дах ничтожна. Тогда к исправлению ошибок следует прибегать только в случае нечетности общей суммы ($p_4=1$). В противном случае число атомов углерода предсказывается по исходным битам $d_3d_2d_1$. Результаты обучения и проверки битов суммарной четности приведены в табл. 4.19. Результирующие коды Хэмминга (7, 4) показали долю правильных предсказаний: 95,4% по варианту 1 и 92,4% по варианту 2 на массиве углеводородов и 67,0% на полном массиве исходных данных.

Данные о прогнозировании числа атомов углерода по разным классификационным схемам приведены в табл. 4.20. Следует отметить, что случайное угадывание числа атомов углерода дает правильный ответ в одном случае из семи (14,3%) для массива углеводородов с числом атомов углерода от 4 до 10 (включительно) и в одном

Таблица 4.20

**Сводка результатов прогнозирования числа атомов углерода
при помощи разных классификационных схем**

Классификационная схема	Доля верных предсказаний, %	
	углеводороды ^а	полный массив данных ^б
Параллельная	93,0	76,0
Ветвящееся дерево	95,4	
Ветвящееся дерево (параллельные w_j) . . .	94,2	76,3
Бинарные коды		
вариант 1: 3 бита	83,7	57,8
6 битов	93,0	68,5
7 битов	95,4	67,0
вариант 2: 3 бита	80,2	
6 битов	93,6	
7 битов	92,4	

^а $C_4 - C_{10}$; при случайном угадывании доля верных предсказаний составила $1/7$ (14,3%).

^б $C_3 - C_{10}$; при случайном угадывании доля верных предсказаний составила $1/8$ (12,5%).

случае из восьми (12,5%) для массива с числом атомов углерода от 3 до 10. Все результаты предсказаний оказались намного лучше, чем при случайном угадывании. Это означает, что классификаторы обучились многому. Результаты прогнозирования для углеводородов были во всех случаях лучше, чем для полного массива. Это легко объяснить разнородностью состава последнего. Для массива исходных спектров результаты прогнозирования всеми методами

оказались одинаковыми. Исключение составляет лишь двоичный код из трех битов. Тем не менее сравнительный анализ позволяет сделать поучительные выводы.

В схеме параллельного соединения подавляющее большинство неправильных классификаций было следствием единственной ошибки в одном из 6—7 бинарных классификаторов. Действительно, для углеводов все неправильные классификации были как раз такого рода. Здесь число неверных классификаций равнялось числу ошибочных решений, принимавшихся бинарными классификаторами. Десять ошибок пришлось на границу между нулем и единицей в двоичном слове, в двух случаях нуль попал между единицами (никакое решение не принималось). Для полного массива исходных данных 10 из 96 неверных классификаций были следствием более чем одной ошибки бинарных классификаторов. Нуль между единицами встретился три раза. Эти результаты согласуются с общим принципом, согласно которому число ошибок, допускаемых любым набором бинарных классификаторов, не может быть больше суммы двоичных ошибок и меньше числа ошибок лучшего бинарного классификатора.

Использование весовых векторов, обученных принимать параллельные решения при ветвящейся классификации, улучшает результаты прогнозирования на обоих массивах данных. Те два спектра из массива углеводов, для которых нуль оказался между единицами при параллельном соединении классификаторов, были правильно классифицированы этой схемой. Это можно объяснить удачным выбором точек ветвления. При ином выборе (если бы, например, точка ветвления 1 отделяла векторы с числами атомов углерода 10, 9, 8 и 7 от векторов с числами атомов углерода 6, 5 и 4) ветвящаяся схема дала бы такие же результаты, что и параллельное соединение. В случае полного массива спектров один из трех векторов с нулями между единицами был классифицирован правильно. Таким образом, поскольку ветвящаяся схема использует при каждой классификации не все бинарные классификаторы, достигаемая ею прогнозирующая способность не хуже, чем при параллельном соединении классификаторов для одинаковых весовых векторов и одной и той же контрольной выборки.

Прогнозирование по ветвящейся схеме лучше использует весовые векторы, обученные классифицированию на специальных выборках из обучающей совокупности, чем классификация параллельными весовыми векторами, каждый из которых обучали на всей обучающей совокупности. Это и неудивительно, поскольку в первом случае каждая точка ветвления «встречает» набор векторов обра-

зов, гораздо более «родственных» обучающей последовательности, чем во втором.

Для обучения всех весовых векторов, необходимых при классификации с помощью различных двоичных кодов, потребовалось не так уж много коррекций. Это означает, что множества с разными числами атомов углерода оказались линейно разделимыми; предвидеть это заранее было трудно. Самым неточным методом прогнозирования числа атомов углерода на обоих массивах оказалась классификация с трехбитовыми бинарными кодами. Большая часть неправильных классификаций была следствием единственной ошибки среди трех битов. Это позволяло надеяться на значительное улучшение результатов при использовании методов с исправлением ошибок. Действительно, все коды, исправляющие ошибки, улучшали показатели. Интересно отметить, что при классификации с помощью трехбитовых бинарных кодов расхождение результатов по вариантам 1 и 2 составило 2,5%, а использование добавочных корректирующих разрядов $c_1c_2c_3$ привело к результатам сопоставимой точности со всего лишь 0,6%-ным разбросом по прогнозирующей способности после коррекции.

Ошибочные контрольные биты в шестибитовых кодах не позволяют достигать оптимальной классификации как из-за пропуска ошибок, так и из-за «исправления» правильных битов. Бит четности общей суммы в коде Хэмминга (7, 4) частично устраняет подобные ошибочные «исправления». Но и в этом случае прогнозирующая способность не достигает идеального уровня, что создает новые трудности.

Для массива углеводородов C_4 дополнительный вектор показал одинаковую распознающую способность в обоих вариантах по числу атомов углерода, но по-разному сказывался на общей прогнозирующей способности. Если по первому варианту доля правильных предсказаний несколько возросла, то по второму варианту она была хуже, чем в любой схеме с 6 или 7 бинарными классификаторами. На полном массиве данных семибитовый код дает несколько худший результат, чем шестибитовый. Поэтому благоприятный эффект кода с исправлением однократных и обнаружением двукратных ошибок во многом зависит от того, какие объекты контрольной выборки приводят к неверному ответу при проверке общей четности.

Самой высокой прогнозирующей способности удалось достичь применением к массиву углеводородов ветвящейся схемы и одного из вариантов кода Хэмминга (7, 4). На полном массиве исходных данных лучший результат дала ветвящаяся схема. Надо полагать, что на других массивах при решении классификационных задач

оптимальным может быть один из этих двух методов. Однако для окончательного анализа необходимы эмпирическая проверка методов и сравнение результатов.

Применение ветвящейся схемы и кодов, исправляющих ошибки, включает три шага: разбиение объектов массива данных на нужные подвыборки, обучение бинарного классификатора образов отдельно на каждой такой подвыборке и окончательное прогнозирование комбинированием отдельных бинарных классификаторов. (Следует отметить, что в подобных устройствах можно использовать любые бинарные классификаторы. Повышения прогнозирующей способности индивидуальных бинарных классификаторов можно добиться, например, применением ненулевых порогов в процессе обучения или использованием многоуровневых пороговых логических элементов.) Параллельный метод проще, поскольку он не требует разбиения объектов массива на подвыборки.

При обучении бинарных классификаторов объекты обучающей выборки выгодно распределять равномерно между положительной и отрицательной категориями, чтобы исключить систематические отклонения при классификации. В этом отношении параллельный метод имеет большой недостаток. Как видно из табл. 4.18, кроме средних порогов, распределение объектов по категориям является в принципе неравномерным, особенно для первого и последнего порогов. Для ветвящейся схемы положительная и отрицательная категории комплектуются равномернее (табл. 4.19), поскольку некоторые объекты отбрасываются в процессе отбора. Что же касается бинарного метода, то здесь обучение каждого классификатора проводится на всей обучающей выборке с сбалансированным распределением объектов между положительной и отрицательной категориями (табл. 4.20). Более того, по мере того как число категорий возрастает, например по числу возможных атомов углерода в молекуле, распределение приближается к оптимальному (50%-ному) уровню даже в тех случаях, когда спектры исходного массива распределяются по соединениям с разным числом атомов углерода довольно неравномерно.

Двоичный код с исправлением ошибок, параллельное соединение и ветвящаяся схема требуют наличия шести бинарных классификаторов, если массив данных необходимо разбить на восемь категорий. При расширении масштабов классификации, например за счет новых категорий по числу атомов водорода и углерода в соединениях менее ограниченного массива данных, двоичные коды предпочтительнее, поскольку эта схема проще и способна обеспечить высшую прогнозирующую способность, так как можно использовать мини-

мальное число бинарных классификаторов. Число бинарных классификаторов, необходимых при ветвящейся схеме или параллельном соединении, пропорционально числу категорий, в то время как для двоичных кодов число требующихся классификаторов увеличивается намного медленнее. Например, добавлением одного бинарного классификатора (d_4) к кодам Хэмминга, изучавшимся в настоящем исследовании, число категорий по числу атомов углерода можно было бы довести до 15. Дополнительные исследования возможностей прогнозирования по молекулярному весу показывают, что коды Хэмминга (10, 4) или (11, 5) позволяют подразделять соединения с числом атомов углерода от 4 до 10 на 47 категорий. По ветвящейся схеме и при параллельном соединении для этого потребовалось бы 46 бинарных классификаторов, т. е. более чем в четыре раза больше. Высшую полную прогнозирующую способность, равную 80,8%, показал код (10, 4). Эта величина намного превосходит долю успешных предсказаний при случайном угадывании ($1/47$, или 2,1%).

Эффективность использования двоичных кодов при решении важных в химии задач определяется тем, удастся ли построить соответствующие бинарные классификаторы с достаточно высокой прогнозирующей способностью.

Кусочно линейная классификация

Кусочно линейные классификаторы можно использовать при классификации на несколько категорий. Эксперименты подобного рода описаны в работе [10]. Опыты проводились на том же массиве масс-спектров низкого разрешения, о котором речь шла в предыдущем разделе при описании кусочно линейных классификаторов.

В табл. 4.21 приведены результаты обучения кусочно линейного классификатора по величине отношения числа атомов углерода к числу атомов водорода (категории $2n+2$, $2n$, $2n-2$, $2n-4$, $2n-6$). Результаты сравнивались с характеристиками бинарного классификатора. Прогнозирующая способность кусочно линейного классификатора для многих категорий, усредненная по пяти опытам, составляла 97,2%. При разбиении на индивидуальные категории прогнозирующая способность изменялась от 98,3% для категории $2n+2$ до 92,4% для категории $2n-4$. Как показал сравнительный анализ, результаты для пяти бинарных классификаторов, построенных независимо для каждого конкретного значения C:H, были хуже, особенно для малочисленных категорий. Так, например, хотя бинарный классификатор, отделяющий соединения категории $2n-4$

Таблица 4.21

Множественная классификация по отношению С.Н в сравнении с бинарной классификацией

Множественная классификация			Бинарная классификация				
прогнозирующая способность, %	прогнозирующая способность, %		бинарная задача	прогнозирующая способность, %		состав контрольной выборки	
	по отдельным категориям	состав контрольной выборки		полная	по категориям	положительная	отрицательная
средняя					положительная		
97,2 (по пяти испытаниям)	$(2n + 2) : 98,3$	45/187	$2n + 2$ / другие категории	97,4	95,4	98,0	44
	$2n : 97,1$	73/187	$2n$ / другие категории	94,6	96,4	94,1	74
	$(2n - 2) : 98,2$	34/187	$2n - 2$ / другие категории	96,2	92,8	97,1	37
	$(2n - 4) : 92,4$	10/187	$2n - 4$ / другие категории	96,1	65,3	96,9	7
	$(\leq 2n - 6) : 95,9$	25/187	$\leq 2n - 6$ / другие категории	98,0	88,6	98,8	16
							171
							180
							150
							113
							143

от всех остальных, давал правильные ответы в 96,1% случаев, доля правильных классификаций для соединений действительно этой категории составила всего 65,3%. Таким образом, узнавая спектр неизвестного соединения категории $2n-4$, множественный классификатор дает почти на 30% более надежное решение, чем бинарный. Вполне возможно, что в случае бинарной классификации, когда объемы категорий далеко не одинаковы, систематическое отклонение идет в сторону бóльшей категории.

Классификация по K ближайшим соседям

Второй подход к задаче распознавания для случая нескольких категорий — классификация по K ближайшим соседям (КБС). В статье [15] описано применение метода КБС для интерпретации спектров ЯМР.

В методе КБС образ классифицируется согласно правилу большинства голосов по K ближайшим соседям в n -мерном пространстве. Если классифицируемый образ неизвестен, в качестве ближайших соседей берут только объекты обучающей выборки. Расчеты сводятся к вычислению и просмотру матрицы расстояний. В качестве расстояния можно взять расстояние в n -мерном евклидовом пространстве между i -й и j -й точками:

$$d_{ij} = \left[\sum_{k=1}^n (x_{ik} - x_{jk})^2 \right]^{1/2}. \quad (4.28)$$

Можно воспользоваться и любой другой метрикой, например взвешенным расстоянием.

Дальнейшее применение метода КБС рассмотрено в работе [16]. Авторы этой работы классифицировали неподвижные фазы в газожидкостной хроматографии по индексу Ковача при помощи данных, полученных для ряда различных пробных растворов. Расстояние между i -й и j -й фазами вычисляли по формуле

$$d_{ij} = \left[\sum_{k=1}^m (\Delta I_{ik} - \Delta I_{jk})^2 \right]^{1/2}, \quad (4.29)$$

где ΔI — разность индексов Ковача для сквалана* и интересующей нас фазы. Для каждой жидкой фазы наименьшее расстояние d имела ближайшая фаза, а наибольшее d — дальняя. Это расстояние

* Сквалан — неароматический углеводород ($C_{14}H_{30}$); используется как стандартная неподвижная фаза в газожидкостной хроматографии. —Прим. перев.

было введено для того, чтобы ответить на вопрос, может ли данную конкретную фазу заменить другая. Авторы составили таблицу расстояний между 226 обычными жидкими фазами и 12 ближайшими к каждой из них.

Классификация по методу наименьших квадратов

Третий способ классификации объектов на несколько категорий — метод наименьших квадратов [9].

Весовой вектор \mathbf{W} , построенный для бинарной классификации, представляет линейную комбинацию всех или некоторых векторов обучающей выборки. Скалярное произведение весового вектора и вектора образа $(\mathbf{W} \cdot \mathbf{X}_i)$ есть скаляр s_i , знак которого показывает, к какой категории следует отнести i -й образ. Принцип метода классификации на несколько категорий сводится к построению такого весового вектора, который давал бы величину s_i , модуль и знак которой свидетельствовали бы о принадлежности образа к одной из нескольких категорий. Истинной категории s_i^* , к которой принадлежит i -й образ, можно приписать любую величину. Например, при разделении соединений, имеющих 0, 1, 2, 3 или 4 атома кислорода, в качестве s_i^* соответствующих соединений можно выбрать 0, 1, 2, 3, 4. Метод наименьших квадратов применяется для расчета весов, позволяющих вычислять значения s_i , минимизирующие величину $(s_i - s_i^*)^2$:

$$Q = \sum_{i=1}^n (s_i - s_i^*)^2 = \sum_{i=1}^n \left(\sum_{j=1}^m w_j y_{ij} - s_i^* \right)^2, \quad (4.30)$$

где m — номер положения в масс-спектре, n — число образов в обучающей выборке, Q — сумма квадратов отклонений.

Нормальные уравнения для минимизации Q получают дифференцированием Q по каждому из весов и приравниванием полученных производных нулю:

$$\frac{\partial Q}{\partial w_k} = 2 \sum_{i=1}^n \sum_{j=1}^m (w_j y_{ij} - s_i^*) (y_{ik}) = 0, \quad (4.31)$$

где $k = 1, 2, 3, \dots, m$.

Нормальные уравнения решают обычным путем, чтобы определить весовой вектор \mathbf{W} , удовлетворяющий критерию наименьших квадратов.

В случаях когда число категорий мало по сравнению с числом образов в каждой из категорий, как при определении числа атомов

кислорода в молекуле по масс-спектрам, каждое значение s_i относится к ближайшей по смыслу категории. Если, например, категориям соединений с 0, 1 и 2 атомами кислорода приписываются значения s_i^* , равные соответственно 0, 1 и 2, величину $s_i = 1,68$ классифицируют как соединение с двумя атомами кислорода в молекуле. Однако произвольный выбор числа 1,5 в качестве значения, отделяющего категорию с одним атомом кислорода от категории с двумя атомами, может оказаться неоптимальным. Поэтому после процедуры наименьших квадратов применяют подпрограмму «лучшей линии», рассчитывающую лучшую разделяющую линию между двумя категориями для векторов обучающей выборки. Если категорий много, как при классификации по молекулярному весу, наиболее существенно значение s_i , ближайшее к правильному ответу, а не сама категория. Для задач с множеством категорий, таких, например, как задача о числе атомов водорода, нахождение лучшей линии может оказаться трудоемким и не дать существенного выигрыша.

Время, необходимое для расчетов по методу наименьших квадратов, пропорционально квадрату размерности задачи и по меньшей мере первой степени числа образов. Когда категорий мало, классификация по методу наименьших квадратов обычно требует больше времени, чем использование набора бинарных классификаторов. Тем не менее, поскольку объем вычислений по методу наименьших квадратов не зависит от числа категорий, его эффективность возрастает с увеличением их числа. Метод наименьших квадратов применим, впрочем, только в том случае, когда категории можно количественно упорядочить.

Классификацию по методу наименьших квадратов проверяли в два этапа с разными объемами массивов данных, заимствованных из таблиц Американского нефтяного института. Рассматривались только интенсивности, превышающие 1% интенсивности максимального пика. Это объясняется возможностями двух применявшихся вычислительных систем. Первый метод решал задачу максимизации распознавания, т. е. способности правильно классифицировать уже предъявлявшиеся классификатору образы. При этом обращалось внимание на уменьшение объема вычислений. Данные состояли из 130 масс-спектров низкого разрешения соединений с молекулярной формулой $C_{1-5}H_{1-12}O_{0-2}N_{0-2}$. Рассматривалось 79 положений m/e , так что вместе с $(d+1)$ -й координатой размерность пространства была равна 80.

В качестве примера рассматривалась задача классификации по числу атомов кислорода, которое могло быть равно 0, 1 или 2.

После каждого расчета проводилось нахождение лучшей линии. Во всех случаях, если это не оговаривалось особо, интенсивности нормировались таким образом, чтобы максимально они были равны 100; затем из них извлекали квадратный корень.

Описанный выше метод наименьших квадратов позволил построить весовой вектор, правильно классифицировавший 123 из 130 спектров. Для категорий с 0, 1 или 2 атомами кислорода успешное распознавание составило 94,6%. (Поправка за счет лучшей линии снизила число ошибочных классификаций с девяти до семи.) Ясно, что в случае трех категорий случайное угадывание дало бы прогнозирующую способность 33%.

Метод наименьших квадратов применялся также для определения типов углеводородов и структуры «средней» молекулы в таких сложных смесях, как бензин [16]. Была составлена система уравнений, по одному на каждое соединение с библиотечным спектром:

$$y_i = a_0 + \sum_{j=1}^n a_j x_{ij}, \quad (4.32)$$

где i изменяется от 1 до m (m — число соединений в выборке). Иначе говоря, величина x_{ij} из обучающей выборки — интенсивность пика в положении $m/e=j$ соединения с номером i , y_i — значение интересующей нас величины для i -го соединения, а a_j — коэффициент весового вектора, определяющего то или иное свойство смеси.

Если объем обучающей выборки m больше числа свободных параметров $n+1$, параметры могут быть найдены при помощи метода наименьших квадратов. Авторы этого исследования подчеркивают, что система уравнений является общей, и величины x_{ij} при желании можно брать из разных источников. Впрочем, данные, взятые в качестве x_{ij} , должны быть аддитивными для смеси. В качестве y_i может быть любое интересующее нас свойство смеси. Примерами служат процентный состав отдельных компонентов, процентное содержание группы соединений, например ароматических, или число метильных групп. Для каждого свойства, т. е. для каждого множества значений y_i , необходимо вычислить коэффициенты a_j . Затем значения a_j подставляют в уравнение для определения величины интересующего нас свойства данной смеси, для которой измерены значения x_j .

Метод наименьших квадратов применялся для определения некоторых особенностей «средней» молекулы бензина. Данные включали масс-спектры, спектры ЯМР, показатели преломления и плот-

ности. Основное внимание уделялось масс-спектрам. Для построения системы использовались 342 углеводорода с числом атомов углерода от 5 до 12. Систему проверяли на синтезированных смесях. Результаты этого исследования широко обсуждались. Особенно подчеркивалось, что найденное процентное содержание углерода и водорода прекрасно согласовывалось со стандартными результатами, полученными обычным химическим анализом. Отмечалось также, что этот подход особенно удобен для аналитического определения в случае, когда число компонентов смеси превышает число сделанных для нее анализов, и что при этом приходится обращать почти вырожденную матрицу.

СПИСОК ЛИТЕРАТУРЫ

1. *Jurs P. C. et al.*, Anal. Chem., **41**, 690 (1969).
2. *Jurs P. C. et al.*, Anal. Chem., **42**, 1387 (1970).
3. *Jurs P. C.*, Anal. Chem., **43**, 22 (1971).
4. *Wangen L. E., Frew N. M., Isenhour T. L.*, Anal. Chem., **43**, 845 (1971).
5. *Sybrandt L. B., Perone S. P.*, Anal. Chem., **43**, 382 (1971).
6. *Jurs P. C.*, Jap. Anal. (Bunseki Kagaki), **21**, 1276 (1972).
7. *Pietrantonio L., Jurs P. C.*, Pattern Recognition, **4**, 391 (1972).
8. *Дуда Р., Харп П.*, Распознавание образов и анализ сцен, «Мир», М., 1976.
9. *Kowalski B. R. et al.*, Anal. Chem., **41**, 695 (1969).
10. *Frew N. M., Wangen L. E., Isenhour T. L.*, Pattern Recognition, **3**, 281 (1971).
11. *Jurs P. C., Kowalski B. R., Isenhour T. L.*, Anal. Chem., **41**, 21 (1969).
12. *Felty W. L., Jurs P. C.*, Anal. Chem., **45**, 885 (1973).
13. *Peterson W. W.*, Error-Correcting Codes, MIT Press, Cambridge, Mass., 1972.
14. *Lytle F. E.*, Anal. Chem., **44**, 1867 (1972).
15. *Kowalski B. R., Bender C. F.*, Anal. Chem., **44**, 1405 (1972).
16. *Leary J. J. et al.*, J. Chromatogr. Sci., **11**, 201 (1973).
17. *Tunnicliff D. D., Wadsworth P. A.*, Anal. Chem., **45**, 12 (1973).

ОТБОР ПРИЗНАКОВ

Во многих проблемах распознавания образов образы разных классов настолько перемешаны между собой, что для их разделения на категории приходится использовать нелинейные методы. Поэтому общую задачу классификации целесообразно разделить на две части, первая из которых сводится к такому упрощению общей задачи, которое позволяет решить вторую часть. Таким образом, первоочередная задача отбора признаков заключается в уменьшении размерности без ущерба для разделения. При надлежащем и эффективном отборе признаков размерность обрабатываемых данных снижается до такого уровня, на котором не так трудно брать ту или иную разделяющую функцию.

Обработку исходных данных для отбора признаков можно осуществлять в двух направлениях:

- 1) сокращение числа исходных данных до приемлемого объема перед разделением;
- 2) исследование векторов образов для выявления тех дескрипторов (признаков), которые наиболее важны при решении задачи классификации.

Поясним разницу между этими двумя направлениями на примере. Когда нас интересуют, скажем, органические соединения с молекулярным весом ~ 300 , мы подбираем сведения о масс-спектрах для значений m/e , лежащих приблизительно в диапазоне 12—300. Это дает 288 дескрипторов. Естественно, среди них будут и положения m/e , которые не соответствуют пикам. Такие положения только увеличивают размерность пространства изображений и не несут никакой информации, поэтому при формировании векторов образов их отбрасывают. В этом заключается сущность первого направления отбора признаков. Второе направление реализуется в том случае, когда пытаются определить, какое подмножество положений m/e коррелирует с интересующей химической особенностью рассматриваемых органических соединений. Поскольку в центре наших рассуждений находится бинарная классификация,

типичную химическую задачу можно сформулировать так: соответствует или не соответствует некий спектр соединению, в структуре которого имеется кольцо. Таким образом, используя приемы отбора признаков, можно попытаться выяснить, какие положения m/e важны и какие несущественны для решения данной задачи.

При отборе признаков по первому направлению используется информация, содержащаяся только в исходных данных, например статистические параметры, вычисленные из таких данных. Этот подход тесно связан с предварительной обработкой данных и с полным правом может считаться ее составной частью. Отбор признаков по второму направлению предполагает также использование в качестве критериев при отборе важных признаков результатов обучения разделению.

Главная теоретическая трудность при отборе признаков, как и при предварительной обработке исходных данных, заключается в том, что делать выводы о полученных результатах приходится по показаниям классифицирующего звена всей распознающей системы, т. е. использовать всю систему, что создает дополнительные затруднения.

Следует отметить, что единственным надежным способом выявления оптимального подмножества, состоящего из m признаков, из всей совокупности n признаков было бы расчетное определение вкладов в формирование образов для всех таких подмножеств, а их число, как известно, равно $C_m^n = n!/(n-m)!m!$. Перебрать же исчерпывающим образом все такие случаи не представляется возможным даже для совокупности данных умеренного объема. Поэтому приходится прибегать к помощи эвристических методов. Применительно к химическим данным, как правило, используют специальные методы отбора признаков, поскольку, как выяснилось, они дают положительный результат при решении задач отбора признаков по уже упоминавшимся двум направлениям.

В литературе по распознаванию образов описаны разнообразные способы отбора признаков [1—5]. Однако только некоторые из них были применены к химическим данным. Прежде всего это объясняется тем, что распознавание образов в массиве химических данных по характеру рассматриваемой информации носит непараметрический характер. Известные методы отбора признаков пригодны в основном для обработки таких совокупностей данных, функции распределения для которых либо уже известны, либо поддаются расчету. Например, анализ главных компонент получил широкое распространение в области распознавания образов, однако применительно к химическим данным его не используют, поскольку он пред-

назначен прежде всего для рассмотрения случайных величин и тех случаев, когда важная информация связана с дисперсией.

Ниже пойдет речь об исследовании методов отбора признаков применительно к химическим данным.

Об одной из первых попыток отбора признаков для масс-спектров низкого разрешения сообщается в статье [6]. Эти данные и применявшийся способ обучения весовых векторов по методу наименьших квадратов были подробно описаны в гл. 4 настоящей книги. В использованных спектрах было закодировано по 80 положений m/e .

В задачу исследования входило обучение весового вектора определению числа атомов кислорода в молекуле для случаев, когда оно было равно 0, 1 или 2. Настройка весового вектора по методу наименьших квадратов позволила правильно классифицировать 123 из 130 спектров, что соответствовало распознающей способности 94,6%. (Поправка с учетом линии наилучшего совпадения сократила число ошибочных расчетов от девяти до семи.) Затем была исследована процедура отбора признаков.

Любая попытка уменьшить объем исходных данных и одновременно сократить число подгоняемых параметров усложняет задачу распознавания, однако это ведет к экономии расчетного времени. Так, уменьшение размерности всего в два раза намного сокращает такое время. Вопрос о том, какие положения m/e целесообразнее отбрасывать, трудно решить, ибо идеальное решение означало бы неосуществимый расчет вкладов всех возможных их сочетаний. Однако логично исключать все положения, которые меньше всего отражаются на результатах расчета. Поэтому были исследованы два следующих метода исключения положений m/e : 1) по наименьшему весу и 2) по минимальному кумулятивному влиянию на результат принятия решения. Во втором случае использовался критерий R_j — произведение веса на сумму амплитуд в соответствующем положении m/e . Это была попытка количественно рассчитать вклад каждого признака совокупности данных в общий результат принятия решения. Признаки, дававшие наименьшие вклады, исключали. На рис. 5.1 сравниваются результаты исключения признаков группами по 15 положений и пересчитанная после таких сокращений распознающая способность.

Результаты обоих методов отбора признаков оказались сопоставимыми, хотя по некоторым показателям отбор по критерию R_j был несколько лучше. Интересно отметить, что при отборе признаков по критерию R_j исключение первых 45 положений с оставлением всего лишь 35 возможных положений (со спектром, содержа-

шим в среднем ~ 15 пиков) почти не влияло на распознающую способность. Даже после отбрасывания всех масс, кроме четырех, вычисление $(d + 1)$ -го члена* давало правильный ответ в 70% случаев, что намного лучше, чем при случайном угадывании (33%). Дополнительная проверка с исключением признаков группами: по 30 положений на основе критерия R_j показала после двух итераций (отбрасывалось 60 признаков) распознающую способность 71,5%, что сопоставимо с результатом для четырех итераций с исключением 15 положений за один прием. Таким образом, хороших результатов можно достичь независимо от того, какие позиции отбрасываются.

Эти же два метода отбора признаков — по весу и по критерию R_j — рассматриваются в статье [7] применительно к бинарным классификаторам образов, корректируемым через обратную связь, а также в работе [8], посвященной распознаванию образов на основе данных из разных источников (результаты для этих данных были приведены в табл. 3.2).

ОТБОР ПРИЗНАКОВ ПО ЗНАКУ ВЕСОВЫХ ВЕКТОРОВ [9]

Рассмотренные выше методы отбора признаков носят алгоритмический характер; причем число признаков, исключенных на каждом этапе, устанавливалось заранее. Эти алгоритмы не позволяют определить, когда же следует прекращать отбор признаков. Признаки исключались даже в тех случаях, когда их отбрасывание значительно ухудшало распознающую способность классификатора. Это обстоятельство стимулировало разработку динамического метода отбора признаков, итог которого зависит от условий, складывающихся в процессе такого динамического отбора. Иными словами, признаки отбрасываются до тех пор, пока не останется ни одного признака, который не давал бы существенного вклада в общее решение. На этой стадии отбор признаков прекращается. Динамический отбор успешно использовался при обработке масс-спектров, однако прежде чем приступить к изложению самого метода, следует описать массив исходных данных.

Масс-спектрометрические данные были взяты из таблиц Американского нефтяного института, составленных по научно-исследовательскому проекту 44. Использовались 630 спектров низкого раз-

* Имеется в виду число слагаемых при вычислении скалярного произведения. — *Прим. ред.*

решения сравнительно небольших органических молекул с общей формулой $C_{1-10}H_{1-24}O_{0-4}N_{0-2}$. В каждом спектре учитывались только пики с интенсивностью, большей 1% интенсивности высшего пика. В большинстве случаев спектры содержали по 15—40 таких пиков, а их общее число составило 17 137. Интенсивности нормировали извлечением квадратного корня.

Из всего массива в 630 спектров 300 спектров, взятых случайно, были включены в обучающую выборку, а остальные 330 — в контрольную. Максимальная величина отношения m/e , соответствующая пику в спектре, оказалась равной 195. Таким образом, поскольку в этом случае $d = 195$, векторы X и Y имели размерность 195. Использование линейных решающих поверхностей предполагало независимость всех координат и отсутствие взаимодействия между слагаемыми. Для всех этих спектров нашлись 40 положений m/e , которые не соответствовали пикам, и поэтому размерность удалось снизить до 155. Дальнейшее сокращение размерности было достигнуто путем исключения из всего массива спектров еще 36 положений m/e с числом пиков менее 10. Всего из 17 137 исходных пиков при этом отборе было исключено 111 пиков, т. е. 0,6% общего числа, причем остальные пики распределялись по 119 положениям. Как показала последующая проверка, использование всех без исключения исходных пиков не могло существенным образом упростить задачу классификации.

Масс-спектрометрические данные, состоящие из 119 признаков каждого образа, исследовали по программе отбора признаков, в основу которой был положен метод линейной обучающейся машины. При этом проводили следующие операции. Два весовых вектора обучали обнаруживать наличие и отсутствие кислорода. Все компоненты одного из них считались первоначально равными +1, тогда как всем компонентам другого приписывали значение —1. Затем эти два вектора обучали классифицировать объекты обучающей выборки. Оценив прогнозирующую способность каждого из них, исключали те положения m/e , которые для классификации не существенны. Делалось это путем сравнения компонент двух весовых векторов, соответствовавших одному положению m/e ; если обе компоненты имели одинаковый знак, то такую позицию признавали хорошо коррелирующей и сохраняли, причем отрицательный знак означал отсутствие кислорода, а положительный — его наличие. Те же положения m/e , для которых компоненты двух весовых векторов имели разные знаки, считали неясными и исключали. Проверив таким образом все положения m/e , процесс обучения начинали снова на спектрах уменьшенной размерности и повторяли его

до тех пор, пока процедура отбора признаков не переставала выявлять неясные положения m/e .

В табл. 5.1 приведены результаты отбора признаков при классификации по наличию или отсутствию кислорода. Классификаторы

Таблица 5.1

**Обучение распознаванию присутствия кислорода
с отбором признаков**

Число положений m/e	ВВ=+1		ВВ=-1		Средняя прогно- зирующая способность, %
	число коррек- ций через обратную связь	прогнозирую- щая способность, %	число коррек- ций через обратную связь	прогнозирую- щая способность, %	
119	236	92,1	219	93,3	92,7
69	179	93,6	221	93,3	93,4
51	208	94,9	223	94,9	94,9
43	202	94,2	256	94,9	94,5
40	217	92,7	217	94,2	93,4
38	184	93,9	203	92,7	93,3
37	199	94,6	213	93,9	94,2
31	235	94,6	202	93,3	93,9
					Среднее: 93,8
			+	—	Всего
Обучающая выборка			82	218	300
Контрольная выборка			92	238	330

образов обучены обнаруживать наличие кислорода в соединении по его спектру независимо от типа образуемой кислородом группы. Число спектров в обучающей и контрольной выборках указано в нижней части таблицы. Обучающая выборка состояла из спектров 82 кислородсодержащих и 218 бескислородных соединений. В первой колонке таблицы приведены данные о числе положений m/e , рассматриваемых на каждом этапе процесса отбора признаков. Во второй колонке указано число коррекций через обратную связь, использованных в ходе всего процесса обучения полному распознаванию соединений обучающей выборки весовым вектором с начальным значением +1. В четвертой колонке — данные об анало-

гичном параметре для весового вектора с начальным значением —1. В третьей и пятой колонках указана прогнозирующая способность обоих весовых векторов, показанная ими при распознавании 330 соединений контрольной выборки на каждом этапе процесса отбора признаков.

В результате семи итераций число признаков для каждого образа снизилось от 119 до 37. Несмотря на такое уменьшение размерности, число коррекций через обратную связь, необходимых для обучения, оставалось приблизительно постоянным. Таким образом, полное машинное время на обучение сократилось, потому что каждая классификация была сопряжена с расчетом меньшего объема в пространстве уменьшенной размерности. Даже когда из 195 исходных положений осталось всего 37, средняя прогнозирующая способность была высокой.

В табл. 5.2 приведены результаты второго опыта, идентичного

Таблица 5.2

Обучение распознаванию наличия кислорода с отбором признаков

Число положений m/e	ВВ = +1		ВВ = -1		Средняя прогнозирующая способность, %
	число коррекций через обратную связь	прогнозирующая способность, %	число коррекций через обратную связь	прогнозирующая способность, %	
119	210	90,0	208	93,3	91,7
74	187	93,3	174	92,7	93,0
53	179	93,3	187	92,7	93,0
45	165	92,7	281	92,1	92,4
42	161	92,1	221	93,0	92,5
38	158	92,1	223	93,3	92,7
31	210	93,0	192	93,9	93,5
					Среднее: 92,7

первому (см. табл. 5.1), однако объекты обучающей выборки теперь «показывали» классификатору в иной последовательности, что приводило к другим решающим поверхностям. В двух этих опытах были получены сопоставимые результаты. Когда при помощи двух вариантов отбора признаков число положений m/e для каждого образа довели до 37 в одном случае и до 38 в другом, из двух перечней

признаков был составлен единый перечень общих признаков. Их число составило 31. Затем на совокупности из 31 положения *m/e* провели обучение двух вариантов процедуры отбора признаков. Дополнительных неясных положений ни тому, ни другому варианту выявить не удалось, поэтому обучение на этом прекратили; окончательная прогнозирующая способность составила соответственно 93,9 и 93,5% (см. табл. 5.1 и 5.2). Интересно отметить, что в обоих случаях способность классификаторов правильно относить совершенно неизвестный образ к нужному классу была почти максимальной (93,9 и 93,5%) среди всех наблюдений. По-видимому, исключение неясных положений *m/e* при разделении не ухудшает характеристики классификатора.

Из 31 положения *m/e*, выбранного в процессе отбора признаков, 14 имели положительные компоненты весового вектора, т. е. коррелировали с наличием кислорода. В табл. 5.3 эти значения *m/e* перечислены вместе с соответствующими им структурными элементами (привязку проводили по сводке фрагментов, составленной из опы-

Таблица 5.3

Вероятные фрагменты на 14 положениях *m/e*, коррелирующие с наличием кислорода

<i>m/e</i>	Фрагменты
14	CH ₂
27	C ₂ H ₃
31	CH ₃ O
37	C ₃ H
38	C ₃ H ₂ , C ₂ N
43	C ₂ H ₃ O, C ₃ H ₇ , CH ₃ N ₂ , C ₂ H ₅ N
45	C ₂ H ₅ O, C ₂ H ₇ N (п)
46	C ₂ H ₆ O
59	C ₃ H ₇ O, C ₂ H ₅ O ₂ , C ₂ H ₇ N ₂ , C ₂ H ₅ NO (п)
69	C ₄ H ₅ O, C ₅ H ₉
73	C ₄ H ₉ O, C ₄ H ₁₁ N (п)
83	C ₅ H ₇ O, C ₆ H ₁₁
100	C ₅ H ₁₀ NO, C ₆ H ₁₄ O, C ₆ H ₁₂ O(п)
135	C ₈ H ₇ O ₂ , C ₉ H ₁₁ O, C ₈ H ₉ NO

ликованных масс-спектрометрических данных [10]). Реакции перегруппировки помечены буквой «п» в круглых скобках. Многие из приведенных значений m/e отвечают ожидаемым, особенно это относится к положениям 31, 45, 59, 73. В большинстве случаев они соответствуют только кислородсодержащим фрагментам. Однако на других положениях m/e , например 14, 27, 37 и 38, кислородсодержащие фрагменты не фигурируют. По-видимому, эти пики обусловлены преимущественно кислородсодержащими соединениями из обучающей выборки и используются обучающей машиной для классификации соединений обучающей выборки, которые нельзя было распознать исходя только из кислородсодержащих фрагментов.

В табл. 5.4 и 5.5 приведены результаты поиска методом отбора признаков корреляции между масс-спектрометрическими признаками и наличием азота в составе молекул. В этих случаях метод отбора признаков позволил сократить их число соответственно до

Таблица 5.4

Обучение распознаванию присутствия азота с отбором признаков

Число положений m/e	BV = +1		BV = -1		Средняя прогнозирующая способность, %
	число коррекций через обратную связь	прогнозирующая способность, %	число коррекций через обратную связь	прогнозирующая способность, %	
119	187	93,0	171	91,8	92,4
69	150	93,3	130	93,3	93,3
55	145	93,3	142	92,7	93,0
51	140	93,3	109	92,1	92,7
47	162	92,1	95	92,7	92,4
44	143	93,0	108	93,3	93,2
43	137	93,3	120	93,3	93,3
37	128	93,9	133	93,6	93,7
36	134	93,3	141	93,3	93,3
					Среднее: 93,0

	+	—	Всего
Обучающая выборка	38	262	300
Контрольная выборка	43	287	330

Таблица 5.5

Обучение распознаванию присутствия азота

Число положений m/e	BB=+1		BB=-1		Средняя прогнозирующая способность, %
	число коррекций через обратную связь	прогнозирующая способность, %	число коррекций через обратную связь	прогнозирующая способность, %	
119	184	93,0	159	92,1	92,6
71	150	93,3	136	92,4	92,9
62	151	93,6	145	93,0	93,3
56	141	94,2	142	93,0	93,6
49	132	94,6	123	93,0	93,8
47	131	93,0	117	94,2	93,6
44	115	93,0	120	93,6	93,3
43	121	93,3	124	94,6	93,9
42	126	94,6	122	93,9	94,2
37	125	93,9	114	95,2	94,5
36	147	93,6	112	94,6	94,3
					Среднее: 93,6

43 и 42. Два перечня признаков имели 37 одинаковых положений m/e , на которых было проведено два варианта метода отбора признаков, выявившего в каждом случае по одному неясному положению. Таким образом, осталось 36 признаков, причем обучающаяся машина по одним только этим признакам правильно классифицировала все объекты обучающей выборки и соответственно 93,3 и 94,3% объектов контрольной выборки.

Половина из 36 отобранных признаков коррелировала с наличием азота в составе молекул (табл. 5.6). Перечень признаков содержит больше аномалий, чем в случае кислорода, обусловленных, по-видимому, теми же причинами, что и прежде. В данном случае значение m/e 14 отсутствует, но зато появилось положение m/e 28. Неясно, свидетельствует ли это о том, что исходные образцы были загрязнены примесью азота, или же о том, что обучающаяся машина использовала для m/e 28 иную информацию, хотя вероятность загрязнения азотом по крайней мере некоторых образцов весьма значительна.

Таблица 5.6

**Вероятные фрагменты на 18 положениях m/e , коррелирующие
с наличием азота**

m/e	Фрагменты	m/e	Фрагменты
15	CH_3	64	C_5H_4
27	C_2H_3	75	C_6H_3 , $\text{C}_3\text{H}_7\text{O}_2$, $\text{C}_2\text{H}_7\text{N}_2\text{O}$
28	N_2 , CH_2N , C_2H_4	76	CH_2NO_3 , C_6H_4
30	CH_4N	93	$\text{C}_6\text{H}_5\text{O}$, C_7H_9 , $\text{C}_6\text{H}_7\text{N}$ (п)
41	$\text{C}_2\text{H}_3\text{N}$, C_3H_5	94	$\text{C}_6\text{H}_6\text{O}$ (п)
44	CH_2NO , $\text{C}_2\text{H}_6\text{N}$, $\text{C}_2\text{H}_4\text{O}$ (п)	96	C_7H_{12}
46	$\text{C}_2\text{H}_6\text{O}$, NO_2	106	$\text{C}_7\text{H}_6\text{O}$, $\text{C}_7\text{H}_8\text{N}$, $\text{C}_8\text{H}_{10}\text{O}$ (п)
52	$\text{C}_3\text{H}_2\text{N}$, C_4H_4	116	$\text{C}_8\text{H}_6\text{N}$, C_9H_8 , $\text{C}_6\text{H}_{14}\text{NO}$
60	$\text{C}_2\text{H}_6\text{NO}$, $\text{C}_2\text{H}_4\text{O}_2$	121	$\text{C}_8\text{H}_5\text{O}$, $\text{C}_7\text{H}_5\text{O}_2$

Метод отбора признаков, основанный на использовании обучающих вычислительных устройств, показывает, что информация о вхождении кислорода или азота в состав молекул локализуется на сравнительно немногих положениях m/e масс-спектров низкого разрешения. Рассматриваемая процедура позволяет обучить классификатор распознаванию всех объектов обучающей выборки и довольно большую часть объектов контрольной выборки, используя лишь какую-то долю всех имеющихся в масс-спектре признаков. При более обстоятельном анализе выяснилось, что состав отбираемых признаков зависит от фрагментации этих больших органических молекул в масс-спектрометре и что некоторые фрагменты, не содержащие атомов кислорода или азота, все же важны для обнаружения присутствия последних в составе молекул.

Рассмотренные выше работы легли в основу другого исследования, целью которого было изучение возможностей итерационного обучения распознаванию образов по методу наименьших квадратов. Этот метод, подробно описанный в гл. 4, сочетался с отбором признаков по знаку весовых векторов, причем массив масс-спектрометрических данных был большего объема [11].

Массив данных состоял из 450 масс-спектров низкого разрешения в записи на магнитной ленте, представленных Управлением атомной энергии Англии. Спектры были заимствованы из таблиц Американского нефтяного института. Каждый спектр включал большое число пиков, интенсивности которых соответствовали последовательным положениям m/e . Интенсивность самого низкого пика со-

ставляла 0,01% интенсивности самого высокого пика. Общее число положений m/e , имеющих по крайней мере 10 пиков по всем 450 спектрам, было равно 132; поэтому максимальная размерность d векторов образов составляла 132. Все спектры принадлежали органическим соединениям с небольшим молекулярным весом общей формулы $C_{1-10}H_{1-22}O_{0-4}N_{0-2}$. Интенсивности нормировали по формуле

$$I' = 10 \lg(I \cdot 1000),$$

где I — исходная интенсивность, а I' — нормированная интенсивность (I' соответствует x_{ij} ; i означает номер спектра, а j — положение m/e). Ради удобства хранения данных значения I' округляли до целого числа. Новые значения нормированных интенсивностей находились в интервале 10—50.

Используемую в итерационном методе наименьших квадратов систему нормальных уравнений можно компактно записать в матричной форме. Система имеет единственное нетривиальное решение в том случае, если линейные уравнения независимы. Ранг матрицы коэффициентов такой же, как и размерность пространства образов. Поскольку число операций, требующихся при решении системы линейных уравнений, пропорционально n^3 (n — ранг матрицы), целесообразно размерность матрицы сводить к минимуму.

Итерационный метод наименьших квадратов был испытан при решении нескольких задач. Первая из них заключалась в классификации по наличию кислорода в органических соединениях с небольшим молекулярным весом, уже исследовавшихся ранее. Эту задачу решали обучением с отбором признаков и исправлением ошибок через обратную связь. Число признаков было сокращено от 132 до 31. Обучающаяся система безошибочно распознавала все объекты обучающей выборки; прогнозирующая способность на объектах контрольной выборки составила 93,9%.

Вектор Y_i считали равным +1, если i -й спектр указывал на присутствие кислорода, и —1 в противоположном случае. Начальные значения компонент весового вектора выбирали следующим образом: если компонента w_j имела значение p , то для компоненты w_{j+1} брали значение $-p$. В данной задаче компонента w_1 считалась равной либо +0,01, либо —0,01. Как выяснилось, эти величины оказались достаточными для того, чтобы значения скалярного произведения находились в разумном интервале (—2,5, +2,5). Минимизация расстояния между кластерами улучшала сходимость.

В табл. 5.7 приведены средние значения нескольких весовых векторов после обучения в разных условиях. Если среднее значе-

Таблица 5.7

Компоненты весового вектора при классификации по наличию кислорода

m/e	Средняя компонента весового вектора	m/e	Средняя компонента весового вектора
14	0,0208	46*	0,0081
15	—0,0250	52	—0,0178
17*	0,0275	59	0,0116
18*	—0,0003	63	—0,0215
24**	0,0007	67**	0,0102
25	—0,0003	69*	—0,0009
27	0,0049	70	—0,0026
30	—0,0183	73	0,0048
31	0,0211	83	0,0038
37	0,0042	84*	—0,0051
38**	—0,0035	86*	0,0002
39**	0,0195	91*	—0,0039
40	—0,0358	100	0,0084
43	0,0140	128	—0,0063
44	—0,0056	135*	0,0046
45	0,0080		

ние больше нуля (положительное), то соответствующее положение m/e должно коррелировать с присутствием кислорода. Если же среднее значение меньше нуля (отрицательное), то соответствующее положение должно коррелировать с отсутствием кислорода. Корреляции, установленные по методу наименьших квадратов, сравнивались с корреляциями, найденными способом с исправлением ошибок через обратную связь. Корреляция для 22 пиков была одинаковой, тогда как 9 пиков характеризовались разной корреляцией. Общая согласованность корреляций, обеспечиваемая этими двумя методами, позволяет с большей точностью установить, какие пики коррелируют с присутствием кислорода, а какие — с его отсутствием, и подтверждает, что обоим методам свойственна определенная надежность.

Быстрая сходимость решения в этой задаче с предварительным отбором признаков предполагает возможность их дальнейшего от-

бора. Проверка подтвердила такую возможность — число признаков удалось довести до 22. Значения m/e , которые были дополнительно исключены, в табл. 5.7 помечены одной звездочкой.

Метод отбора признаков сводился к использованию двух разных исходных весовых векторов. Естественно, что обучающая выборка в этих случаях была одинаковой. После обучения сопоставляли знаки компонент весовых векторов. Если знаки оказывались одинаковыми, то соответствующее положение m/e сохраняли, если же они были разными — соответствующее положение исключали. Отбор признаков по этому принципу не отразился ни на распознающей, ни на прогнозирующей способности классификатора.

В табл. 5.8 приведены результаты классификации по наличию

Таблица 5.8

Задача классификации по наличию кислорода

Число положений m/e	Число объектов обучающей выборки		Распознающая способность, %	Число объектов контрольной выборки		Прогнозирующая способность, %
	+	—		+	—	
31	42	108	99,3	26	274	98
	39	111	100,0	81	219	96
	43	107	98,7	77	223	98
22	43	107	99,3	77	223	98
18	43	107	99,3	77	223	98

кислорода для 31 положения m/e , а также после дополнительного отбора признаков. Классификатор почти безошибочно распознавал объекты обучающей выборки и показывал высокую прогнозирующую способность ($\sim 98\%$) независимо от числа сохраняемых признаков.

Этот факт подтверждает предположение, что исключенные признаки не дают почти никакого вклада в решение задачи.

Отсутствие согласованности корреляций для отдельных пиков, выявляющееся при обучении этими двумя методами, дало основания полагать, что пики с несогласующимися корреляциями можно фактически исключить, как это делается при отборе признаков. Такое исключение позволило довести число сохраняемых признаков от 22

до 18. В табл. 5.7 четыре исключенные положения помечены двумя звездочками. Результаты подобного исключения, как это видно из табл. 5.8, почти не отразились на способности классификатора к распознаванию и прогнозированию. Для оставшихся 18 пиков была достигнута полная согласованность с результатами, полученными методом с исправлением ошибок через обратную связь. Этот вывод подтверждает предположение о том, что определенные пики позволяют судить об отсутствии атомов кислорода в молекулах, тогда как другие идентифицируют присутствие кислорода.

Второй проверочной задачей была классификация по наличию атомов азота в молекулах органических соединений небольшого молекулярного веса. В данном случае удалось добиться полного распознавания, что позволило произвести отбор признаков. В результате число признаков было сокращено от 132 до 43, из которых были выбраны только те, которые соответствовали 31 наименьшему значению m/e .

Таблица 5.9

Компоненты весовых векторов при классификации по наличию азота

m/e	Средняя компонента весового вектора	m/e	Средняя компонента весового вектора
12	—0,0220	53	—0,0228
13**	0,0340	55**	0,0254
15**	—0,0180	59*	0,0001
16**	0,0164	60*	0,0001
27	0,0002	64*	0,0005
28	0,0358	74	—0,0009
29	—0,0707	75*	—0,0001
30	0,0436	76	0,0137
31	—0,0056	85*	—0,0006
39*	—0,0009	86*	0,0002
41	—0,0101	87**	0,0001
43**	0,0001	91	—0,0004
44	0,0121	92*	—0,0003
45	0,0007	94*	0,0004
46**	—0,0001	96*	—0,0001
52	0,0216		

Как и в случае с кислородом, считалось, что вектор $Y_i = +1$, если i -й спектр указывал на присутствие азота, и $Y_i = -1$ в противоположном случае. Исходные компоненты весового вектора выбирали такими же, как и в рассмотренном выше случае.

В табл. 5.9 приведены результаты обучения. Весовые векторы определялись при разных условиях обучения, в том числе на разных обучающих выборках для разных исходных весовых векторов. Были вычислены средние компоненты весовых векторов и установлены корреляционные связи с наличием или отсутствием азота. Они были сопоставлены с корреляциями, установленными итерационным методом с исправлением ошибок через обратную связь. И в данном случае обнаружилась хорошая согласованность корреляций: на 21 случай соответствия корреляций пришлось всего 10 случаев несоответствия.

В табл. 5.10 приведены результаты решения задачи классификации по наличию атомов азота в молекулах для 31 исходного пика и после дополнительного отбора признаков. Исключенные признаки помечены одной звездочкой в табл. 5.9. И здесь не наблюдалось снижения распознающей или прогнозирующей способностей. Последняя составила 96,7% для 31 признака и 98% для 21 признака.

Корреляционные связи с наличием или отсутствием атомов азота в молекулах, установленные разными способами обучения, хорошо

Таблица 5.10

Задача классификации по наличию азота

Число положений m/e	Число объектов обучающей выборки		Распознающая способность, %	Число объектов контрольной выборки		Прогнозирующая способность, %
	+	—		+	—	
31	12	138	99,3	20	280	95
	8	142	100	24	276	97
	9	141	99,7	23	277	94,7
	11	139	99,3	21	279	96,7
21	8	142	98,7	24	276	96
	17	133	98,7	15	285	97
	20	130	98,0	12	288	97
14	12	138	98,7	20	280	97

согласовывались. Пики с разноречивыми корреляциями исключали, и обучение продолжали на остальных пиках. Исключенные положения помечены в табл. 5.9 двумя звездочками. Как показывают результаты, приведенные в табл. 5.10, исключенные пики давали незначительный вклад в решение задачи. 98%-ная прогнозирующая способность была достигнута путем сокращения признаков до 14, что составило $\frac{1}{3}$ числа, полученного в случае метода с исправлением ошибок через обратную связь.

Одна из трудностей, встретившихся при решении задачи классификации по наличию азота, была связана с распределением спектров по выборкам. Более 90% использовавшихся спектров принадлежали соединениям, не содержащим азот. Поэтому решающую гиперплоскость пришлось придвинуть к кластеру азотсодержащих соединений. В результате этого распознающая и прогнозирующая способности оказались намного выше для соединений, не содержащих азот, чем для азотсодержащих молекул. Чтобы избежать эту трудность, обучающие выборки пришлось обогащать азотсодержащими соединениями путем простой замены не содержащих азот соединений азотсодержащими. Это привело к значительному улучшению расположения решающей гиперплоскости. Прогнозирующая способность в отношении азотсодержащего кластера повысилась от 40 до 75%. Разумеется, полностью преодолеть эту трудность не удалось, так как число азотсодержащих соединений было слишком мало. Лучшим выходом было бы максимально возможное пополнение совокупности исходных данных азотсодержащими соединениями.

Методом отбора признаков по знаку весовых векторов пользовались и авторы работы [12] при исследовании вольт-амперных характеристик стационарного электрода (СЭ-полярограмм), которое было описано в гл. 4. Из массива исходных данных первоначально было отобрано 133 признака, затем это число сократили отбором по знаку весовых векторов до 57 практически без всякого ущерба для распознающей способности. Способность к разделению по таким СЭ-полярограммам одно- и двухкомпонентных соединений оказалась для обучающей выборки равной приблизительно 96%, причем число неясных случаев составляло ~ 5 —6%. Дальнейшее сокращение числа признаков осуществлялось отбором по методу, изложенному в предыдущем разделе настоящей главы.

ОТБОР ПРИЗНАКОВ ПРИ ПОМОЩИ МЕТРИКИ РАССТОЯНИЙ

Разработан метод отбора признаков по критерию расстояния, который позволяет определять подлежащие исключению признаки [13].

Если известно какое-то решение для той или иной конкретной обучающей выборки, то каждой ее точке \mathbf{X}_i будет соответствовать расстояние d_i по нормали к решающей поверхности. Если точка расположена с «правильной» стороны от этой поверхности, то расстояние считается положительным; в противном случае оно рассматривается как отрицательное. Тогда полутолщину решающей поверхности можно определить как минимальное для обучающей выборки значение d_i , вычисляемое по формуле

$$d_i = \pm \frac{|\mathbf{W} \cdot \mathbf{X}_i|}{|\mathbf{W}|}, \quad |\mathbf{W}| \text{ — норма вектора } \mathbf{W};$$

$$t = \min d_i.$$

Определенное таким образом значение t представляет максимальную полутолщину конкретной решающей поверхности, при которой все еще обеспечивается сходимость решения, т. е. правильная классификация всех объектов обучающей выборки.

Чтобы оценить важность того или иного дескриптора, в процедуре отбора признаков было предусмотрено временное исключение этого дескриптора из совокупности данных. После этого были определены все расстояния d_i и найдены значения t . Предполагалось, что исключение важного дескриптора должно привести к значительному уменьшению t , тогда как отбрасывание постороннего дескриптора вызовет лишь незначительное изменение t . После выполнения таких расчетов для всех дескрипторов совокупности данных найденные значения t сравнивались. Этот алгоритм предполагает, что максимальная величина t соответствует самому малозначащему дескриптору, который подлежит исключению из совокупности данных. Такую процедуру можно неоднократно повторять, исключая дескриптор за дескриптором. Но, как оказалось, наилучшие результаты достигаются в том случае, когда в процессе отбора признаков весовой вектор периодически пересчитывается.

Проверка данного алгоритма была проведена на выборке данных, полученных при моделировании реальных ИК-спектров. Каждый синтезированный спектр состоял из 128 дескрипторов. При формировании отдельного спектра все дескрипторы полагали равными в исходном состоянии 100%-ному коэффициенту пропускания. Затем

на спектр в соответствии с законом Бэра накладывали некоторое конкретное число пиков поглощения гауссовой формы. Вершину каждого такого пика относили к случайно выбранной длине волны. Интенсивность пиков и полную ширину, соответствующую половине максимальной интенсивности, выбирали случайно, не выходя за пределы установленных диапазонов. Выборка данных состояла из 500 спектров, каждый из которых содержал 20 случайно размещенных гауссовых пиков с интенсивностями 40—80% и случайно выбранными в диапазоне 0,1—0,4 мкм значениями полной ширины, соответствующей половине максимальной интенсивности. Эти параметры были выбраны для того, чтобы как можно точнее смоделировать реальные ИК-спектры. Моделированные ИК-спектры оказались удивительно похожими на реальные.

Спектры синтезированной выборки данных в действительности представляют набор случайных фонов, на которые можно накладывать различные признаки для проверки эффективности разных способов обучения. Преимущества подобной выборки перед совокупностью настоящих спектров заключаются в том, что ее составитель имеет возможность поставить задачу обучения в тщательно контролируемых условиях, когда ему заранее известно, какие признаки важны и информацию какого рода они несут.

Из совокупности синтезированных данных была составлена обучающая выборка в 200 спектров. На половину этих спектров был наложен пик с интенсивностью 30%, полной шириной 0,4 мкм, соответствующей половине максимальной интенсивности, и серединой при длине волны 4,4 мкм. Разумеется, некоторая часть прочих синтезированных спектров должна содержать пики в данной области. Чтобы сократить объем вычислений, было решено использовать только по 50 первых дескрипторов (на длинах волн 2,0—6,9 мкм). Затем были сформированы весовые векторы, способные отличать спектры с добавленными на длине волны 4,4 мкм пиками от спектров без такого добавления.

О необходимости вызова программы обучения для перестройки весового вектора судили по следующим критериям: сначала $t_{\text{исх}}$ полагали равным значению t до первого обращения к алгоритму отбора признаков. Первый этап заключался в том, чтобы повторно обращаться к этому алгоритму, пока t не приобретет отрицательное значение. Затем вызывали программу обучения, положив $t = t_{\text{исх}}$. Если при этом достигалась сходимость весового вектора, то переходили ко второму этапу. Если же сходимость не обеспечивалась, то брали новое значение t , равное $\frac{1}{2} t_{\text{исх}}$, и опять вызывали программу обучения. Если сходимость достигалась, то переходили к

очередному этапу; если же сходимости не было, то вновь вызывали программу обучения, положив $t = 0$. Если сходимость отсутствовала и в этом случае, то дальнейших вычислений не проводили. Перед каждым этапом проверяли прогнозирующую способность решающей поверхности на контрольной выборке из 300 (150 положительных и 150 отрицательных) объектов. Результаты такого отбора признаков приведены в табл. 5.11.

Таблица 5.11

Результаты отбора признаков

Номер этапа	$t_{\text{исх}}$	Число дескрипторов, исключенных за данный этап	Число оставшихся дескрипторов	Прогнозирующая способность, %
0	2,63	—	50	98,51
1	2,63	23	27	97,55
2	1,31	11	16	97,23
3	1,31	1	15	97,24
4	1,31	1	14	97,90
5	1,31	1	13	97,89
6	1,31	1	12	97,51
7	1,31	1	11	98,59
8	0,66	1	10	97,24
9	0,66	1	9	97,25
10	0,66	1	8	96,25
11	0,66	1	7	96,92
12	0,66	0	7	—

Два первых этапа позволили легко исключить приблизительно $\frac{2}{3}$ дескрипторов. Затем на каждом этапе исключалось по одному дескриптору. Сначала многие компоненты весовых векторов имели близкие к нулю значения. А поскольку исключение дескрипторов было равноценно приписыванию соответствующей компоненте нулевого значения, отбрасывание такой компоненты почти не влияло на величину скалярного произведения и почти не изменяло решающую поверхность. Когда за первые два шага было исключено 34 дескриптора, оставшиеся дескрипторы все еще имели значительную вели-

чину, хотя и не влияли существенным образом на прогнозирующую способность, находясь далеко от 25-го дескриптора. Поэтому дальнейшее исключение дескрипторов настолько изменяло решающую поверхность, что требовалось переобучение весового вектора.

После восьмого этапа осталось 10 дескрипторов при прогнозирующей способности 97,2%. Этими дескрипторами оказались следующие (мкм): 2,0 (+); 2,2 (—); 3,0 (—); 4,0 (—); 4,3 (+); 4,4 (+); 4,5 (+); 5,4 (—); 5,6 (+) и 6,2 (—). Знаками плюс и минус в круглых скобках указано, коррелирует или не коррелирует конкретный интервал с наличием пика, наложенного при длине волны 4,4 мкм.

Был опробован также второй вариант программы отбора признаков, предусматривавшей исключение сразу двух дескрипторов до проверки того, не требует ли отрицательное значение t переобучения весового вектора. Полученные результаты сопоставимы с данными, приведенными в табл. 5.11, хотя было затрачено много меньше машинного времени. Программа отбора признаков с попарным исключением дескрипторов позволила за четыре этапа снизить число последних до восьми. При наличии всего восьми оставшихся признаков прогнозирующая способность составила 96,3%. Этими признаками были следующие (мкм): 2,4 (—); 4,0 (—); 4,3 (+); 4,4 (+); 4,5 (+); 5,4 (—); 5,8 (+) и 6,2 (—). Здесь корреляционные связи полностью подобны корреляциям, выявленным в первом случае.

Наконец, для сравнительной оценки на этой выборке данных был осуществлен отбор признаков по программе учета знаков весовых векторов. Программа предусматривала обучение двух весовых векторов; всем компонентам одного из них приписывается первоначальное значение +1, а всем компонентам другого — значение —1. После обучения знаки отдельных компонент сопоставляют, отбрасывая те из них, которые имеют разные знаки. Обучение продолжают до тех пор, пока не прекращается исключение признаков. В рамках данной задачи это привело к отбрасыванию всего 12 из 50 признаков, но обеспечило высокую прогнозирующую способность — 97,8% при 38 оставшихся дескрипторах.

Затем рассмотренная программа отбора признаков по значениям t была опробована на массиве данных, состоявшем из 500 ИК-спектров простых органических соединений с общей формулой $C_{3-10}H_{2-22}O_{0-3}N_{0-2}$. В этот массив были включены первые 500 ИК-спектров растворов (из таблиц, представленных фирмой «Садлер ризёрч лабораториз»), отвечавшие указанному критерию по составу. Каждый спектр был преобразован в цифровую форму с интервалом между отсчетами 0,1 мкм в диапазоне 2,0 — 14,7 мкм, что дало в целом 128 дескрипторов. Коэффициенты пропускания

измерялись с точностью до целого процента. Если самая интенсивная линия в спектре превосходила уровень, соответствующий 5%-ному пропусканию, то спектр нормировали в предположении действия закона Бэра, так что самая интенсивная линия соответствовала 5%-ному пропусканию. Наконец, все дескрипторы были преобразованы с соблюдением масштаба в целые числа от 0 до 31, отвечающие коэффициентам пропускания 0 — 100%.

На базе массива из 500 реальных ИК-спектров были составлены отображающие карты весовых векторов для химических соединений нескольких классов. Для каждого случая массив этих данных разбивали следующим образом. Сначала для всего массива из 500 спектров определялось число положительных объектов. В обучающую выборку было включено $\frac{2}{3}$ положительных объектов наряду с вдвое большим числом случайно выбранных отрицательных объектов. Остальные спектры вошли в состав контрольной выборки.

В отношении каждого класса химических соединений соблюдался следующий порядок. Отбор признаков был проведен по программе обучения с ненулевым порогом. Отбор по описанной выше программе включал повторные вызовы с попарным исключением дескрипторов. Когда значение t становилось отрицательным, переобучали весовой вектор \mathbf{W} и определяли прогнозирующую способность. Этим способом были исследованы три класса химических соединений: карбоновые кислоты, сложные эфиры и первичные амины.

Карбоновые кислоты

Обучающая выборка состояла из 40 положительных и 80 отрицательных объектов, тогда как контрольная делилась на 21 положительный и 359 отрицательных объектов. Результаты обучения и отбора признаков обобщены в табл. 5.12. Прогнозирующая способность была высокой (94,5%) даже в том случае, когда осталось всего 18 дескрипторов. Правда, при последующем отборе признаков она стала заметно снижаться. Когда же осталось 6 признаков, сходимость уже не достигалась, так что отбор признаков на этом этапе пришлось прекратить.

На рис. 5.2 изображена карта весового вектора с 36 координатами для карбоновых кислот. Как и раньше, масштаб компонент весовых векторов выбран произвольно, а сами компоненты нанесены в том же направлении, что и для ИК-спектров. На этой карте ряд участков представляет интерес и поэтому рассматривается ниже в порядке перехода слева направо.

Таблица 5.12

Обучение и отбор признаков для карбоновых кислот

Число признаков	Число коррекций через обратную связь	Прогнозирующая способность, %
128	753	95,9
34	272	95,6
22	487	93,5
18	409	94,5
16	1143	91,7
14	2049	90,7
12	1508	91,2
10	1842	92,6
8	2838	88,5
6	—	—

На участке длин волн 2,4—2,8 мкм наблюдается явная отрицательная корреляция с карбоновыми кислотами. Данная область соответствует валентным колебаниям свободных гидроксильных групп. Однако в недиссоциированных кислотах гидроксилы обычно связаны водородной связью, что смещает поглощение в область более длинных волн. Таким образом, в данной группе дескрипторов свободные гидроксильные группы как кислоты не классифицируются.

Все 8 дескрипторов на длинах волн 3,2, 3,3 и в диапазоне 3,8—4,3 мкм положительно коррелируют с карбоновыми кислотами. В этой области спектры карбоновых кислот обычно обнаруживают широкую полосу поглощения, соответствующую водородной связи с участием гидроксила. Поскольку такой пик поглощения характерен в основном карбоновые кислоты, не составило большого труда включить данную информацию в программу обучения бинарного классификатора образов.

Дескрипторы при длинах волн 5,8 и 5,9 мкм обнаруживают положительные корреляции. Этот участок соответствует валентным колебаниям карбонильной группы.

Остальные дескрипторы не обнаруживают явной корреляции с какой-то конкретной особенностью карбоновых кислот. Обычно они имеют отрицательные значения и, по-видимому, обусловлены

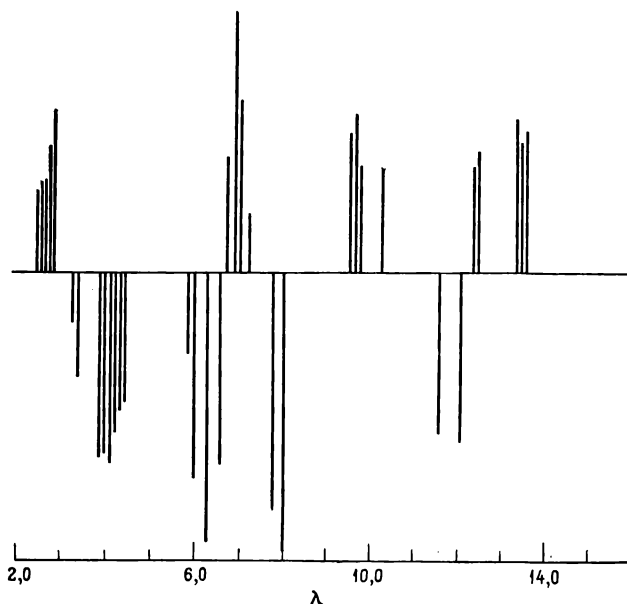


Рис. 5.2. Схема координат весового вектора для карбоновых кислот.

особенностями отрицательных объектов обучающей выборки. Экзанацционная проверка весового вектора, у которого осталось всего 18 компонент, не выявила новых корреляций, кроме уже рассмотренных.

Сложные эфиры

Обучающая выборка содержала 40 положительных и 80 отрицательных объектов, тогда как контрольная — 20 положительных и 360 отрицательных объектов. Полученные на этих выборках результаты обобщены в табл. 5.13. Прогнозирующая способность с уменьшением числа дескрипторов медленно постепенно убывала, а когда осталось 8 дескрипторов, сходимость весового вектора уже не достигалась.

Таблица 5.13

Обучение и отбор признаков для сложных эфиров

Число признаков	Число коррекций через обратную связь	Прогнозирующая способность, %
128	720	96,6
36	458	95,5
24	604	94,4
14	541	93,5
12	319	93,2
10	516	91,5
8	—	—

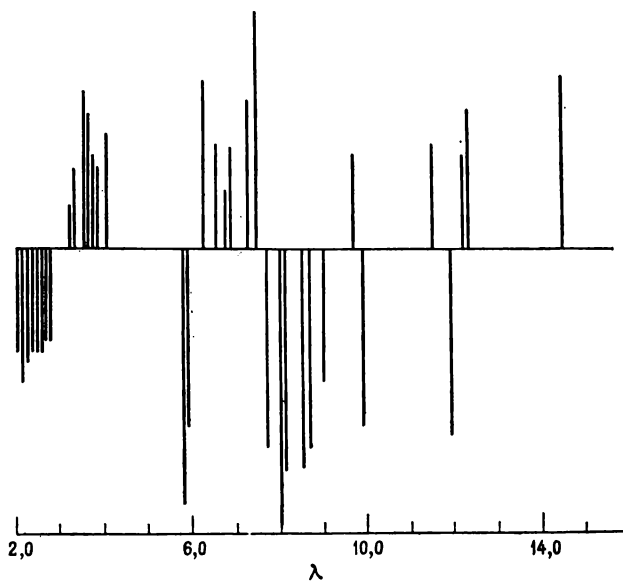


Рис. 5.3. Схема координат весового вектора для сложных эфиров.

На рис. 5.3 изображена карта весового вектора с 36 координатами для сложных эфиров. Легко понять положительную корреляцию между эфирами и дескрипторами на участках, соответствующих длинам волн 5,7; 5,8; 7,9; 8,0; 8,4; 8,6 и 8,9 мкм, потому что здесь проявляются обычные валентные колебания эфирной группы. Отрицательную корреляцию для дескрипторов на длинах волн 3,1; 3,2 и 3,5—3,8 мкм можно, по-видимому, приписать тому, что эти дескрипторы особенно полезны для исключения валентных колебаний С—Н альдегидов и возможных валентных колебаний гидроксильных групп карбоновых кислот, участвующих в образовании водородной связи. Сильную корреляцию между сложными эфирами и дескрипторами при 2,0 и 2,7 мкм трудно объяснить. Отнюдь не исключено, что здесь мы имеем дело с приборным артефактом.

Первичные амины

Обучающая выборка включала 38 положительных и 76 отрицательных объектов, тогда как контрольная — 20 положительных и 366 отрицательных объектов. Результаты обучения и отбора признаков приведены в табл. 5.14. Прогнозирующая способность была

Таблица 5.14

Обучение и отбор признаков для первичных аминов

Число признаков	Число коррекций через обратную связь	Прогнозирующая способность, %
128	598	95,2
32	686	95,5
24	432	96,0
18	366	95,1
12	3560	92,3
10	3991	92,6
8	—	—

высокой до тех пор, пока число оставшихся дескрипторов не стало меньше 18, затем она начала снижаться.

На рис. 5.4 изображена карта весового вектора с 32 координатами для первичных аминов. Группу дескрипторов на участке

3,6—4,0 мкм, положительно коррелирующих с первичными аминами, трудно объяснить. Не исключено, что они обусловлены колебаниями ионов NH_4^+ , образующихся при попадании воды в образцы.

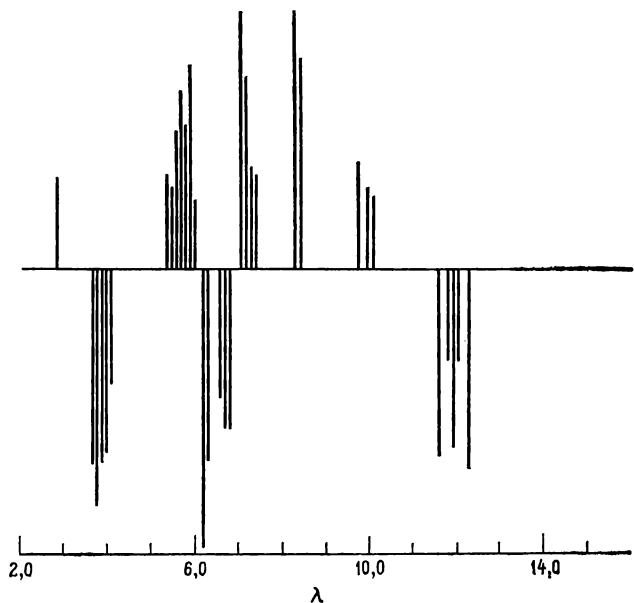


Рис. 5.4. Схема координат весового вектора для первичных аминов.

На группе дескрипторов в интервале 5,3—5,9 мкм проводилось обучение по отношению к характеристикам соединений других классов. Дескрипторы на длинах волн 6,1 и 6,2 мкм коррелируют с деформационными колебаниями N—H первичных аминов. Положительно коррелирующие дескрипторы на более длинных волнах могут быть отнесены к известным широким полосам поглощения аминов.

СПИСОК ЛИТЕРАТУРЫ

1. Tou J. T., Pattern Recognition, 1, 3 (1968).
2. Levine M. D., Proc. IEEE, 57, 1391 (1969).
3. Fukunaga K., Koontz W. L. G., IEEE Trans. C-19, 311 (1970).
4. Patrick E. A., Fischer F. P., II, IEEE Trans., IT-15, 577 (1969).
5. Во многих монографиях, список которых приведен в конце книги, есть главы об отборе признаков.

-
6. Kowalski B. R. *et al.*, Anal. Chem., **41**, 695 (1969).
 7. Jurs P. C. *et al.*, Anal. Chem., **41**, 690 (1969).
 8. Jurs P. C. *et al.*, Anal. Chem., **41**, 1949 (1969).
 9. Jurs P. C., Anal. Chem., **42**, 1633 (1970).
 10. McLafferty F. W., Advan. Chem. Ser., **40** (1963).
 11. Peitran Antonio L., Jurs P. C., Pattern Recognition, **4**, 391 (1972).
 12. Sybrandt L. B., Perone S. P., Anal. Chem., **44**, 2331 (1972).
 13. Preuss D. R., Jurs P. C., Anal. Chem., **46**, 520 (1974).

ДОПОЛНИТЕЛЬНЫЕ ПРЕОБРАЗОВАНИЯ

Как отмечалось в гл. 1, цель любой процедуры распознавания образов — преобразование пространства образов в классифицирующее пространство, т. е. осуществление таких преобразований данных, которые переводят их в нужные категории. Эту задачу можно трактовать как отображение пространства $(d + 1)$ -й размерности в пространство гораздо меньшей размерности, чаще всего в одномерное. Линейные классификаторы образов, оперирующие с каждым измерением независимо, были довольно подробно рассмотрены в предыдущих главах настоящей книги. Однако во многих случаях точки образов не обладают свойством линейной разделимости. Успешное решение задачи классификации образов в подобных ситуациях требует либо использования решающей поверхности более высокого порядка, либо такого преобразования исходных данных, которое превращает их в линейно разделимое множество. (Это утверждение предполагает, что неразделимость отражает истинную природу исходных данных, а не является просто следствием неадекватности дескрипторов и т. п.).

Последний подход, предполагающий преобразование исходных данных в линейно разделимое множество, весьма удобен, поскольку он позволяет воспользоваться обстоятельно разработанными приемами линейной классификации. Поэтому в настоящей главе рассматриваются преобразования, которые уже не оперируют независимо с каждым измерением. Были предприняты попытки использования некоторых таких преобразований при обработке масивов химических данных для упрощения классификации.

Если бы всегда заранее было известно правильное преобразование, то, разумеется, решение любой задачи методами распознавания образов свелось бы к тривиальному случаю. В действительности же правильное преобразование известно лишь в редких случаях, так что приходится вести «игру» для выбора такого преобразования, которое обеспечивает приемлемую точность приближения. Пока еще не создана строгая схема выбора правильного преобразования.

Поэтому здесь нужна интуиция ученого. Подобная ситуация наблюдается во многих других областях, например успешное решение дифференциальных уравнений часто зависит от способности математика правильно выбрать необходимое преобразование.

Нужно иметь в виду, что линейная обучающаяся машина — это просто один из методов определения коэффициентов линейной весовой функции. Если предположить, что какая-то аналитическая функция связывает исходные данные с категориями, то весовая функция в виде полинома должна идеально классифицировать все данные:

$$s = w_1x_1 + w_2x_2 + \dots + w_{d+1}x_{d+1} + w_{11}x_1^2 + w_{22}x_2^2 + \dots \\ \dots + w_{12}x_1x_2 + w_{13}x_1x_3 + \dots \quad (6.1)$$

В первом приближении линейную разделяющую (классифицирующую) функцию можно рассматривать как полиномиальную функцию. Следующим шагом логически должна стать разделяющая функция второго порядка, т. е. квадратичная функция. Эту функцию можно записать в таком виде, где отдельные члены разделены и d весов служат коэффициентами при членах x_j^2 , еще d весов — коэффициентами при членах x_j , $d(d-1)/2$ весов — коэффициентами при членах $x_j x_k$ (перекрестных членах, когда $j \neq k$) и дополнительно еще один вес, который коэффициентом уже не является, но связывает добавочный $(d+1)$ -й член. Таким образом, эту функцию можно представить в следующем виде:

$$s = w_1f_1(X) + w_2f_2(X) + \dots + w_mf_m(X) + w_{m+1}, \quad (6.2)$$

где $m = d + d + d(d-1)/2$. Важно отметить, что весовые коэффициенты входят в данное выражение линейно. Следовательно, реализовать квадратичную разделяющую (дискриминантную) функцию можно одним из двух способов:

$$\begin{array}{l} \text{Квадратичная} \\ X \rightarrow \text{разделяющая} \rightarrow \text{Классификация.} \\ \text{функция} \\ \\ X \rightarrow \text{Квадратичный} \rightarrow \text{Линейная} \\ \text{препроцессор} \quad \text{разделяющая} \rightarrow \text{Классификация.} \\ \text{функция} \end{array}$$

Это свойство является общим для довольно широкого класса функций, называемых Ф-функциями [1]. Квадратичная функция пред-

ставляет простой случай такой Φ -функции. К числу Φ -функций относятся, например, всевозможные полиномы.

Многие сложные разделяющие функции можно реализовать сочетанием подходящего препроцессора или преобразователя с линейной разделяющей функцией. Понимание этого важного вывода привело к проведению ряда исследований различных препроцессоров, речь о которых пойдет ниже.

ГЕНЕРИРОВАНИЕ ПЕРЕКРЕСТНЫХ ЧЛЕНОВ

Применение методов распознавания образов в масс-спектрометрии на первых порах почти всегда проводилось с использованием пороговых логических элементов. Такие распознающие системы принадлежат к категории линейных систем, поскольку масс-спектрометрические пики считаются в данном случае не зависящими друг от друга. Между тем теория масс-спектрометрии, равно как и фундаментальные основы классификации образов, позволяют предположить, что при подобной классификации можно было бы успешно использовать взаимодействия второго порядка (перекрестные члены, учитывающие зависимости между пиками). В статье [2] сообщается об использовании меры подобия к данным масс-спектрометрии низкого разрешения для вывода перекрестных членов двух типов: внутригрупповых (для объектов одной выборки) и межгрупповых (для объектов нескольких выборок). Показано, что для полученных таким образом межгрупповых перекрестных членов существует большая вероятность корреляции с теми молекулярными признаками, которые можно положить в основу разбиения на категории. Это предположение было реализовано в виде классификаторов образов на пороговых логических элементах, проверявшихся на нескольких выборках масс-спектрометрических данных. Как оказалось, перекрестные члены расширяют возможности систем классификации образов либо ускоряя сходимость, либо повышая прогнозирующую способность этих систем, либо же обеспечивая и то и другое одновременно.

На рис. 6.1 изображен масс-спектр в общем виде помеченного графа. Поясним используемую нами символику. Узлы ($v_1, v_2, \dots, v_i, v_j, \dots, v_p$) изображают пики в конкретном масс-спектре, а стрелками (ребрами), соединяющими такие узлы, указаны пути взаимодействия, которое ведет к возникновению ионов фрагментов. Так, v_1 означает материнский, или молекулярный, ион, из которого образуются все другие ионы. Для графа с перенумерованными узлами можно составить матрицу смежности $A = [a_{ij}]$. Такая матрица

будет квадратной и симметричной. Для графа с p узлами это будет матрица из $p \times p$ элементов, которые определяют следующим образом: $a_{ij} = 1$, если узел v_i расположен по соседству с узлом v_j ;

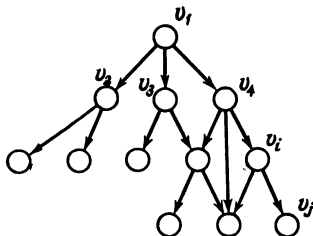


Рис. 6.1. Ациклический направленный граф в общем виде.

в противном случае $a_{ij} = 0$. Матрица смежности $p \times p$ представляет полное и однозначное отображение графа с p узлами.

Данные многих типов (такие, как интересующие нас масс-спектрометрические данные) можно выразить в векторной форме. Действительно, каждый спектр из выборки масс-спектрометрических данных можно записать как $X = x_1, x_2, \dots, x_p$, где каждая отдельная компонента вектора соответствует одному пику в спектре; так, x_{31} характеризует интенсивность пика с m/e 31. Из совокупности подобных векторов, характеризующих выборку данных, можно вычислить следующие величины: b_{ii} — число векторов в выборке данных с ненулевым членом x_i ; b_{ij} — число векторов с ненулевыми значениями x_i и x_j . Например, для массива масс-спектрометрических данных из 100 векторов значение $b_{15,15}$ будет равно 50, если половина векторов имеет ненулевую компоненту x_{15} . Значение $b_{15,30} = 40$ означает, что 40 векторов имеют пики в обоих положениях m/e 15 и m/e 30.

Эти величины используют при вычислении элементов c_{ij} матрицы сходства для всех попарных сочетаний пиков. Если все векторы в выборке данных имеют размерность p , матрица сходства для всех попарных сочетаний пиков будет квадратной с числом элементов, равным $p \times p$. Каждый элемент вычисляют по формуле

$$c_{ij} = \frac{b_{ij}}{b_{ii} + b_{jj} - b_{ij}}. \quad (6.3)$$

Элемент c_{ij} характеризует степень связи между i -й и j -й компонентами набора векторов. Эта мера сходства особенно удобна в при-

менении к масс-спектрам низкого разрешения, поскольку она точно связана с позициями компонент векторов, а не с их амплитудами.

Такая матрица сходства состоит из элементов c_{ij} ; значения c_{ij} , вычисленные по формуле (6.3), лежат в пределах от нуля до единицы; чем больше c_{ij} , тем сильнее зависимость между членами x_i и x_j . Эту матрицу можно преобразовать в матрицу смежности путем сравнения каждого значения c_{ij} с порогом T , принимая затем $c_{ij} = 1$, если $c_{ij} > T$, и $c_{ij} = 0$ во всех остальных случаях. Можно исследовать число ненулевых элементов полученной таким образом матрицы смежности как функцию порога. Каждая 1, фигурирующая в составленной пороговым нормированием матрице смежности, соответствует отдельному перекрестному члену, появляющемуся в выборке данных достаточно часто, чтобы превзойти пороговое значение. Подобные перекрестные члены могут служить полезным признаком для пороговых логических элементов при разделении данных в целях классификации. Следовательно, такие члены можно рассматривать как полезные при классификации признаки. Эти признаки явно относятся к внутригрупповым, поскольку они вводятся для элементов множества векторов в целом.

В основе подхода к составлению межгрупповых признаков также лежит рассмотренная выше матрица сходности для попарных сочетаний всех членов. При этом методе все множество исходных данных разбивают на два подмножества, которые классификатор образов должен научиться распознавать. Затем по формуле (6.3) для каждого из двух подмножеств составляют матрицу сходства с членами c_{ij}^1 и c_{ij}^2 . После этого определяют абсолютную величину разности между этими двумя матрицами:

$$\Delta c_{ij} = |c_{ij}^1 - c_{ij}^2|. \quad (6.4)$$

Мы получили матрицу с элементами Δc_{ij} , где величина каждого элемента выражает разность мер сходства перекрестных членов, образованных i -й и j -й компонентами двух подмножеств данных. Этим методом выбора межгрупповых признаков пользовались при работе с масс-спектрами низкого разрешения для отбора перекрестных членов, подлежащих включению в совокупность признаков, представляемых системе классификации образов.

Для массива из 630 масс-спектров низкого разрешения по формуле (6.3) была составлена матрица сходства для всех попарных сочетаний компонент. Затем эту матрицу преобразовали пороговым нормированием для различных значений порога T . Кривые a и b на рис. 6.2 представляют графики зависимости числа перекрестных членов (или ребер), имеющих в совокупности данных, от ве-

личины порога T . При построении этих кривых по оси абсцисс брали те значения x , которые указаны в верхнем ряду. Кривая a относится к перекрестным членам, образующимся при всевозможных

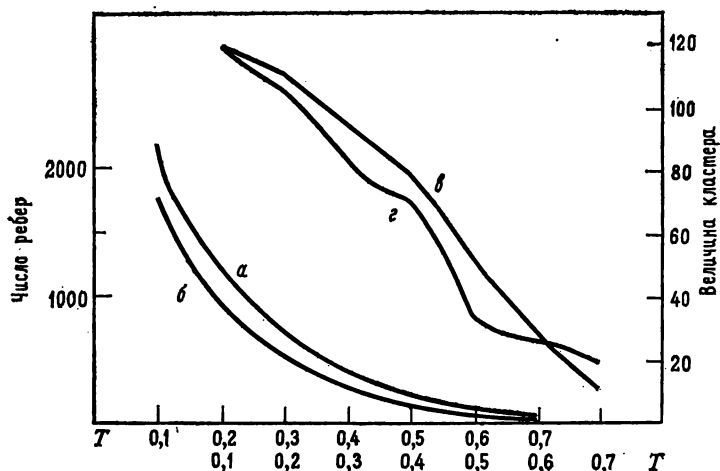


Рис. 6.2. Графики изменения числа ребер в массиве данных (кривые a , $б$) и объема самого крупного кластера (кривые $в$, $г$) в зависимости от величины порога T для масс-спектров с 119 положениями m/e .

a , $в$ — для всех положений m/e ; $б$, $г$ — для положений m/e , различающихся не менее чем на 10 единиц.

сочетаниях положений m/e , тогда как кривая $б$ характеризует число перекрестных членов, образующихся попарным сочетанием таких положений m/e , которые различаются между собой не менее чем на 10 единиц. Форма обеих кривых одинакова. В случае 119 положений m/e число всех возможных перекрестных членов второго порядка равно $(119)(118)/2 = 7021$, так что 1246 перекрестных членов с пороговыми значениями $T > 0,2$ образуют только какую-то долю всех возможных сочетаний.

Кривые $в$ и $г$ на рис. 6.2 характеризуют число узлов в самом крупном кластере как функцию величины порога. Кластер определяют как совокупность таких узлов, которые соединены между собой ребром. Переход от одного узла подобного класса к другому можно осуществить последовательным перемещением по ребрам. Как показывает график (рис. 6.2), число узлов в самом крупном кластере тем меньше, чем выше порог T .

Описанным методом поиска межгрупповых признаков были обработаны масс-спектры соединений нескольких характерных химических категорий для выявления представляющих интерес перекрестных членов, коррелирующих с двумя классами химических соединений. Множество из 450 масс-спектров было разбито на два подмножества. Затем по формуле (6.3) были вычислены соответствующие матрицы сходства для попарных сочетаний компонент и, наконец, по формуле (6.4) была найдена разность между двумя этими матрицами. Результаты трех таких расчетов обобщены в табл. 6.1. Эти расчеты проводились для следующих вопросов, допускающих только два варианта ответа: 1) содержится или не содержится в соединении кислород? 2) больше или меньше 14 число атомов водорода в молекуле? 3) больше или меньше двух отношение числа атомов водорода к числу атомов углерода? Выбранные в этом исследовании категории (классы) типичны для той информации, какую полезно знать о соединении, когда перед обучающейся вычислительной машиной ставится задача распознать соединение по его спектру.

В верхней части табл. 6.1 приведены числа перекрестных членов для обоих подмножеств при некоторых значениях порога T в интервале 1,00—0,80. Число перекрестных членов для разных категорий изменяется в широких пределах при каждом значении T . Исследованное для каждой категории число перекрестных членов указано в шестой строке таблицы (это члены с высшими значениями c_{ij} , по которым вычисляли значения Δc_{ij}). По результатам расчетов были отобраны отдельные перекрестные члены с наибольшими значениями Δc_{ij} . Число отобранных таким образом перекрестных членов для каждой категории указано в седьмой строке. В нижней части табл. 6.1 приведены данные о профилях шести наборов исследованных перекрестных членов. Величина и интервалы изменения значений c_{ij} и Δc_{ij} с переходом от одной категории к другой изменяются в весьма широком диапазоне. Наибольшие значения Δc_{ij} принадлежат перекрестным членам контрольного набора для определения присутствия или отсутствия кислорода (со средними значениями Δc_{ij} , равными 0,74 для кислородсодержащих молекул и 0,66 для бескислородных соединений).

Чтобы проверить, в какой степени эти перекрестные члены коррелируют с молекулярными признаками, на основе которых проводилось разделение на категории, их вводили в классификатор образов на пороговых логических элементах. Классификатор испытывался на выборке из 450 спектров по 132 положениям m/e в каждом. Были выполнены три цикла экспериментов для трех вариантов

Таблица 6.1

Межгрупповые перекрестные члены

Порог	Кислород		$H > 14$ (164)	$H \leq 14$ (286)	$2C:H < 1$ (141)	$2C:H \geq 1$ (309)
	содержится (120)	отсутствует (165)				
1,0	2	0	10	0	21	0
0,95	28	65	283	15	81	21
0,90	69	137	559	38	131	68
0,85	136	273		79	194	124
0,80	220	432		125	292	242
Число исследованных членов	220	220	200	200	292	242
Число отобранных членов	13	14	11	10	10	11
Значения Δc_{ij}						
максимальное	0,769	0,688	0,498	0,652	0,551	0,367
среднее	0,738	0,658	0,456	0,560	0,514	0,313
минимальное	0,701	0,613	0,413	0,393	0,501	0,301
Значения c_{ij}						
максимальное	0,886	0,912	0,987	0,903	0,932	0,852
среднее	0,818	0,898	0,972	0,791	0,914	0,826
минимальное	0,805	0,881	0,968	0,753	0,901	0,814

классификации, показанных в табл. 6.1. В начале каждого эксперимента линейный классификатор образов обучали, используя все 132 положения m/e . В обучающую выборку включили 250 случайно выбранных спектров, а оставшиеся 200 спектров — в контрольную выборку. Размерность пространства образов уменьшали отбором признаков по знаку компонент весовых векторов путем исключения тех положений m/e , которые несущественны для проводимой классификации (без ущерба для скорости сходимости или прогнозирующей способности). Затем отобранные в соответствии с изложенной выше процедурой перекрестные члены размещали по вакантным положениям и снова обучали классификатор, периодически проверяя прогнозирующую способность. Результаты данного исследования представлены в табл. 6.2. Испытания проводили в три цикла на разных обучающих выборках *A*, *B*, *C*. Во всех трех случаях при обучении и определении прогнозирующей способности предусматривали мертвую (свободную) зону шириной 50 единиц. Интенсивности всех пиков подвергали логарифмическому преобразованию. Интенсивности для перекрестных членов вычисляли умножением интенсивностей для двух соответствующих положений m/e с последующим извлечением квадратного корня из произведения, чтобы не выходить за пределы нормировки.

В табл. 6.2 обобщены результаты такой процедуры классификации на присутствие кислорода. Классификатор образов здесь должен обучаться выявлению присутствия кислорода в любой функциональной группе в составе соединения, масс-спектр которого является объектом классификации. В первой колонке перечислены обозначения трех обучающих выборок, использованных в опытах. Во второй и третьей колонках таблицы приведены данные о числе линейных положений m/e и числе перекрестных членов, вводимых в классификатор. В четвертой колонке — данные о числе коррекций через обратную связь, необходимых для достижения полного (100%-ного) распознавания; каждое из двух чисел относится к одному из двух дублирующих опытов, которые были проведены с разными исходными весовыми векторами (первый для компонент, равных +1, а второй для компонент, равных —1). В пятой колонке приведены данные о показателе скорости сходимости, достигаемом при введении перекрестных членов. Этот показатель выражается отношением общего числа коррекций через обратную связь при обучении с перекрестными членами к их общему числу при обучении только на линейных членах. Таким образом, если данный показатель равен единице, то скорость сходимости не изменяется. Когда же он равен 0,5, сходимость при обучении с перекрестными членами

Таблица 6.2

Обучение классификации на присутствие кислорода в составе соединения

Обучающая выборка	Число положений m/ϵ	Число пере- крестных членов	Число коррекции через обрат- ную связь	Показатель скорости сходимости	Прогнозирующая способность, %		Число отказов	Число исключенных признаков
					фактическая	средняя		
A	95	—	157/151	0,54	96,9/98,5	97,7	6/8	11
	95	27	79/89		95,4/95,4	95,4	3/3	19/1
B	95	—	185/164	0,77	98,4/97,9	98,1	10/9	12
	95	27	147/122		98,0/96,9	97,5	5/4	13/1
B	95	—	284/320	0,58	98,9/98,4	98,7	11/8	13
	95	27	173/167		99,5/99,0	99,3	5/10	49,0

Обучающая выборка Обучающая выборка Контрольная выборка

	Обучающая выборка		Контрольная выборка	
	+	—	+	—
A	74	176	46	154
B	68	182	52	148
B	73	177	47	153

улучшается в два раза по сравнению с обучением только на линейных членах. В шестой колонке представлена прогнозирующая способность двух классификаторов образов, а в седьмой — средняя прогнозирующая способность для этих двух случаев. В восьмой колонке указано число образов из 200 объектов контрольной выборки, которые остаются не классифицированными, поскольку соответствующие скалярные произведения попадают в мертвую зону. В последней колонке приведены данные о числе признаков распознаваемых образов, исключенных после обучения по программе отбора признаков. (Исключались признаки, для которых компоненты соответствующих весовых векторов имели противоположные знаки.) Две цифры в этой колонке при обучении с перекрестными членами означают число исключенных линейных и перекрестных членов. Состав обучающей и контрольной выборок указан в нижней части таблицы.

При обучении классификации на присутствие кислорода в составе соединения программа выделения линейных признаков сокращала число положений m/e от 132 до 95. При наличии 95 положений классификатор образов быстро сходился к полному распознаванию, при этом прогнозирующая способность составила 97,7, 98,1 и 98,7%. С введением 27 перекрестных членов (тех же, что и в табл. 6.1) характеристики классификаторов образов изменялись. В опытах с обучением классификации на присутствие кислорода показатель скорости сходимости находился в интервале 0,54—0,77. Такой значительный выигрыш в сходимости свидетельствует о том, что отобранные перекрестные члены сильно коррелировали с присутствием или отсутствием кислорода в составе соединения. В случае обучающих выборок B и B прогнозирующая способность очень мало изменялась при введении перекрестных членов, а для выборки A она немного снижалась. Анализ данных, приведенных в восьмой колонке, показывает, что классификаторы образов при введении перекрестных членов дают меньше отказов, чем при распознавании только по линейным признакам. Данные, представленные в последней колонке, указывают на то, что программа отбора признаков исключает во всех случаях введения перекрестных членов больше линейных дескрипторов, чем при обучении только на линейных членах. Как показала программа отбора признаков, почти все перекрестные члены были полезными при классификации.

В контрольных опытах, подобных представленным в табл. 6.2, но проведенных с отбором перекрестных членов экспериментатором или на основе значений c_{ij} (внутригрупповые перекрестные члены), было установлено, что такие члены при классификации образов

несущественны. Следовательно, отбор межгрупповых перекрестных членов по значениям Δc_{ij} позволяет сформировать признаки второго порядка, улучшающие характеристики классификатора образов.

ПРЕОБРАЗОВАНИЕ ФУРЬЕ [3]

Преобразование Фурье чаще всего трактуют как такое преобразование, которое связывает временную область с частотной областью. Так, положительная часть результата преобразования Фурье от конечной косинусоидальной волны есть просто пик, центр которого соответствует частоте волны. И вообще преобразование Фурье дает как раз частоту, приходящуюся на единицу аргумента X , если $f(X)$ — преобразуемая функция. (Общая теория преобразования Фурье и примеры ее приложений изложены в монографии Брейсуэлла [4].) В интерферометрии X выражает расстояние, тогда как в масс-спектрометрии X соответствует отношению массы к заряду. Таким образом, это преобразование можно рассматривать как частотный анализ исходного масс-спектра.

Другая трактовка преобразования Фурье выявляет все его преимущества при передаче информации в области преобразованного аргумента. Любая точка в этой области — взвешенная сумма всех точек исходной области, следовательно, данные о той или иной точке исходного пространства (координата которой есть отношение массы к заряду) распределяются по всему спектру в области Фурье. Об этой особенности преобразования Фурье говорят как о его усредняющем свойстве. Поэтому ошибка или потеря битов при передаче информации в области Фурье минимально искажают исходный спектр, тогда как потеря одного бита при передаче информации как функции аргумента m/e может привести к утрате всякого смысла такого спектра. Как будет показано ниже, эта особенность преобразования Фурье успешно используется при классификации образов.

Чтобы упростить анализ, введем некоторые обозначения. Как правило, результат преобразования Фурье действительных данных представляет комплексную величину, действительную часть которой образует косинусоидальная составляющая, а мнимую — синусоидальная составляющая разложения. Результат преобразования Фурье от функции $f(X)$ выражается в виде

$$G(v) = \int_{-\infty}^{\infty} f(X) e^{i2\pi v X} dX \quad (6.5)$$

или

$$G(v) = \int_{-\infty}^{\infty} f(X) \cos 2\pi v X dX + i \int_{-\infty}^{\infty} f(X) \sin 2\pi v X dX = \\ = G_c(v) + iG_s(v). \quad (6.6)$$

На практике функция $f(X)$ изменяется в конечных пределах и вне этих пределов считается равной нулю, поэтому интегрирование проводится на интервале конечной длины. Из формулы (6.6) легко получить еще две величины для представления данных в области Фурье: фазовый спектр

$$\Phi(v) = \operatorname{arctg} \frac{G_s(v)}{G_c(v)} \quad (6.7)$$

и спектр интенсивности

$$I(v) = [G_c(v)^2 + G_s(v)^2]^{1/2}. \quad (6.8)$$

Применительно к классификации образов преобразованные данные содержат ту же информацию, но в форме, упрощающей задачу создания классификатора или способствующей ее решению линейными методами. Теперь мы должны показать, как использовать быстрое преобразование Фурье (БПФ) при классификации образов по действительным данным (масс-спектрам низкого разрешения) и как применять усредняющее свойство этого преобразования для сокращения размерности.

Такое исследование было проведено на массиве данных из 630 масс-спектров низкого разрешения, заимствованных из таблиц Американского нефтяного института. Это были масс-спектры 387 углеводородов и 243 соединений типов C_nH_m , $\text{C}_n\text{H}_m\text{N}$ и $\text{C}_n\text{H}_m\text{ON}$ с небольшим молекулярным весом (меньше 200 а. е. м.).

Преобразование Фурье выполнялось при помощи программы SHARE (SDA 3465), разработанной Кули. Все 630 масс-спектров после преобразования записали на диск ЭВМ типа IBM 360/75 в межуниверситетском вычислительном центре, что позволяло провести их быстрое считывание. Программы обучения предусматривали стандартную обработку преобразованных масс-спектров. Для этих универсальных программ было необходимо только, чтобы исходные данные представлялись в векторной форме, а то, в каком именно пространстве отображены эти данные, не имело значения.

Были составлены четыре разные выборки данных по 630 образов в каждой. Компоненты таких образов рассчитывали разными

способами: 1) по косинусоидальной части разложения (действительная часть G); 2) по синусоидальной части разложения (мнимая часть G); 3) по фазовым спектрам и 4) по спектрам интенсивности.

Число точек при расчетах с помощью алгоритма БПФ должно быть равно 2^N , где N — целое положительное число. Поэтому 200 координат каждого спектра пришлось пополюнять еще 56 положениями с нулевыми пиками, чтобы довести общее число координат до 256. Алгоритм БПФ давал набор из 512 точек — по 256 для косинусоидальной и синусоидальной частей разложения. При работе с действительными данными важно то, что косинусоидальная часть является четной функцией, а синусоидальная — нечетной. Таким образом, в каждом случае вся информация заключается в 128 точках. Остальные 128 точек являются излишними, поскольку их без труда можно получить исходя из симметрии.

Массив данных состоял из четырех выборок по 630 образов и каждый образ имел 128 компонент.

Чтобы узнать, сохраняется ли основная информация в области Фурье, пришлось построить классификаторы образов для преобразованных данных и сравнить их с классификаторами, созданными для исходных спектров (интенсивности пиков как функции положения m/e). В обоих случаях категории, на которые подразделяли образы, были одинаковыми. Наиболее интересные результаты такого сопоставления приведены в табл. 6.3. В каждом случае положительная категория состояла из соединений с указанными в таблице значениями отношения числа атомов углерода к числу атомов водорода (C:H) в молекуле, а отрицательная — из всех прочих соединений. Исходный массив состоял из 387 углеводородов, из которых 200 случайно отобранных были включены в обучающую выборку, а остальные 187 — в контрольную. Наиболее важный результат проведенного исследования состоит в том, что преобразование Фурье оставляет основную информацию и что ее довольно легко извлекать из преобразованных данных. Скорость сходимости (выражаемая числом коррекций через обратную связь, необходимых для обеспечения полной сходимости) и прогнозирующая способность (процентная доля верных ответов при классификации неизвестных образов) сопоставимы с соответствующими показателями при классификации образов по исходным масс-спектрам.

Авторы рассматриваемого исследования подходили к оценке полезности обученных весовых векторов, руководствуясь двумя требованиями. Во-первых, если весовой вектор предназначен для замены обычных систем информационного поиска, то он должен достигать полной сходимости, свидетельствующей о линейной раз-

делимости, за приемлемое число коррекций через обратную связь. Во-вторых, обученный весовой вектор, если он предназначен для классификации неизвестных образов, должен характеризоваться достаточной прогнозирующей способностью.

Как выяснилось в ходе предшествующего исследования [5], пять из 43 категорий углеводов, для которых ожидалась линейная разделимость масс-спектров низкого разрешения, остались неразделенными за 2000 коррекций на обучающей выборке из 200 соединений. В табл. 6.4 сравниваются результаты пробной классификации для всех этих пяти категорий с применением к исходным данным преобразования Фурье и обычной классификации по непреобразованным масс-спектрам. Категории «этил», «*n*-пропил» и «винил» охватывают соединения, в состав которых входят эти группы. Категория «атомы углерода в боковой цепи» соответствует соединениям, в которых число связей углерод — углерод не меньше трех. Следует отметить, что при использовании фазовых спектров сходимость достигалась во всех случаях, тогда как классификация по данным в любой иной форме сходимости не обеспечивалась. Следовательно, применение преобразования Фурье к исходным данным позволяет получить ответы на некоторые вопросы, на которые непреобразованные исходные данные ответов не давали (по крайней мере в выбранных пределах сходимости). Этот результат удовлетворяет первому требованию, предъявляемому системам информационного поиска. Однако прогнозирующая способность во всех случаях была низкой, т. е. второе требование не выполнялось.

При решении задач классификации образов и разработке систем информационного поиска часто бывает целесообразно сокращать размерность исходных данных. Такое сокращение резко снижает потребности как в машинном времени, так и в объеме памяти. Информация, содержащаяся в той или иной точке координатной оси, на которой откладываются значения m/e , при преобразовании Фурье распределяется (в результате его усредняющего свойства) по всем позициям в области Фурье. Поэтому некоторые компоненты в области Фурье можно принять равными нулю или исказить как-то иначе, не исключая возможности восстановить обратным преобразованием исходный масс-спектр, но с более высоким уровнем шума. (В этом заключается одно из преимуществ передачи информации в области Фурье.) Задачи накопления информационных данных и их обработки, в ходе которой возможны ошибки и произвольный пропуск части данных, эквивалентны задачам передачи зашумленной информации. Поэтому образы, подвергшиеся преобразованию

Таблица 6.4

Результаты классификации в области Фурье для случаев, по которым не было сходимости при обычной классификации исходных спектров

Положительная категория	Число коррекций до сходимости				Доля правильных классификаций, %					
	масс-спектры	разложения			фазовые спектры	масс-спектры	разложения			
		косинусо-идальное (а)	синусо-идальное (б)	а) + б)			косинусо-идальное (а)	синусо-идальное (б)	а) + б)	
C=C	> 2000 ^а	> 10 000	> 10 000	> 10 000	160	78	73	75	72	75
Винил (CH ₂ =CH—)	> 2000	> 10 000	> 10 000	> 10 000	199	80	79	76	65	74
Этил	> 2000			> 10 000	645	73			62	67
н-Пропил	> 2000			> 10 000	422	72			60	81
Больше двух атомов углерода в боковой цепи	> 2000			> 10 000	217	62			62	66

^а После этого числа коррекций обучение прекращали, не достигнув сходимости.

Фурье, в некоторых случаях предпочтительны по сравнению с обыкновенными образами, например масс-спектрами.

В рассматриваемом исследовании размерность образов сокращали несколькими способами и строили классификаторы образов для масс-спектров и для полученной в результате преобразования Фурье косинусоидальной (действительной) компоненты. Задача состояла в отборе из совокупности углеводородов соединений с отношением числа атомов углерода к числу атомов водорода (С:Н) в молекуле, равным 1:2. Во всех случаях использовалась обучающая выборка из 200 спектров углеводородов, среди которых 81 соединение имело отношение С:Н, равное 1:2. Остальные 187 углеводородов, в том числе 74 соединения с отношением С:Н, равным 1:2, составляли контрольную выборку. Результаты этой части исследования обобщены в табл. 6.5. (Во всех случаях расчеты прекращали по истечении заранее установленного машинного времени.) В каждом случае данные после преобразования Фурье обеспечивали возможность более значительного сокращения размерности образов без существенного ущерба для времени сходимости или прогнозирующей способности. Результаты исключения масс-спектрометрических данных были различными. Из нескольких опробованных способов наиболее благоприятным было отбрасывание координат по критерию наименьшей средней величины, по крайней мере в отношении скорости сходимости до последнего этапа, когда оставалось 16 измерений (координат). В двух из трех случаев с использованием преобразования Фурье прогнозирующая способность фактически была несколько выше лучшего результата, достигаемого при классификации масс-спектрометрических данных. Таким образом, размерность данных после преобразования Фурье допускает значительное произвольное сокращение, прежде чем проявится ухудшение прогноза, тогда как существенное уменьшение размерности исходных масс-спектров возможно только в том случае, если оно производится на базе какого-нибудь логического критерия, например минимизации средней амплитуды пиков.

Дальнейшему исследованию возможностей использования преобразования Фурье для предварительной обработки масс-спектрометрических данных посвящена статья одного из авторов настоящей книги [6]; массив данных состоял из 450 масс-спектров низкого разрешения со 132 координатами (положениями m/e) в каждом.

Вычислительная процедура обработки исходных масс-спектров состояла из следующих операций. Поскольку исходные спектры имели пики до координаты m/e 200, размерность преобразуемых векторов образов была равна 200. Интенсивности пиков преобразо-

Таблица 6.5

**Сравнительные результаты сокращения размерности масс-спектров
и спектров Фурье тремя способами**

	Размерность	Критерий исключения					
		по максимальной величине		случайный выбор		по наименьшей средней величине	
		число коррекций до сходимости	прогнозирую- щая способ- ность, %	число кор- рекций до сходимости	прогнозирую- щая способ- ность, %	число кор- рекций до сходимости	прогнозирую- щая способ- ность, %
Спектры Фурье (ко- синусоидальная ком- понента)	128	75	95	75	95	75	95
	96	79	95	81	96	92	97
	64	162	98	103	95	147	98
	48	319	97	118	94	293	98
	32	2696	97	112	93	2088	97
	16	>8000	80	>8000	94	>8000	97
Масс-спектры	182	74	94	74	94	74	94
	96	116	96	142	98	74	94
	65	231	95	156	96	73	95
	48	>2666	90	781	94	82	95
	32	.		<4000	82	126	95
	16	.				575	94

ывали логарифмически. Преобразование всех спектров проводили с помощью алгоритма БПФ [7], что давало 256-мерный комплексный вектор. Оставляли только действительную часть результата преобразования Фурье. Поскольку 256 компонент действительной части этого вектора симметричны по отношению к середине вектора, сохраняли лишь первую их половину. Следовательно, образ после преобразования Фурье имел 128 компонент, что приблизительно равнозначно 132 исходным координатам (значениям m/e). Разумеется, масс-спектрометрические образы характеризуются гораздо большим числом компонент с нулевой интенсивностью, чем в образах после преобразования Фурье.

Первый этап рассматриваемого исследования заключался в отборе признаков для спектров, подвергшихся преобразованию Фурье, и оценке характеристик классификаторов образов в зависимости от числа оставленных дескрипторов. Как уже отмечалось, отбрасывание тех или иных дескрипторов для образов после преобразования Фурье скорее равнозначно частичной потере информации о всех исходных масс-спектрах, чем потере всей информации о части образов [8].

Результаты классификации образов, проведенной с отбором признаков для спектров после преобразования Фурье, представлены в табл. 6.6. Решалась химическая задача, в которой классификаторы обучали подразделять соединения на категорию с отношением числа атомов углерода к числу атомов водорода в молекуле больше 2 и категорию, для которой это отношение не больше 2. Таким образом, соединения с отношением $C:H > 2$ (алканы, амины и т. п.) образовывали одну категорию, а соединения с отношением $2(nC) \geq nH$ (алкены, кетоны, ароматические соединения и т. п.) входили в другую. Очевидно, что по масс-спектрам низкого разрешения такое разделение сразу провести невозможно.

В первой колонке табл. 6.6. приведены условные обозначения обучающих выборок в трех параллельных опытах, каждая из которых включала 250 спектров, случайно выбранных из исходного массива в 450 спектров. Остальные 200 спектров составляли во всех случаях контрольную выборку. Во второй, третьей, четвертой и пятой колонках указаны некоторые характеристики классификаторов образов до отбора признаков. В начальный момент времени любой образ после преобразования Фурье имеет 124 дескриптора, потому что первые четыре дескриптора, имеющие огромную величину, только затрудняют решение задачи. Обучение классификаторов проводили с порогом $Z = 10$. Для обучения классификаторов правильному распознаванию всех объектов обучающих выборок

требовалось от 600 до 1000 коррекций через обратную связь. На каждой обучающей выборке обучали по два бинарных классификатора. Точные числа коррекций, необходимых для обучения пары таких классификаторов, приведены в третьей колонке табл. 6.6: сначала указано число коррекций для одного весового вектора, а затем — для другого. В четвертой колонке таблицы указано число спектров, оставшихся неклассифицированными из-за того, что соответствующее скалярное произведение попадало в мертвую зону. В пятой — средняя прогнозирующая способность (доля правильных классификаций) для двух бинарных классификаторов; при переходе от одной обучающей выборки к другой она незначительно изменяется.

В колонках с шестой по девятую приведены характеристики классификаторов образов после интенсивного отбора признаков. Для трех обучающих выборок из первоначальных 124 дескрипторов оставалось только 76, 70 и 78 соответственно. Число коррекций через обратную связь, необходимых для обучения классификатора безошибочному распознаванию всех объектов обучающих выборок, либо оставалось приблизительно таким же, как и в предшествующем случае, либо было несколько меньше. Прогнозирующая способность классификаторов образов, действовавших по сокращенному перечню дескрипторов, была практически такой же, как и для спектров, подвергшихся преобразованию Фурье до отбора признаков. Следовательно, отбор признаков, ускоряя вычисления, несколько не ухудшает характеристики классификаторов образов.

В табл. 6.7 сопоставлены характеристики классификаторов образов, когда на их ввод в качестве векторов образов поступают исходные масс-спектры и масс-спектры после преобразования Фурье. Классификаторы обучали подразделять соединения на те же химические категории, что и в предыдущем случае (см. табл. 6.6). Данные в левой части таблицы показывают, что в случае обучения классификаторов на масс-спектрометрических образах, содержащих после отбора признаков по 111 координат, для полного распознавания требовалось от 1200 до 2000 коррекций через обратную связь при прогнозирующей способности 93—95% на выборках из совершенно неизвестных объектов. В правой части таблицы помещены данные о характеристиках классификаторов образов, обучавшихся на спектрах, подвергнутых преобразованию Фурье. Последние содержали по 76, 70 и 78 дескрипторов на образ в трех обучающих выборках соответственно. Здесь для сходимости требовалось 600—1000 коррекций, т. е. приблизительно в два раза меньше, чем в предыдущем случае. Во всех трех случаях прогно-

зирующая способность стала выше (увеличение составило соответственно 0,4, 0,4 и 1,0%). Таким образом, классификаторы, анализирующие спектры Фурье после отбора признаков, по своим характеристикам выгодно отличаются от классификаторов, на вход которых поступают исходные (непреобразованные) данные.

В рассматриваемом исследовании была проведена также проверка надежности пороговых логических элементов при обучении на исходных масс-спектрометрических данных и на спектрах Фурье. Как выяснилось, более высокой надежностью обладают бинарные классификаторы образов при обучении на спектрах Фурье.

Алгоритм быстрого преобразования Фурье был применен при решении задачи расшифровки спектров ЯМР методами распознавания образов [9]. В этом исследовании в качестве массива данных использовались автокорреляционные функции моделированных спектров ЯМР. Автокорреляция устраняет трансляционную дисперсию спектров и делает их более подходящими для расшифровки методами распознавания образов. Автокорреляционная функция $A(x)$ функции $F(t)$ записывается в виде

$$A(x) = \int F(t) F(t+x) dt. \quad (6.9)$$

Функция $F(t)$ описывает непрерывный спектр ЯМР. Автокорреляционную функцию можно приближенно представить в виде ряда и вычислить непосредственно по определяющей формуле. Однако расчеты удобнее проводить, если эту функцию вычислять по этапам в следующей последовательности: 1) выполнить преобразование Фурье для функции $F(t)$, чтобы найти функцию $G(x)$; 2) умножить функцию $G(x)$ на комплексно сопряженную ей функцию, чтобы получить спектр мощности $|G(x)|^2$ исходной функции; 3) обратным преобразованием Фурье по найденному спектру мощности $|G(x)|^2$ найти автокорреляционную функцию. Эту процедуру из трех этапов удобно выполнять при помощи алгоритма БПФ. В задачу данного исследования входила проверка возможности сокращения таким путем размерности векторов образов, т. е. возможности отбора признаков.

Исследование проводилось на линейчатых спектрах, вычисленных по химическим сдвигам и постоянным взаимодействия, для которых рассчитывали автокорреляционные функции. Массив данных состоял из 634 образов, каждый из которых имел 236 координат. Несколько линейных разделяющих функций обучали классификации соединений по такой структурной особенности, как при-

существование или отсутствие n -пропильной группы. Результаты этого обучения оказались очень хорошими.

Предварительную обработку масс-спектрометрических данных пробовали также осуществлять путем преобразования Адамара [10]. Это преобразование аналогично преобразованию Фурье, но разлагает функции не на синусоидальные, а на «квадратно-волновые» компоненты. Классификацию осуществляли по масс-спектрам низкого разрешения, предварительно подвергнутым преобразованию Адамара. Категории определяли по числу атомов углерода в молекулах углеводов (шесть, семь или восемь).

ФАКТОРНЫЙ АНАЛИЗ [11]

К дополнительным методам преобразования исходных данных следует отнести также факторный анализ. Основу этого метода составляет диагонализация корреляционной матрицы для нахождения ее собственных значений, по которым производится оценка «важности» соответствующих собственных векторов. При этом предполагается, что исходные данные и классы, на которые эти данные подразделяются, связаны друг с другом через дисперсию (среднее отклонение) распределения вероятностей.

В последнее время факторный анализ успешно применяется при изучении основных свойств растворенных веществ и неподвижных фаз по временам удерживания в газохроматографических исследованиях [12, 13], растворов по химическим сдвигам в спектроскопии ЯМР [14—16], а также в полярографии [17].

Цель факторного анализа — заменить исходное множество таким набором меньшего числа переменных, который должным образом описывал бы или выражал бы исходное множество. Что же именно следует понимать под «должным образом», устанавливает экспериментатор. Возможно, что для этого достаточно воспроизвести данные в пределах ошибки эксперимента или не слишком строго отобразить главные отклонения, пренебрегая более слабыми эффектами.

Анализ начинают с матрицы наблюдений D , например с набора масс-спектров. Исходные экспериментальные данные нормируют путем вычитания средней интенсивности и последующего деления на среднеквадратичное отклонение по каждому положению m/e . Из нормированных данных вычисляют корреляционную матрицу C , которая отображает связь отклонений от среднего в каждой позиции со всеми остальными позициями:

$$C = D^T D. \quad (6.10)$$

Корреляционную матрицу $m \times m$ можно преобразовать в корреляционную матрицу в ортогональной системе координат, составленную из линейных комбинаций коэффициентов корреляции в матрице C с таким расчетом, чтобы по возможности неравномернее распределить дисперсию, которую отображает матрица C . Иными словами, первая координатная ось, т. е. линейная комбинация, должна «вмещать» столько дисперсии, сколько возможно для единственной оси. Вторая линейная комбинация должна «содержать» вторую по величине дисперсию с учетом того ограничения, что вторая ось ортогональна первой. Линейные комбинации, выражающие максимальную дисперсию, составляют до тех пор, пока не будет «исчерпана» вся дисперсия. Эту ортогональную систему координат находят диагонализацией корреляционной матрицы, чтобы определить набор собственных значений E и набор сопряженных собственных векторов B :

$$B^{-1}CB = E. \quad (6.11)$$

Квадратные корни из собственных значений соответствуют средне-квадратичным отклонениям, а сами собственные значения — дисперсиям по осям сопряженных собственных векторов. Таким образом, корреляционную матрицу можно получить перестановкой членов уравнения (6.11):

$$C = BEB^{-1} = BEB^T. \quad (6.12)$$

Поскольку B — ортогональная матрица, обратная ей матрица образуется при транспонировании, следовательно, уравнение (6.12) можно записать в виде

$$C = e_1 b_1 b_1^T + e_2 b_2 b_2^T + \dots + e_m b_m b_m^T. \quad (6.13)$$

Однако, по мере того как собственные значения e_i приближаются к нулю, включение сопряженных собственных векторов в приближенное выражение для корреляционной матрицы дает все меньший и меньший вклад. Задача уменьшения числа собственных векторов, требующегося для воспроизведения корреляционных связей в исходных переменных, эквивалентна задаче уменьшения размерности исследуемого пространства изображений, и поэтому сжатие информации есть следствие выбора для представления случайных изменений более оптимальной системы координат, нежели исходная.

Можно лучше понять смысл записи корреляционной матрицы через собственные векторы с учетом важности изменений интенсивности линий в масс-спектре, если оси собственных векторов так

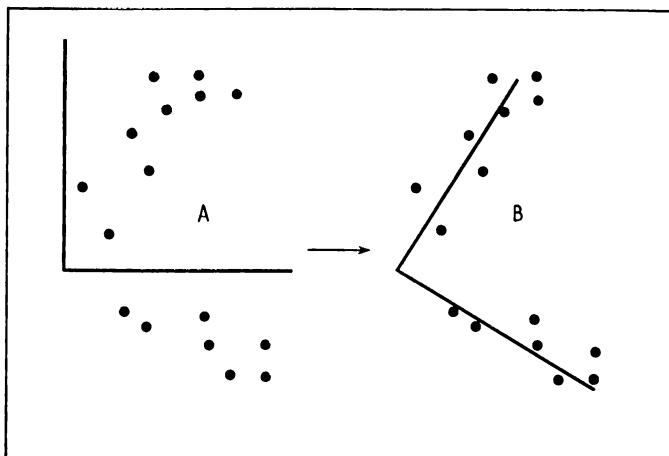


Рис. 6.3. Множество точек в двух системах координат, различающихся только поворотом осей.

повернуть в пространстве, чтобы плотность распределения собственных векторов была максимальной. Этот процесс иллюстрируется на рис. 6.3. Обе системы координат А и В характеризуют расположение точек, однако система В для расшифровки данных проще, поскольку здесь векторы в пространстве становятся больше по величине, но число их уменьшается,

В рассматриваемом исследовании для упрощения интерпретации систему координат поворачивали ортогонально, а попыток осуществлять поворот на такие важные векторы данных, как потенциалы ионизации или возбуждения, не предпринималось. После поворота удалось легко определить значащие массы в каждом факторе (фактор — собственный вектор, отнесенный к корню квадратному из сопряженного с ним собственного значения). Так, наибольшие коэффициенты в 26-м факторе принадлежали массам 45, 31 и 19. Поскольку эти массы обычно ассоциируются с кислородом, можно полагать, что 26-й фактор коррелирует с наличием кислорода в составе изучаемого соединения. Чтобы проверить, в какой степени данный

фактор (равно как и другие) коррелирует с наличием в молекулах определенной функциональной группы, был использован метод упорядочения факторов по отношению к функциональным группам, предполагающий вычисление скалярного произведения фактора и каждого масс-спектра с суммированием произведений P_{jk} по всем масс-спектрам класса. Это было сделано для каждого фактора по формуле

$$P_{jk} = \frac{1}{n} \sum_{i=1}^{n_h} F_j S_i, \quad (6.14)$$

где F_j — j -й фактор и S_i — i -й компонент в нормированном масс-спектре для класса k . Нормированные масс-спектры вычисляли по формуле

$$S = \frac{I - \bar{I}}{\sigma}, \quad (6.15)$$

где I — интенсивность пика, \bar{I} — средняя интенсивность пика для данного значения m/e и σ — соответствующее среднеквадратичное отклонение. Результаты расчетов по формуле (6.14) располагали в ряд по убыванию абсолютной величины и выявляли наиболее значимые массы в наибольших по абсолютной величине факторах по отношению ко всем функциональным группам.

В табл. 6.8 приведены результаты расчетов по формуле (6.14) для семи функциональных групп. В случае карбонильной группы ($C=O$) наибольшую и приблизительно одинаковую значимость имеют два фактора — 18-й и 34-й, тогда как для гидроксильной группы самыми важными оказались 26-й, 40-й и 41-й факторы. (Второстепенные факторы отличаются от главных гораздо меньшим значением скалярного произведения, см. третью колонку табл. 6.8.) 26-й фактор имеет значимость и для простых эфиров; это указывает на то, что в простом эфире и гидроксиде фрагменты образуются аналогичными путями, чего нельзя сказать о карбонильных группах.

Изменение интенсивности пиков в масс-спектрах 81 нитросоединения выявило связь с двумя факторами — 10-м и 24-м, тогда как для 58 аминов из массива данных наиболее важными были четыре фактора — 10-й, 4-й, 19-й и 37-й. Поскольку среди значимых для аминов факторов нет 24-го, его следует считать характерным для нитросоединений.

Таблица 6.8

Связь функциональных групп с факторами для семи химических классов

Функциональная группа	Номер фактора	Значимость фактора [формула (6.14)] ^а	Значимые массы в факторе
C ₆ H ₅	15	—1,24	120, 105
	8	1,19	134, 133, 119
	21	—0,86	92, 91
	1	—0,65	117, 116, 90, 89
	31	0,58	129, 102
	32	0,50	122, 77
C=O	18	0,65	88, 61, 60
	34	—0,63	74, 73, 28
C—O—C	26	—0,80	73, 45, 31, 19
	9	—0,64	87, 75, 59, 47
OH	26	—1,60	73, 45, 31, 19
	40	—1,00	59, 33, 31
	41	—0,94	62, 32, 31
N	10	—0,51	123, 108, 18, 17
	13	—0,47	58
	24	0,40	76, 61, 46, 30, 17
NH ₂ , NH	10	—0,72	123, 108, 18, 17
	4	—0,62	98, 83, 70, 69, 56, 55, 42, 41, 39, 27
	19	0,55	72, 44
	37	0,53	96, 95, 94, 93
C _n H _{2n+2}	16	1,31	113, 99, 71, 57, 54
	23	—0,62	57
	25	—0,56	85, 84, 43

^а Отрицательное значение не обязательно означает, что связанные с факторами массы отрицательно коррелируют с функциональной группой. Если значения отрицательны, то двойное отрицание приводит к положительной корреляции между функциональными группами и массами.

Для насыщенных углеводородов наиболее значимым является 16-й фактор, тогда как 23-й и 25-й играют меньшую роль. Поскольку суммы произведений факторов на нормированные спектры делят на число спектров в соответствующем классе, появляется возможность оценки факторов по их отношению к *любой* функциональной группе. Так, среди всех факторов для всех групп по значимости наиболее важен 26-й фактор, который для гидроксильных групп в два раза более существен, чем для простых эфиров. Действительно, приведенные в третьей колонке табл. 6.8 данные свидетельствуют о том, что все три «гидроксильных» фактора связаны с гидроксильными группами в большей степени, чем каждый из «эфирных» факторов с эфирами. Отсюда можно сделать вывод, что фрагментация простых эфиров протекает легче, чем гидроксильных групп, тем более, что в эфирах атом кислорода имеет более разнообразное структурное окружение.

10-й фактор, вообще говоря, важнее для аминов, чем для нитросоединений (0,72 и 0,51 соответственно), так что значимые для него массы приписывают аминам.

Если для данной функциональной группы известны важные факторы, положения m/e для нее находят непосредственно как массы с наибольшими коэффициентами в факторе. Эти коэффициенты равны корням квадратным из дисперсии в том положении, которое можно приписать данному фактору.

Факторный анализ позволяет найти ответ на вопрос: как присутствие функциональных групп отражается на масс-спектрах. Для этого было проведено суммирование по формуле (6.14) произведений для 42 факторов по каждой из 35 функциональных групп или структурных параметров. Масс-спектры имели форму нормированных данных, поэтому каждый спектр можно рассматривать как направленное расстояние до среднего для всех спектров. Факторы тоже имеют смысл направленных расстояний до среднего значения данных, найденных как ортогональные координаты максимального направленного расстояния до среднего (значения). Таким образом, произведение спектра на фактор служит мерой сходства двух векторов по направлению и расстоянию. Суммирование этих величин по всем спектрам того или иного класса позволяет сделать вывод о степени родства (связанности) между соединениями данного класса и осью изменения данных. Суммирование же по всем направлениям этого изменения позволяет определять, как данная функциональная группа влияет на распределение фрагментов в масс-спектрометрических образах. Приведенные в табл. 6.9 данные показывают, какие структурные особенности наиболее

Таблица 6.9

Соотношения между функциональными группами и факторами

Функциональная группа или структурный параметр	Число спектров	Сумма произведений в формуле (6.14)
1. Фенильная группа	62	12,88
2. Не меньше 3 двойных связей	75	11,82
3. Амин	58	11,66
4. Азот	81	9,60
5. Не меньше 2 двойных связей	108	9,48
6. C_nH_{2n+2}	89	9,83
7. Тройная связь	42	8,65
8. Не меньше 2 атомов кислорода	86	8,62
9. $C-O-C$	57	8,59
10. OH	33	8,25
11. Атомы углерода	86	7,71
12. Атомы углерода, не связанные с водородом	103	7,35
13. Больше 4 метильных групп	106	7,15
14. Не меньше 9 атомов углерода	184	6,48
15. Кислород	174	6,29
16. C_nH_{2n}	154	6,24
17. Карбоциклическая группа	76	6,13
18. Не меньше 15 атомов водорода	205	5,69
19. Не меньше 2 точек разветвления	189	5,60
20. Винильная группа	75	5,52
21. Не меньше 8 атомов углерода	273	5,05
22. Не меньше 3 метильных групп	220	4,86
23. Не меньше 1 двойной связи	241	4,74
24. Не меньше 2 этильных групп	107	4,54
25. Не меньше 1 точки разветвления	309	4,30
26. n -Пропильная группа	141	4,00
27. Не меньше 13 атомов водорода	313	3,97
28. Не меньше 7 атомов углерода	351	3,95
29. Атомы углерода, не связанные с водородом	285	3,20
30. Не меньше 1 этильной группы	287	3,00
31. Не меньше 11 атомов водорода	405	2,84
32. Не меньше 6 атомов углерода	447	2,59
33. Не меньше 2 метильных групп	420	2,41
34. Не меньше 9 атомов водорода	495	1,94
35. Не меньше 1 метильной группы	543	1,12

сильно сказываются на распределении фрагментов в исследованных 630 масс-спектрах.

Присутствие фенильной группы наиболее сильно влияет на масс-спектры. Второе место среди значимых функциональных групп занимает азот, а третье — структуры насыщенных углеводо-

родов. Другими словами, наличие этих трех групп смещает спектры на значительное расстояние от спектра, усредненного по всему массиву данных, тогда как наличие в молекуле одной метильной группы мало изменяет спектры по сравнению с усредненным спектром.

Из всех функциональных групп самой сильной зависимостью с факторными осями характеризуется фенильная группа, затем следуют такие параметры, как наличие до трех двойных связей (охватывающее и фенильную группу), амины, азот, наличие до двух двойных связей и структуры насыщенных углеводородов.

КОМПЛЕКСНАЯ НЕЛИНЕЙНАЯ РАЗДЕЛЯЮЩАЯ ФУНКЦИЯ [18]

Комплексная нелинейная разделяющая функция есть результат применения обобщенного преобразования Уолша для разбиения пространства признаков на области решений [19]. В случае обобщенного преобразования Уолша первого порядка, при котором спектральные интенсивности могут принимать 50 целочисленных значений в диапазоне 0—49, каждое измерение (координата) образа, т. е. спектра, преобразуется по формуле

$$T(I) = \left(e^{2\pi \sqrt{150}} \right)^I, \quad (6.16)$$

где I — интенсивность, соответствующая каждому положению m/e преобразуемого спектра. Тогда $\Phi(x)$ есть вектор, характеризующий результат преобразования всех интенсивностей (измерений) в масс-спектре, т. е.

$$\Phi(x) = T(I_1), T(I_2), \dots, T(I_n). \quad (6.17)$$

Используя $\Phi(x)$, можно построить разделяющую функцию вида

$$F(x) = 0 + W^* \Phi(x). \quad (6.18)$$

Здесь W^* — вектор, комплексно сопряженный вектору W , а последний представляет сумму всех векторов $\Phi(x)$ для спектров обучающей выборки, т. е.

$$W = \frac{W_A}{a} - \frac{W_B}{b}, \quad (6.19)$$

где a и b — числа спектров соответственно в категориях А и В, а W_A и W_B — компоненты весового вектора, определяемого вектор-

ным суммированием преобразованных спектров соединений категорий А и В. Поскольку W — вектор, сопряженный ему вектор W^* получают переменной знака мнимой части на обратный, поэтому здесь не нужна какая-то новая операция, и в матричном обозначении все остается в прежнем виде. Постоянная θ связана с относительными количествами и дисперсиями спектров для категорий А и В, образующих обучающую выборку. Таким образом, функция $F(x)$ выражается комплексным числом и нелинейна по отношению к компонентам масс-спектра. При прогнозировании соединения, которым соответствуют положительные действительные части $F(x)$, относят к одной категории, а соединения с отрицательными такими частями — к другой.

Как видно из формулы (6.18), чтобы применить разделяющую функцию, нужно вычислить $\Phi(x)$ и W^* . Преобразованные спектральные интенсивности могут принимать только 50 значений [см. соотношение (6.16)], поэтому эти интенсивности можно вычислить и хранить до использования в расчетах решающей поверхности, а не вычислять каждый раз для любого нового спектра. Таким путем достигается большая экономия машинного времени. Вектор W вычисляют непосредственно по формуле (6.19).

Чтобы составить комплексную нелинейную разделяющую функцию, предназначенную для классификации соединений по их спектрам, берут соединения двух классов и преобразуют их масс-спектры по формуле (6.16). Считая один класс соединений обучающей выборки положительным, а другой отрицательным, алгебраическим сложением преобразованных спектров находят весовой вектор W , который затем можно использовать для предсказания по формуле (6.18) категории соединений, не вошедших в обучающую выборку. Если, например, отрицательный класс состоит из соединений с молекулярной формулой C_nH_{2n} , а положительный охватывает все прочие соединения, то соединение, для которого расчетное значение $F(x)$ отрицательно, считается принадлежащим классу соединений с молекулярной формулой C_nH_{2n} . В табл. 6.10 приведены данные прогнозирования для 630 соединений с разбиением на указанные в этой таблице категории.

Общую прогнозирующую способность (долю верных предсказаний) определяли путем обучения на всех 630 соединениях, вычитали вклад прогнозируемого соединения в весовой вектор и пытались отнести это соединение к должной категории. Затем такой вклад снова прибавляли к весовому вектору и вычитали вклад следующего соединения. Этим способом была произведена классификация всех 630 соединений. Таким образом, обучающая выборка

Таблица 6.10

Прогнозирующая способность комплексной нелинейной разделяющей функции

	Порог	Положительная категория ^а	Отрицательная категория	Доля соединений в наиболее обильной категории, %	Значения θ	Прогнозирующая способность, %	
						без нормировки	с нормировкой
Кислород	1	456	174	72,4	—0,622	87,6	88,3
	2	544	86	86,4	—0,326	90,0	90,0
Карбонильная группа	1	554	76	87,9	—1,614	87,9	88,7
Азот	1	549	81	87,1	—1,111	88,3	90,3
Амин	1	572	58	90,8	—1,422	91,4	92,9
—C=C—	1	389	241	61,8	—0,429	80,0	82,5
	2	522	108	82,9	—1,466	94,9	95,2
	3	555	75	88,1	—2,014	98,3	89,3
C_nH_{2n}	—	476	154	75,6	—1,955	91,6	96,8
C_nH_{2n+2}	—	541	89	86,0	—1,777	95,9	95,6
Метил	1	87	543	82,8	1,022	87,5	88,7
Этил	1	341	287	54,5	—0,148	71,4	77,1
	2	523	107	83,1	—0,340	86,5	86,0
Фенил	1	568	62	90,2	—2,311	96,5	96,8
Углерод	5	105	525	83,4	0,458	91,6	92,1
	6	183	447	71,0	0,177	84,0	86,4
	7	279	351	55,7	0,088	77,3	85,2
	8	357	273	56,7	0,014	84,3	88,1
	9	446	184	70,8	0,192	86,7	85,4
	10	544	86	86,4	0,177	90,5	90,5
Водород	9	135	495	78,6	0,311	85,6	87,8
	11	225	405	64,3	0,148	80,2	81,6
	13	317	313	50,3	—0,888	78,9	77,8
	15	425	203	67,5	—0,800	85,6	84,0
	17	501	129	79,5	—0,666	87,8	87,0
	19	554	76	88,0	—0,844	92,1	92,1

^а Положительная категория охватывает соединения, в которых число функциональных групп меньше порога.

состояла из 629 соединений. Одновременно определяли постоянную θ прогнозированием в некотором диапазоне малых приращений, прибавляемых к произведению $W^* \Phi(x)$. Эти приращения откладывали на графике, вычерчивали в зависимости от них процентную долю верных классификаций и максимум кривой считали соответствующим оптимальному значению θ . Полученные оптимальные значения θ для соответствующих категорий приведены в табл. 6.10.

В этой таблице приведены также результаты классификации соединений по преобразованным исходным спектрам без нормировки и с нормировкой их сумм. Нормировка заключалась в том, что сумму интенсивностей пиков для каждого соединения считали равной 100 и исходя из этого пересчитывали интенсивности всех пиков; интенсивность выше 49 полагали равной 49. Повышение прогнозирующей способности, обусловленное такой нормировкой, связано, по-видимому, с влиянием отдельных соединений на весовой вектор.

В табл. 6.10 указана процентная доля соединений в более обширной категории по каждой классификации. Если какое-то соединение каждый раз попадает в более обширную категорию, то прогнозирование можно вести на уровне, равном процентной доле данной категории. Таким образом, эта величина служит показателем того, насколько разделяющая функция обучилась классификации рассматриваемых категорий. Чем больше разница между прогнозирующей способностью и долей более обширной категории, тем лучше разделяющая функция отличает соединения одного класса от соединений другого. Например, комплексная нелинейная разделяющая функция незначительно повышает прогнозирующую способность в отношении карбонильной группы (от 87,9 до 88,7%), однако при обнаружении двойных связей и определении их числа увеличение прогнозирующей способности было гораздо большим.

Доля верных предсказаний может служить критерием доверия к отдельному прогнозу. Так, правдоподобность прогноза о наличии фенильной группы (96,8%) выше, чем правдоподобность прогноза о наличии азота (90,3%). Этот критерий можно дополнительно усовершенствовать. На рис. 6.4 показан график распределения соединений двух классов (соединений с формулой C_nH_{2n} и всех других соединений) в зависимости от значения $F(x)$. Когда значение этой функции находится очень близко к решающей поверхности, «доверие» к отдельному предсказанию минимально. По мере удаления $F(x)$ от решающей поверхности доверие к прогнозам возрастет. Это обстоятельство особенно ценно в тех случаях, когда в отношении одного соединения составляются разнообразные прогнозы. Возникающие при этом противоречия можно разрешить, взяв

предсказание с высшей вероятностью правильности, т. е. тот прогноз, для которого промежуток между $F(x)$ и решающей поверхностью максимальный. Как показывает соответствующее распределение (см. рис. 6.4), это еще, разумеется, не гарантирует полную правильность (безошибочность) подобного решения.

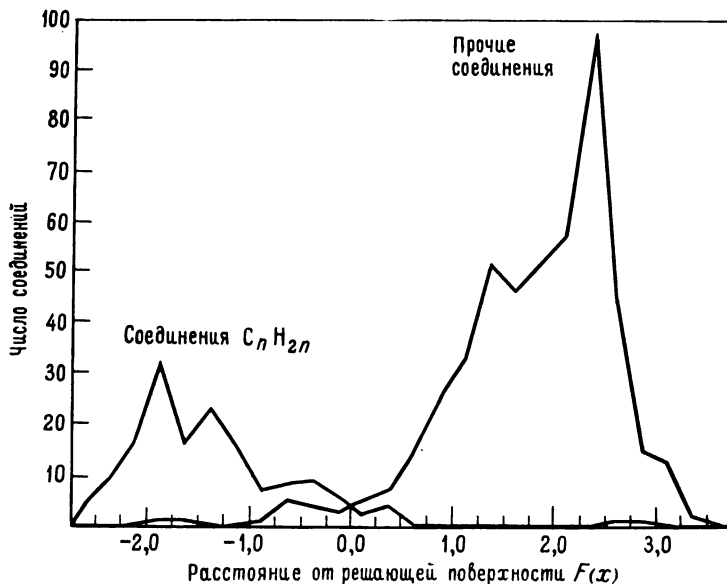


Рис. 6.4. Распределение 630 соединений как функция расстояния до решающей поверхности. Решающая поверхность отделяет соединения $C_n H_{2n}$ от соединений другого состава.

Комплексную нелинейную разделяющую функцию проверяли также на действительной, мнимой и фазовой частях масс-спектров после преобразования Фурье. Полученные результаты оказались сопоставимыми, за исключением фазовой части, для которой прогнозирующая способность была в большинстве случаев ниже (на 1—8%). Однако доля верных предсказаний по фазовой части составила 92,2% для азота и 95,1% для аминов, что представляет значительное достижение. Улучшение классификации по присутствию азота в составе молекулы можно объяснить эффектом «нечетной» массы атомов азота, который выражается в том, что пики,

относящиеся к фрагментам с азотом, не совпадают по фазе с пиками, относящимися к соединениям, не содержащим азот.

СПИСОК ЛИТЕРАТУРЫ

1. Nilsson N. J., Learning Machines. Foundations of Trainable Pattern-Classifying Systems, McGraw-Hill, New York, 1965.
2. Jurs P. C., Appl. Spectrosc., **25**, 483 (1971).
3. Wangen L. E., Frew N. M., Isenhour T. L., Jurs P. C., Appl. Spectrosc., **25**, 203 (1971).
4. Bracewell R., The Fourier Transform and Its Application, McGraw-Hill, New York, 1965.
5. Jurs P. C. et al., Anal. Chem., **42**, 1387 (1970).
6. Jurs P. C., Anal. Chem., **43**, 1812 (1971).
7. Cooley J. W., Tukey J. W., Math. Comput., **19**, 297 (1965).
8. Boulton P. I., Davison E. J., Lang G. R., Symposium on Feature Extraction and Selection in Pattern Recognition, ANL, Argonne, Ill., Oct. 1970.
9. Kowalski B. R., Reilley C. A., J. Phys. Chem., **75**, 1402 (1971).
10. Kowalski B. R., Bender C. F., Anal. Chem., **45**, 2234 (1973).
11. Justice J. B., Ph. D. Thesis, Department of Chemistry, University of North Carolina, Chapel Hill, N. C., 1974.
12. Weiner P. M., Howery D. G., Anal. Chem., **44**, 1189 (1972).
13. Weiner P. M., Howery D. G., Can. J. Chem., **50**, 448 (1972).
14. Weiner P. M., Malinowski E. R., Levinstone A. R., J. Phys. Chem., **74**, 4537 (1970).
15. Weiner P. M., Malinowski E. R., J. Phys. Chem., **75**, 1207 (1971).
16. Weiner P. M., Malinowski E. R., J. Phys. Chem. **75**, 3160 (1971).
17. Howery D. G., Bull. Chem. Soc. Japan., **45**, 2643 (1972).
18. Justice J. B., et. al., Anal. Chem., **44**, 2087 (1972).
19. Uesaka Y., IEEE Trans., SMC-1, 194 (1971).

ОТ МОЛЕКУЛЯРНОЙ СТРУКТУРЫ К СВОЙСТВАМ

Задачи интерпретации химических данных, рассмотренные в предыдущих главах, были связаны с установлением химической структуры из определенных экспериментально величин, причем в основном использовались результаты спектроскопических исследований.

Теперь мы обсудим обратную задачу: определение физических свойств соединений непосредственно из их молекулярной структуры. Эту задачу также можно решить общими методами распознавания образов, которые подробно были рассмотрены ранее. В этом случае необходимо закодировать молекулярные структуры исследуемых соединений в виде векторов образов такого формата, который был бы согласован с возможностями пороговых логических элементов. Решение поставленной нами сейчас задачи относится к области обработки информации о химической структуре.

Именно таким путем изучены физические свойства соединений нескольких типов. Наиболее широко исследовано формирование масс-спектров небольших органических молекул, моделированных непосредственно по соответствующим образом закодированной молекулярной структуре.

ФОРМИРОВАНИЕ МАСС-СПЕКТРОВ [1—3]

Расчетное прогнозирование масс-спектров было основано на двух подходах — с использованием теории квазиравновесия и машинных вычислений по эвристической программе DENDRAL. Разработка теории квазиравновесия базировалась на обоснованном экспериментально предположении, согласно которому образование фрагментов ионов, обуславливающее масс-спектры, можно рассматривать как процесс, который протекает подобно обыкновенной химической реакции с определенной скоростью. Были созданы интересные модели, позволяющие рассчитывать масс-спектры простых молекул [4—6]. В теории квазиравновесия молекулярные

процессы, которые проявляются в масс-спектре, представляются рядом конкурирующих стадий разложения исходных ионов молекулы, находящихся в возбужденном состоянии. Кроме того, считается, что константы скоростей подобных процессов можно вычислить на основе теории абсолютных скоростей реакций. Скорость реакции есть функция внутренней энергии участвующего в ней иона, поэтому, чтобы охарактеризовать распределение внутренней энергии, нужно знать функции, описывающие перенос энергии между ионами. Необходима также информация о путях реакции, энергии активации, параметрах активированных комплексов и электронных состояниях. Все это позволяет построить графики зависимости относительного распределения разных ионов от энергии. Теория квазиравновесия была успешно применена в случае углеводородов и простых соединений с одной функциональной группой.

Подпрограмма PREDICTOR, являющаяся частью эвристической программы DENDRAL, способна прогнозировать главные особенности масс-спектров ациклических органических соединений [7]. Эта подпрограмма опирается на подтвержденные экспериментально теоретические механизмы фрагментации и позволяет предсказывать масс-спектры алифатических кетонов, простых эфиров и аминов, а также давать общее решение для алифатических соединений $C_nH_{2n+v}X$ (X — атомы кислорода, азота или серы, v — валентность X).

В задачу исследования входило формирование масс-спектров низкого разрешения для молекул органических соединений методами распознавания образов исходя из их молекулярной структуры. Каждый бинарный классификатор образов обучали прогнозировать наличие или отсутствие того или иного пика в масс-спектре, соответствующего каждому из 60 положений m/e . Кроме того, вся система давала некоторые сведения об интенсивности пиков для 11 из этих положений.

Массив исходных данных был взят из сводного каталога масс-спектров, записанных на магнитной ленте, в Центре масс-спектрометрических данных при Управлении атомной энергии Англии. На этой ленте был записан также 2261 спектр из таблиц Американского нефтяного института. Было выбрано 600 масс-спектров низкого разрешения соединений состава $C_{3-10}H_{2-22}O_{0-4}N_{0-2}$. После цифрового преобразования интенсивности всех спектров были переведены в интервал 0,01—99,99% путем отнесения к самому высокому пику в спектре. Спектры нормировали с таким расчетом, чтобы полный ионный ток для всех спектров был одинаковым. Таким

образом, 9,5%-ная интенсивность какого-то пика означает, что она соответствует 9,5% полного ионного тока в данном спектре.

Обычно 150 спектров, случайно выбранных из 600, включали в обучающую выборку, а остальные 450 — контрольную.

Кодирование фрагментов

Первое требование при создании бинарного классификатора образов состоит в адекватном представлении образов (молекулярных структур) в подходящем для вычислительной машины виде. Разработано несколько методов подобного описания химических структур; в ряде обзоров излагаются различные приемы такого представления и их особенности (см., например, работу [81]). Разумеется, чем однозначнее описание, тем вероятнее успех использования бинарных классификаторов образов и тем совершеннее моделирование масс-спектров.

Для решения поставленной задачи был выбран такой способ кодирования фрагментов, который предполагает описание соединения как комбинации его основных фрагментов и соотношений между ними. Затем этим признакам приписывают численные дескрипторы. Преимущества данного способа кодирования заключаются в том, что им легко овладеть: он понятен каждому, дает перечень линейных дескрипторов, которые можно вводить в вычислительную машину непосредственно без какой бы то ни было предварительной обработки, требует малого объема машинной памяти. Однако за простоту данного способа приходится расплачиваться не совсем полным описанием молекулярной структуры. Теряется информация о том, какие фрагменты молекулы связаны друг с другом или какой атом фрагмента связан с другим его атомом, т. е. информация о геометрии и стереохимии исследуемой молекулы.

Бинарный классификатор образов создают в несколько последовательных этапов. Сначала из массива данных выбирают ту или иную молекулу (соединение). Затем, исходя из ее пространственной структуры, химик набрасывает от руки плоскую (двумерную) структуру данной молекулы. Третий этап предполагает контрольное сопоставление такой двумерной схемы с перечнем заранее выбранных дескрипторов. Завершив кодирование всех соединений в виде векторов образов, приступают к построению бинарного классификатора образов по принципу обучающейся машины. Если образы обучающей выборки удовлетворяют требованию линейной разделимости, то при помощи обыкновенной программы отбора

Таблица 7.1

Дескрипторы молекулярной структуры

Дескриптор	Тип	Нормировочный множитель	Ограничения
1. Молекулярный вес	Ц	0,05	Немоноциклические соединения
2. Наибольшая длина цепи	Ц	1,00	
3. Наибольшее кольцо	Ц	1,00	
4. Число атомов углерода	Ц	1,00	
5. Число атомов водорода	Ц	0,50	
6. Число атомов кислорода	Ц	3,00	
7. Число атомов азота	Ц	5,00	
8. Число колец+число двойных связей	Ц	1,00	
9. Простой эфир	Б	5,00	
10. Сложный эфир	Б	5,00	
11. Кетон	Б	5,00	Только моноциклические соединения
12. Спирт	Б	5,00	
13. Карбонильная группа	Б	5,00	
14. Кислородная связь	Б	5,00	
15. Гидроксильная группа	Б	5,00	
16. Винильная концевая группа	Б	5,00	
17. Ароматическое соединение	Б	5,00	
18. Наличие бензольного кольца	Б	5,00	
19. Одно бензольное кольцо	Б	5,00	У бензольных колец три
20. Гетероатом в кольце	Б	5,00	
21. Число связей $C=C$	Ц	2,00	
22. Число связей $C\equiv C$	Ц	5,00	
23. Ациклическое соединение (колец нет)	Б	5,00	
24. Число атомов углерода в точках разветвления цепи	Ц	2,00	
25. Число цепей наибольшей длины	Ц	3,00	
26. Число непарных атомов водорода	Б	5,00	
27. Число <i>n</i> -бутильных групп	Ц	5,00	
28. Число метильных групп	Ц	2,00	
29. Число этильных групп	Ц	3,00	
30. Число <i>n</i> -пропильных групп	Ц	5,00	
31. Число атомов углерода, не связанных с водородом	Ц	3,00	
32. $C:H = 2n + 2$	Б	5,00	
33. $C:H = 2n$	Б	5,00	
34. $C:H = 2n - 2$	Б	5,00	
35. $C:H = 2n - 6$	Б	5,00	
36. $C:H = 2n - 4$	Б	5,00	
37. Число групп $-CH_2-$ в цепи	Ц	1,00	
38. $C=C-C-CH_3$; метил в β -положении по отношению к $C=C$	Б	5,00	
39. Группа $-C\equiv N$	Б	5,00	

Продолжение табл. 7.1

Дескриптор	Тип	Нормировочный множитель	Ограничения
40. Группа — NO ₂	Б	5,00	
41. Группа — NH ₂	Б	5,00	
42. Атом азота, связанный не менее чем с двумя атомами углерода	Б	5,00	
43. Изопропильная группа	Б	5,00	
44. Число колец	Ц	4,00	
45. Размер цикла в моноциклическом соединении	Ц	1,00	
46. Наименьшее кольцо	Ц	1,00	Немоноциклическое соединение
47. Конденсированные кольца	Б	5,00	
48. α-Замещение	Б	5,00	Атом азота в кольце
49. γ-Водород	Б	5,00	Только ациклические соединения
50. Карбоксильная группа	Б	5,00	
51. Альдегид	Б	5,00	
52. Две электронодонорные группы (<i>орто</i>)	Б	5,00	Только 6-членные ароматические кольца
53. Две электронодонорные группы (<i>мета</i>)	Б	5,00	То же
54. Две электронодонорные группы (<i>пара</i>)	Б	5,00	» »
55. Неконденсированные кольца	Б	5,00	
56. Две электроноакцепторные группы (<i>орто</i>)	Б	5,00	Только 6-членные ароматические кольца
57. Две электроноакцепторные группы (<i>мета</i>)	Б	5,00	То же
58. Две электроноакцепторные группы (<i>пара</i>)	Б	5,00	» »
59. Донорноакцепторные группы (<i>орто</i>)	Б	5,00	» »
60. Донорноакцепторные группы (<i>мета</i>)	Б	5,00	» »
61. Донорноакцепторные группы (<i>пара</i>)	Б	5,00	» »

признаков сокращают число дескрипторов, необходимых для линейной разделмости.

Второй и третий этапы в принципе можно поменять местами. Возможны и иные способы описания молекулярных структур, например линейное (с помощью линии Висвессера) или графическое изображения. При кодировании фрагментов не обязательно исполь-

зывать единый для всех молекул перечень дескрипторов — его можно систематически видоизменять с учетом особенностей исходного массива данных.

В табл. 7.1 перечислен 61 дескриптор, которые использовали авторы рассматриваемого исследования. В первой колонке таблицы указаны категории дескрипторов, которые в большинстве случаев не нуждаются в дополнительных пояснениях. Тем не менее на отдельных определениях следует остановиться. «Наибольшая длина цепи» (*largest clump*) означает наибольшее число связанных друг с другом атомов углерода в цепи. «Наибольшее кольцо» — максимальное число атомов углерода, кислорода или азота при однократном обходе всего кольца (так, у нафталина наибольшее кольцо включает 10 атомов). «Наименьшее кольцо» — наименьшее число атомов при однократном обходе всего кольца (у нафталина, например, наименьшее кольцо включает 6 атомов). «Число колец и двойных связей» определяют по следующей формуле 9]:

$$\text{Число атомов C} + \frac{\text{Число атомов N}}{2} - \frac{\text{Число атомов N}}{2} + 1.$$

«Простой эфир», «кетон» и «спирт» имеют обычный смысл, но сочетание соответствующих функциональных групп в одной молекуле не разрешено (молекула спирта, например, может содержать только гидроксильные группы). «Сложный эфир», «карбоксильная группа» и «альдегид» означают вхождение соответствующих групп в состав соединения как по отдельности, так и в том или ином сочетании (метилтерефталат, скажем, содержит карбоксильную и сложную эфирную группы). «Карбонильная группа» — наличие в молекуле двойной связи углерод — кислород. «Кислородная связь» означает, что два атома углерода соединены между собой кислородным мостиком. «Одно бензольное кольцо» характеризует принадлежность к моноциклическим соединениям. В классификационных целях считается, что бензольное кольцо имеет три двойных связи. «Число атомов углерода в точках разветвления цепи» означает число атомов углерода, непосредственно связанных не менее чем с тремя другими атомами углерода. Под числами метильных, этильных, *n*-пропильных и *n*-бутильных групп понимаются числа этих функциональных групп, образующихся при разрыве одинарной связи. Категория «число атомов углерода, не связанных с водородом» охватывает четвертичные атомы углерода, не связанные с атомами водорода. «Две электронодонорные группы (*орто*)» — категория, относящаяся к 6-членным ароматическим кольцам, содержащим не менее двух заместителей, из которых два находятся в

орто-положении друг к другу. Deskрипторы других заместителей определяются аналогично. Категория « α -замещение» означает метильную группу в *орто*-положении по отношению к атому азота в кольце. « γ -Водород» — атом водорода в γ -положении по отношению к карбонильной группе в ациклическом соединении. Определения остальных дескрипторов даны согласно классическим представлениям.

Дескрипторы подразделяются на две категории: бинарные (Б) и цифровые (Ц). Бинарные дескрипторы могут принимать только два значения соответственно утвердительному и отрицательному ответам. Цифровые дескрипторы могут иметь значения до 202 (молекулярный вес $C_{10}H_{18}O_4$). Поэтому значения дескрипторов необходимо нормировать. Нормировочный множитель, на который умножали дескриптор, указан в третьей колонке табл. 7.1. Нормировка переводила значения дескрипторов в более удобный диапазон. Все бинарные дескрипторы умножали на 5.

В четвертой колонке указаны ограничения в отношении дескрипторов или оговорены особые случаи, например, категория « γ -водород» относится только к ациклическим соединениям.

В табл. 7.2 приведены дескрипторы нескольких соединений из массива данных. Под каждым соединением указано численное значение, приписанное еще ненормированному дескриптору. Следует отметить, что эти дескрипторы не отражают ни *цис-транс*-изомерии, ни положения функциональных групп в неароматических циклических системах. Выбор дескрипторов производился с таким расчетом, чтобы подразделить структуры на как можно большее число классов. Как будет показано, во многих случаях в перечне дескрипторов содержится достаточный объем информации для линейной разделимости.

Бинарные классификаторы образов обучали на масс-спектрах с 60 положениями *m/e*, перечисленными в первой колонке табл. 7.3. Каждый такой классификатор обучали предсказывать наличие или отсутствие пика, соответствующего данному значению *m/e* в спектрах соединений обучающей выборки. Считалось, что такой пик имеется в том или ином положении, если его интенсивность превосходила определенное критическое значение — пороговую интенсивность (порог).

Для всех 60 положений *m/e* строили весовые векторы относительно пороговой интенсивности, составляющей 0,5% полного ионного тока. Обученный таким образом весовой вектор способен ответить на следующий вопрос: существует ли в данном положении *m/e* масс-спектра интересующего нас соединения пик с такой интен-

Таблица 7.2

Перечни дескрипторов для пяти выбранных соединений

Номер дескриптора	Номер соединения				
	1	2	3	4	5
1	112	88	136	180	118
2	8	2	10	8	9
3	0	0	6	0	6
4	8	4	10	9	9
5	16	8	16	8	10
6	0	2	0	4	0
7	0	0	0	0	0
8	1	1	3	6	5
9	0	0	0	0	0
10	0	1	0	1	0
11	0	0	0	0	0
12	0	0	0	0	0
13	0	1	0	1	0
14	0	1	0	1	0
15	0	0	0	1	0
16	0	0	0	0	0
17	0	0	0	1	1
18	0	0	0	1	1
19	0	0	0	1	0
20	0	0	0	0	0
21	0	0	0	3	3
22	0	0	0	0	0
23	0	1	0	0	0
24	2	0	5	2	2
25	1	2	1	2	1
26	0	0	0	0	0
27	0	0	0	0	0
28	2	2	3	1	0
29	0	1	0	0	0
30	0	0	0	0	0

Продолжение табл. 7.2

Номер дескриптора	Номер соединения				
	1	2	3	4	5
31	0	1	2	4	1
32	0	0	0	0	0
33	1	1	0	0	0
34	0	0	0	0	0
35	0	0	0	0	0
36	0	0	1	0	0
37	4	1	1	0	2
38	0	0	0	0	0
39	0	0	0	0	0
40	0	0	0	0	0
41	0	0	0	0	0
42	0	0	0	0	0
43	0	0	0	0	0
44	1	0	3	1	2
45	6	0	0	6	0
46	0	0	3	0	3
47	0	0	1	0	0
48	0	0	0	0	0
49	0	1	0	0	0
50	0	0	0	1	0
51	0	0	0	0	0
52	0	0	0	0	0
53	0	0	0	0	0
54	0	0	0	0	0
55	0	0	0	0	1
56	0	0	0	0	0
57	0	0	0	0	0
58	0	0	0	1	0
59	0	0	0	0	0
60	0	0	0	0	0
61	0	0	0	0	0

Продолжение табл. 7.2

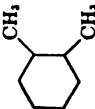

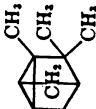
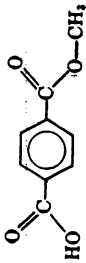
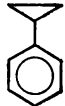
Номер соединения	Номер спектра по каталогу АНИ	Структура	Химическое название
1	220		1,2-Диметилциклогексан
2	326		Этилацетат
3	466		1,7,7-Триметилтрицикло-[2,2,1,0 (2,6)]-гептан (трициклен)
4	1755		Метилтерфталат
5	1963		Циклопропилбензол

Таблица 7.3

Отбор признаков и прогнозирующая способность

m/e	Пороговая интенсив- ность, %	Число оставшихся дескрипторов	Число коррекции через обратную связь	δa	Δb	$\Delta - \delta^B$
29	0,1	15	270	89,7	91,3	1,6
	0,5	18	1246	80,7	92,0	11,3
	1,0	19	275	75,7	89,6	13,9
30	0,5	19	47	88,3	93,3	5,0
31	0,5	13	52	79,5	89,6	10,1
37	0,5	17	27	84,5	91,8	7,3
38	0,5	39	741	63,2	81,1	17,9
39	0,1	14	42	96,3	97,3	1,0
	0,5	14	20	93,7	96,2	2,5
	1,0	13	36	91,2	93,3	2,1
40	0,5	61	>2500	53,7	79,6	25,9
41	0,1	11	7	96,8	96,4	-0,4
	0,5	15	93	88,5	92,7	4,2
	1,0	25	83	82,7	93,8	11,1
42	0,1	18	>2500	87,5	94,0	6,5
	0,5	17	47	79,9	91,8	11,9
	1,0	61	>2500	56,2	72,1	15,9
43	0,1	19	294	85,1	89,8	4,7
	0,5	61	>2500	71,9	90,0	18,1
	1,0	61	>2500	64,7	84,0	19,3
44	0,5	61	>2500	70,5	76,5	6,0
45	0,5	15	42	81,5	90,8	9,3
46	0,5	13	33	96,1	95,3	-0,8
50	0,5	17	139	72,1	90,6	18,5
51	0,1	61	>2500	80,7	85,9	5,2
	0,5	61	>2500	56,1	93,3	37,2
	1,0	14	166	74,1	89,9	15,8
52	0,5	20	149	74,9	90,1	15,2
53	0,1	18	103	87,0	91,7	4,7
	0,5	26	173	56,4	86,3	29,9

Продолжение табл. 7.3

<i>m/e</i>	Пороговая интенсив- ность, %	Число оставшихся дескрипторов	Число коррекций через обратную связь	δ_a	Δ^b	$\Delta - \delta^B$
	1,0	61	2349	63,6	86,7	23,1
54	0,5	61	>2500	70,8	82,2	11,4
55	0,1	35	574	79,2	81,7	2,5
	0,5	61	>2500	66,7	82,7	16,0
	1,0	30	1084	56,6	85,5	28,9
56	0,5	61	>2500	55,6	79,3	23,7
57	0,5	61	>2500	50,9	78,1	27,2
58	0,5	61	>2500	75,7	77,7	2,0
59	0,5	28	1058	87,2	83,3	-3,9
63	0,5	21	70	80,8	92,5	11,7
65	0,5	26	>2500	74,4	92,2	17,8
66	0,5	22	336	83,1	90,6	7,5
67	0,1	29	629	59,1	88,4	29,3
	0,5	29	1328	64,2	86,1	21,9
	1,0	34	>2500	74,8	85,4	10,6
68	0,5	61	>2500	76,1	86,3	10,2
69	0,5	61	>2500	56,1	87,0	30,9
70	0,5	61	>2500	61,2	81,4	20,2
71	0,5	61	>2500	76,0	81,3	5,3
73	0,5	18	251	90,1	89,0	-1,1
74	0,5	34	2265	90,1	86,6	-3,5
75	0,5	24	496	92,3	91,3	-1,0
76	0,5	12	12	94,1	96,3	2,2
77	0,1	15	56	52,0	89,4	37,4
	0,5	21	42	76,7	93,4	16,7
	1,0	18	31	82,5	90,1	7,6
78	0,5	10	161	84,2	94,0	9,8
79	0,1	28	873	54,1	86,7	32,6
	0,5	20	131	79,1	91,3	12,2
	1,0	16	104	87,6	89,9	2,3
80	0,5	18	250	92,9	93,9	1,0

Продолжение табл. 7.3

<i>m/e</i>	Пороговая интенсив- ность, %	Число оставшихся дескрипторов	Число коррекции через обрат- ную связь	δ_a	Δ_b	$\Delta - \delta^B$
81	0,5	61	>2500	80,6	91,3	10,7
82	0,5	25	>2500	84,1	87,8	3,7
83	0,5	61	>2500	74,8	88,8	14,0
84	0,5	61	>2500	72,8	75,8	3,0
85	0,5	61	>2500	86,4	88,5	2,1
91	0,5	11	491	83,2	92,0	8,8
95	0,5	13	141	88,2	93,2	5,0
97	0,5	61	>2500	84,2	87,3	3,1
98	0,5	61	>2500	78,0	76,2	-1,8
100	0,5	18	230	95,8	96,6	0,8
103	0,5	11	69	82,5	89,5	7,0
104	0,5	27	1365	85,1	90,1	5,0
105	0,5	8	58	85,0	92,1	7,1
106	0,5	15	39	91,1	95,5	4,4
115	0,5	18	68	81,8	93,5	11,7
119	0,5	14	26	83,5	95,4	11,9
120	0,5	18	31	83,9	94,7	10,8
121	0,5	13	27	89,6	92,1	2,5
127	0,5	6	5	93,5	93,2	-0,3
128	0,5	13	8	87,9	92,3	4,4
136	0,5	3	5	83,8	82,1	-1,7

^a δ — доля (%) объектов наиболее обширной категории от полного числа объектов выборки.

^b Δ — прогнозирующая способность, %.

^B Разность $\Delta - \delta$.

сивностью, которая превосходит 0,5% полного ионного тока, или же она имеет допороговое значение?

Кроме того, для каждого из 11 специально выбранных положений m/e строили по два весовых вектора относительно пороговых интенсивностей, равных 0,1 и 1,0% полного ионного тока. Первый вектор предсказывал, превосходит ли интенсивность пика 0,1%-ный пороговый уровень или же не превосходит (равна или меньше). Второй вектор давал аналогичные ответы в отношении порога, составившего 1,0% полного ионного тока. Выбор этих 11 положений был сделан с таким расчетом, чтобы в обучающей и контрольной выборках имелось достаточное число соединений с интенсивностью пиков выше 1,0% полного ионного тока. Таким способом было обучено 82 бинарных классификатора образов.

Во второй колонке табл. 7.3 указаны пороговые интенсивности, использованные при обучении каждого из этих 82 пороговых логических элементов. Наличие примесей в образцах, помехи масс-спектрометра, а также присутствие в природных образцах изотопов ^{13}C и ^{15}N обуславливают слабые пики в масс-спектрах. Пороговая интенсивность предотвращает возможность попадания подобного шума в обучающуюся машину, которая весьма чувствительна к помехам при принятии решения. За нижний предел интенсивности, которую может иметь пик, был выбран 0,1%-ный уровень полного ионного тока для соответствующего спектра. Все пики с допороговыми интенсивностями исключались из спектра.

Как уже упоминалось, при помощи подпрограммы для генерации случайных чисел была составлена обучающая выборка из спектров 150 соединений, тогда как остальные 450 вошли в состав контрольной выборки неизвестных соединений. Прежде чем началось обучение, из исходного массива в 600 спектров исключали соединения, молекулярный вес которых, увеличенный на единицу, был меньше позиционной координаты (значения m/e). Было бы нелогично выявлять корреляцию между структурой того или иного соединения и каким-то положением m/e , если молекулярный вес такого соединения был меньше соответствующего значения m/e . Поскольку самый небольшой молекулярный вес среди соединений исходного массива составлял 40, все значения m/e меньше 42 (меньше 41 включительно) этой процедурой не затрагивались. Объем обучающей выборки для таких положений остается равным 150 спектрам, а объем контрольной — 450 спектрам. Для m/e 83 исходный массив сокращается с 600 до 500 соединений; для m/e 98 остается 400 соединений; для m/e 106 остается 302 соединения и для m/e 119 — всего 200 соединений. Для m/e 128 массив данных состоит

из 149 соединений, из которых для m/e 136 остается всего 68 соединений. При сокращении исходного массива уменьшаются и объемы обучающей и контрольной выборок. Так, объем обучающей выборки для m/e 128 уменьшается от 150 до 32 соединений, из которых 4 характеризуются пиками с интенсивностью выше пороговой — 0,5% полного ионного тока, а 28 — ниже пороговой. Контрольная выборка распадается на 14 соединений с «запороговыми» пиками и 103 соединения без пиков, интенсивность которых превосходила бы 0,5% полного ионного тока.

Важно отметить, что от уровня пороговой интенсивности зависит разбиение на категории соединений обучающей и контрольной выборок. В случае m/e 29 число соединений, характеризующихся пиками с интенсивностью выше 0,1% полного ионного тока, составляет 137, а 13 соединений имеют пики с более низкой интенсивностью. При 0,5%-ном пороге в положительную категорию попадает 121 соединение, а в отрицательную — 29. Если же порог выбрать равным 1,0% полного ионного тока, то в положительной категории будет 112, а в отрицательной 38 соединений. Эта же тенденция наблюдалась для обеих выборок и в случае других положений при переходе от одного из трех порогов к другому.

В третьей колонке табл. 7.3 приведено число дескрипторов, оставшихся после отбора признаков по знаку весового вектора, который проводился обучающейся машиной. Такие дескрипторы считались важными свидетельствами наличия того или иного пика в конкретном положении для данного уровня пороговой интенсивности, установленного для исследуемого массива данных. Несмотря на то обстоятельство, что число оставшихся дескрипторов во многих случаях было весьма незначительным и составляло лишь небольшую долю от 61 исходного дескриптора, весовые векторы сохраняли ту же прогнозирующую способность, что и для всего числа исходных дескрипторов.

В четвертой колонке табл. 7.3 дано число коррекций через обратную связь, необходимых для обеспечения сходимости. Если, например, число коррекций >2500 , то это означает, что 2500 итераций не обеспечивали полной обученности. Нельзя считать, что подобные случаи указывают на линейную неразделимость. В отдельных случаях необученные полностью весовые векторы показывали довольно высокую прогнозирующую способность, как, например, для порогового логического элемента, обучающегося на пиках с m/e 65. Однако в нескольких случаях обучения все-таки имела место неразделимость, которая, по-видимому, была следствием недостаточности информации для описания молекул. Чтобы

преодолеть эту трудность, в ходе исследования были выработаны дополнительные дескрипторы. Их включение в табл. 7.1 обусловлено неопределенностью в описании молекул, замеченной обучающейся машиной. К сожалению, не всегда удастся обнаружить все такие неопределенности.

Существует простой способ проверки способности обучающейся машины классифицировать неизвестные соединения путем сравнения ее прогнозирующей способности с долей соединений наиболее обширной категории от числа всех соединений для того или иного положения m/e . Если прогнозирующая способность машины превосходит долю угадываний в том случае, когда соединение всегда относят к наиболее обширной категории, то можно сказать, что машина чему-то научилась в отношении установления связи между образами и категорией, к которой они принадлежат. Так, для m/e 29 доля соединений с пиками, интенсивность которых выше 1%-ной пороговой интенсивности, составляет 75,5%. Если утверждать, что соединение имеет пик в данном положении, то доля разовых верных классификаций составит 75,5%. Аналогично для m/e 128 доля соединений с пиками, интенсивность которых выше 0,5% полного ионного тока, равна 12,1%. Предположив, что ни одно из соединений всей совокупности не имеет пика в данном положении, вы сделаете 87,9% правильных разовых классификаций. В пятой колонке табл. 7.3 приведена доля (%) соединений наиболее обширной категории при условии, что пороговая интенсивность для конкретного положения m/e имеет значение, указанное во второй колонке.

В шестой колонке табл. 7.3 указана прогнозирующая способность весовых векторов для числа дескрипторов, приведенного в третьей колонке. Для каждого положения обучали три пороговых логических элемента на трех разных обучающих выборках. Здесь представлены данные для весовых векторов, показавших максимальную прогнозирующую способность по каждому положению и для каждой пороговой интенсивности. Средняя прогнозирующая способность для всех 82 пороговых логических элементов составила 88,8%.

В седьмой колонке табл. 7.3 указана разность между прогнозирующей способностью обучающейся машины, приведенной в шестой колонке, и фигурирующей в пятой колонке долей соединений наиболее обширной категории от всех соединений данной совокупности. В 73 случаях прогнозирующая способность обучающейся машины превосходит долю соединений наиболее обширной категории; средняя разность оказалась равной 10,4%.

Обнаруженные корреляции между дескрипторами и положениями m/e можно разделить на два типа: истинные корреляции между фрагментами структуры и положениями m/e и случайные корреляции, или артефакты, обучающих выборок. Чтобы свести к минимуму число подобных артефактов, корреляции для трех обучающих выборок рассматривали как единое целое. После завершения обучения и отбора признаков по каждому положению m/e оставалось по три пары весовых векторов и три набора дескрипторов для данного положения. Поскольку имелись три обучающие выборки, не все дескрипторы в трех перечнях были одинаковыми. Если какой-то дескриптор фигурирует во всех трех перечнях, то это должно указывать на сильную корреляцию. И, что еще важнее, когда компоненты каждой пары весовых векторов для таких дескрипторов имеют одинаковые знаки, корреляцию можно считать реальной. (Пара весовых векторов в случае одной и той же обучающей выборки имеет для каждого дескриптора одинаковые знаки, но не обязательно одинаковую величину; это есть следствие метода отбора признаков для сокращения числа дескрипторов.)

В качестве примера в табл. 7.4 приведены результаты обучения

Таблица 7.4

Результаты обучения и отбора признаков для m/e 45^a

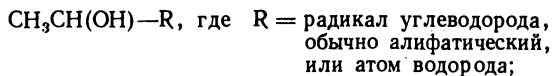
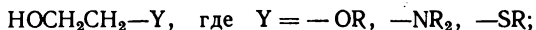
Дескриптор	Корреляция
2. Наибольшая длина цепи	—
6. Число атомов кислорода	+
9. Простой эфир	+
14. Кислородная связь	+
15. Гидроксильная группа (—ОН)	+

^a Дескрипторы, сохранившиеся во всех трех обучающих выборках при использовании 0,5%-ной пороговой интенсивности.

и отбора признаков для m/e 45. Хотя корреляции зависят от наличия примесей и изотопов (так, благодаря естественному содержанию изотопов ^{13}C и ^{15}N фрагменты с m/e 43 и 44 обуславливают пики, свойственные положению m/e 45), эти эффекты в известной мере ослабляются введением пороговой интенсивности, составлявшей в данном случае 0,5% полного ионного тока. Обучение и отбор

признаков на трех обучающих выборках дали три набора из 15, 26 и 15 дескрипторов, среди которых насчитывалось 35 разных дескрипторов. Общими для всех трех выборок были только 5 дескрипторов, которые перечислены в табл. 7.4 с указанием характера соответствующих корреляций. Дескриптор наибольшей длины цепи показал отсутствие корреляции. Дескрипторы простого эфира, кислородной связи и наличия гидроксильной группы свидетельствовали о положительных корреляциях. Подобные корреляции отражают тот факт, что компоненты весовых векторов для этих дескрипторов имеют одинаковый знак по всем трем обучающим выборкам; по дескриптору числа атомов кислорода две обучающие выборки дали положительные результаты и одна — отрицательный. Последнее обстоятельство объясняется, по-видимому, артефактом третьей выборки.

Положению *m/e* 45 соответствуют следующие структурные фрагменты: CNO_2 , $\text{C}_2\text{H}_5\text{O}$ и $\text{C}_2\text{H}_7\text{N}$. Первый фрагмент представляет карбоксильную группу. Поскольку в массиве данных было мало карбоновых кислот, корреляция с этим фрагментом проявлялась слабо. Дескриптор карбоксильной группы при отборе признаков был исключен и не попал ни в одну из трех обучающих выборок. Аналогично фрагмент $\text{C}_2\text{H}_7\text{N}$ никак не связан с дескрипторами, перечисленными в табл. 7.4; он образуется в процессе перегруппировки и едва ли может обуславливать появление многих пиков в данном положении. Поэтому обучающаяся машина исключила все «азотные» дескрипторы, считая их, подобно дескриптору карбоксильной группы, малозначимыми при решении задачи. Но дескрипторы простого эфира, кислородной связи и наличия гидроксильной группы хорошо согласуются со вторым фрагментом ($\text{C}_2\text{H}_5\text{O}$). Согласно Мак-Лафферти [10], этот фрагмент может входить в состав первичных и вторичных спиртов, а также простых эфиров:



Отрицательную корреляцию дескриптора наибольшей длины цепи объяснить не так просто. Можно было бы рассуждать следующим образом. Если в молекуле много атомов углерода связано друг с другом, то вероятность наличия атома кислорода или кислородсодержащего фрагмента мала. То обстоятельство, что дескриптор числа атомов кислорода обнаруживает как положительную, так

и отрицательную корреляции с обучающими выборками, означает, что процесс принятия решения обучающейся машиной носит более сложный характер, нежели простое признание присутствия атомов кислорода в соединении. Пожалуй, более важно учитывать расположение атомов кислорода в соединении, чем их число.

Для общей проверки пригодности данного метода прогнозирования из 600 соединений массива данных случайно выбрали 30 соединений. Для каждого соединения были составлены прогнозы по всем 60 положениям m/e с использованием 82 построенных весовых векторов. Для 49 положений можно рассчитать «бинарный» масс-спектр наличия или отсутствия пиков в этих положениях в том случае, когда в качестве пороговой берется интенсивность, соответствующая 0,5% полного ионного тока. Использование трех значений пороговой интенсивности помогает количественно определить интенсивность пиков в 11 остальных положениях. Были составлены прогнозы для каждого из 30 таких соединений по всем положениям до значения $(m/e)=M+1$, где M — молекулярный вес соединения. Данные табл. 7.5 характеризуют применимость при-

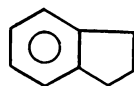
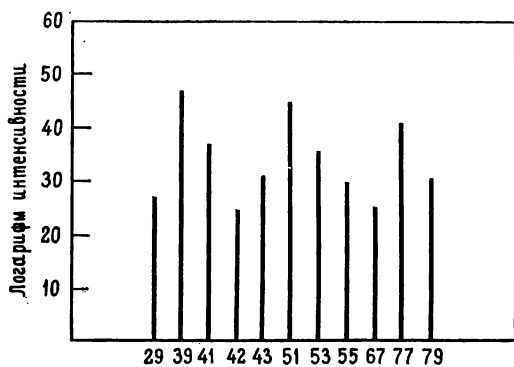
Таблица 7.5

**Прогнозирование полного спектра
(для 30 случайно выбранных соединений)**

Всего ошибок	Число ошибок с «пиковой» стороны	Число ошибок с «беспиковой» стороны	Прогнозирующая способность, %
154	85	69	92,9

ма, описанного в рассматриваемом исследовании. Средняя прогнозирующая способность оказалась равной 93%. Эта цифра несколько выше средней прогнозирующей способности для всех 82 пороговых логических элементов, потому что небольшая часть из 30 случайно выбранных соединений входила в обучающие выборки некоторых из этих элементов. Достигнутая прогнозирующая способность (93%) показывает, что использованные методы могут обеспечить высокую прогнозирующую способность даже в тех случаях, когда сохраняется только небольшая часть дескрипторов.

На рис. 7.1—7.6 показаны предсказанный и реальный масс-спектры с 11 пиками для 6 очень разных соединений из 30 исходных. По горизонтальной оси указаны 11 положений m/e , для реальных



индан C_9H_{10}

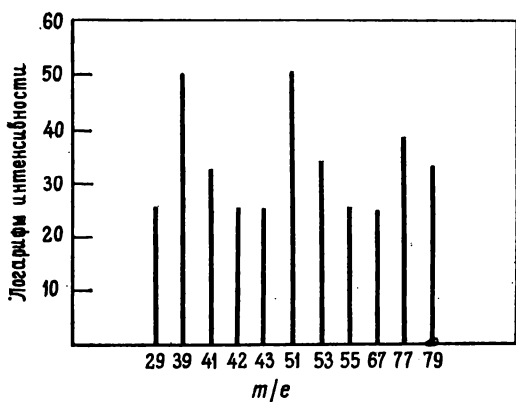


Рис. 7.1. Фактический (вверху) и предсказанный (внизу) масс-спектры с 11 пиками для индана.

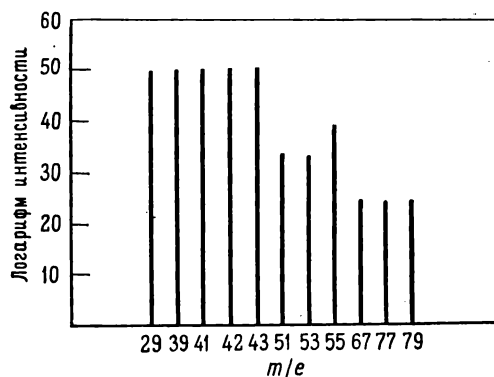
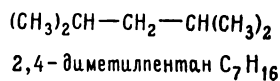
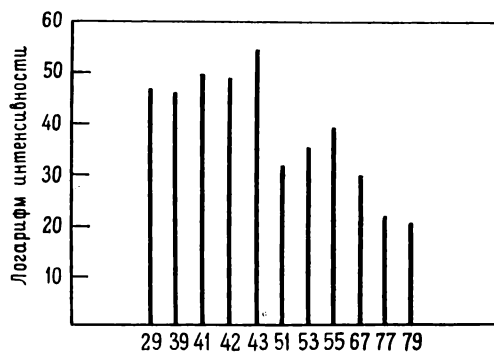


Рис. 7.2. Фактический (вверху) и предсказанный (внизу) масс-спектры с 11 пиками для 2,4-диметилпентана.

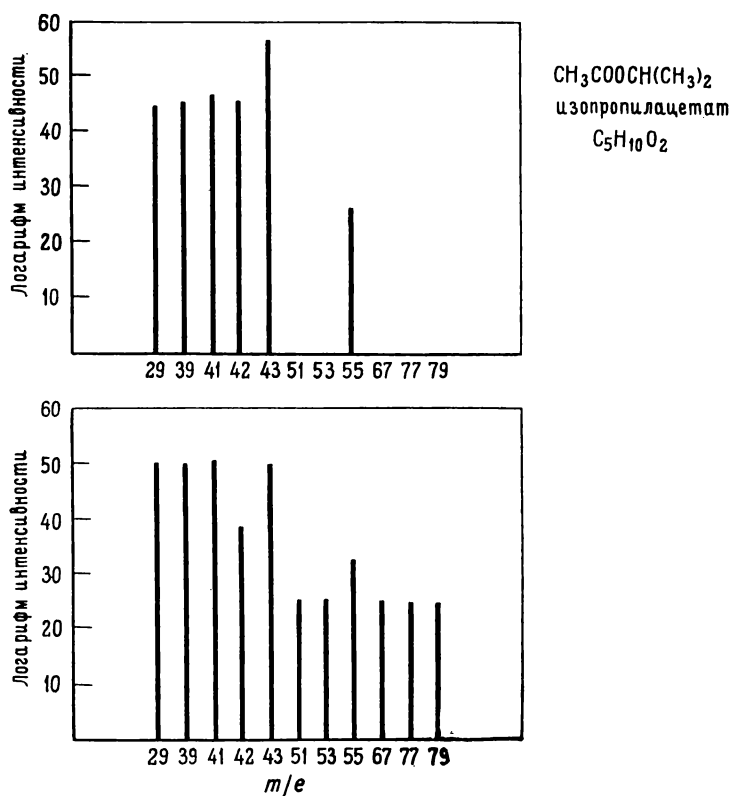


Рис. 7.3. Фактический (вверху) и предсказанный (внизу) масс-спектры с 11 пиками для изопропилацетата.

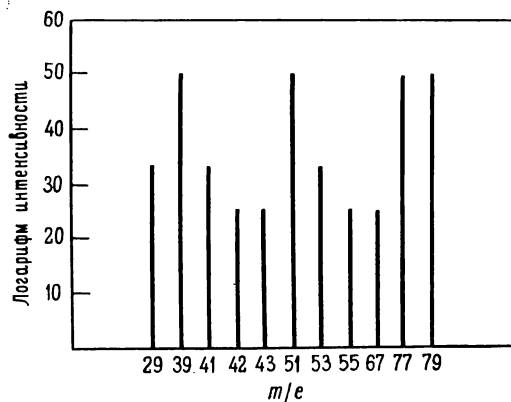
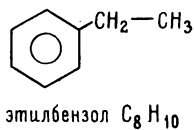
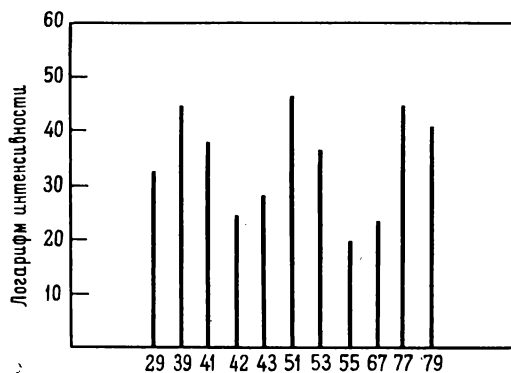


Рис. 7.4. Фактический (вверху) и предсказанный (внизу) масс-спектры с 11 пиками для этилбензола.

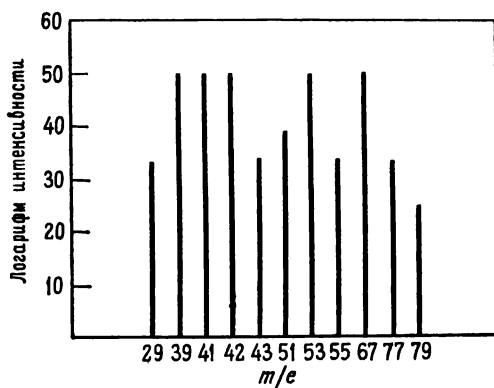
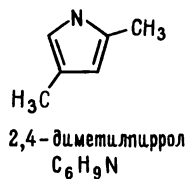
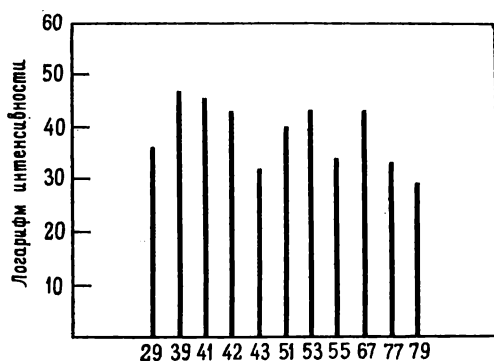
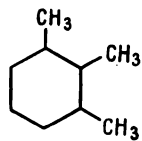
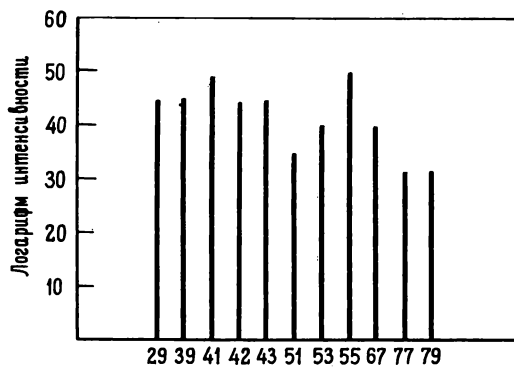


Рис. 7.5. Фактический (вверху) и предсказанный (внизу) масс-спектры с 11 пиками для 2,4-диметилпиррола.



1,2,3-триметилциклогексан C_9H_{18}

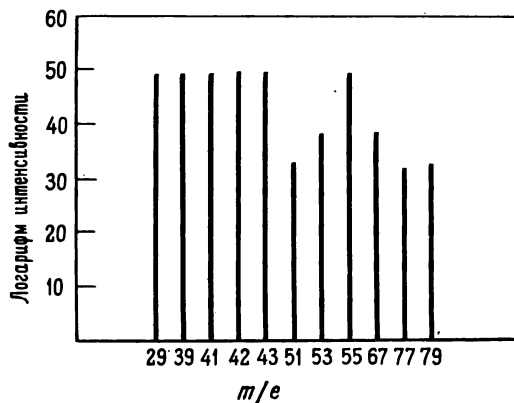


Рис. 7.6. Фактический (вверху) и предсказанный (внизу) масс-спектры с 11 пиками для 1,2,3-триметилциклогексана.

масс-спектров; по вертикальной оси даны значения, вычисленные по формуле

$$I' = 10 \lg(10000 \cdot I),$$

где I — доля (%) полного ионного тока для спектра, приходящаяся на рассматриваемый пик, и I' — интенсивность для реальных спектров. Значение 30 по вертикальной оси соответствует интенсивности, эквивалентной 0,1% полного ионного тока; значения 40 и 50 — интенсивностям, эквивалентным 1,0 и 10,0% полного ионного тока соответственно.

Значения интенсивностей пиков при разных положениях для предсказанных масс-спектров устанавливали следующим образом. (Обучали по три пороговых логических элемента с порогами 0,1, 0,5 и 1,0% полного ионного тока, что соответствует 30, 37 и 40 единицам по логарифмической шкале интенсивности на графиках.) Пик, для которого все три весовых вектора давали положительные скалярные произведения, произвольно приписывали значение 50 (10% полного ионного тока). Если все три весовых вектора давали отрицательные скалярные произведения, то интенсивность считалась равной 25 (0,03%; эта величина была, вероятно, обусловлена шумом). Если же три весовых вектора давали скалярные произведения с разными знаками, то интенсивность полагали равной 34 (в случае расхождения между векторами с порогами 30 и 37) или 39 (в случае расхождения между векторами с порогами 37 и 40).

33 пороговых логических элемента, производившие отраженные на графиках классификации, допустили следующие ошибки. В прогнозах для индана есть две ошибки — положения m/e 77 и 79. В случае 2,4-диметилпиррола все прогнозы подтвердились. В предсказаниях для изопропилацетата были допущены две ошибки — положения m/e 42 и 55. (Пики при m/e 51, 53, 67, 77 и 79 были правильно отнесены к шуму.) В предсказаниях для этилбензола сделана одна ошибка при m/e 41. Все предсказания для 2,4-диметилпентана и 1,2,3-триметилциклогексана были правильными. Все 6 предсказанных масс-спектров оказались удивительно похожими на реальные масс-спектры. Надо отметить, что для наглядности было показано только 11 пиков, классифицировать которые были обучены три пороговых логических элемента. При общей проверке были составлены прогнозы и для 49 остальных положений.

Исследование, результаты которого обобщены в табл. 7.3—7.5 и которое основывалось на кодировании фрагментов молекулярных структур, доказало возможность прогнозирования масс-спектров непосредственно на основе молекулярной структуры. Однако

чтобы добиться лучших результатов, нужно выявить и устранить недостатки, присущие подобному подходу. Одна из трудностей связана с несовершенством кодирования фрагментов, обусловленным недостаточной полнотой описания молекулярной структуры. Другая — с балансированием (уравновешенным распределением) объектов положительной и отрицательной категорий в обучающих выборках. Эти, а также другие трудности находились в центре внимания авторов исследований, которые рассматриваются ниже.

Множественные признаки

Одну из важных проблем представляет составление дескрипторов, которые учитывали бы одновременное вхождение в состав молекулы нескольких фрагментов; такие дескрипторы получили название множественных признаков.

Выше уже обсуждался вопрос о выделении признаков, поэтому снова рассматривать его в общем виде мы не будем.

Основное внимание уделялось совершенствованию описания молекулярных структур. Используемые нами ранее структурные дескрипторы указывали на наличие отдельных структурных фрагментов в молекуле и не выявляли комбинаций структурных фрагментов, за исключением нескольких случаев. Так, 48-й дескриптор, а именно α -замещение, свидетельствует о появлении в молекуле метильной концевой группы и кольца, содержащего атом азота. Данный дескриптор характеризует вместе с тем взаимное расположение этих двух фрагментов. Метильная группа соединена с атомом углерода в кольце, который непосредственно связан с атомом азота. Использование приемов выделения признаков позволяет автоматически составлять дескрипторы, включающие несколько структурных фрагментов как единое целое. Дескриптор подобного рода называют множественным признаком. Множественные признаки, составляемые из структурных фрагментов перечня дескрипторов, не отражают позиционных соотношений. Не исключено, что вывести признаки, описывающие взаимное расположение фрагментов, удастся при ином подходе к изображению топологии молекулы вместо составления перечня дескрипторов. Рассматриваемые ниже множественные признаки указывают на комбинации фрагментов в молекулах, а не на позиционные соотношения между ними.

Используемый нами метод выделения признаков известен под названием алгоритма включения атрибутов. Включение атрибутов характеризует взаимосвязь между ними в той или иной выборке образов. В этом случае атрибут является синонимом дескриптора.

Молекулярные структуры изображаются векторами образов, компонентами которых служат значения атрибутов (дескрипторов). Используемый здесь алгоритм ограничен бинарными атрибутами (ноль и единица). Поэтому предъявляемые алгоритму векторы образов состоят из 40 бинарных дескрипторов, отобранных из перечня дескрипторов, который охватывает 61 структурный фрагмент.

Один атрибут включают в другой во всех тех случаях, когда присутствие первого в каком-либо образе означает и присутствие второго. Любые два атрибута, удовлетворяющих отношению включения, принадлежат одному и тому же признаку. Следовательно, все атрибуты, связанные последовательными операциями включения, можно объединить в единственный признак и рассматривать его как множественный признак, или множественный дескриптор. Таким образом, нужно располагать набором признаков, который группировал бы вместе атрибуты, связанные друг с другом корреляцией взаимного включения. Математически включение атрибутов отображает векторы образов из пространства атрибутов в пространство признаков меньшей размерности.

Вопрос о том, каким путем формируют признаки по отношению включения, систематически изложен в статье Абдали [11], в которой приведено также несколько примеров восстановления символов. Как выяснилось, этот метод позволяет восстанавливать образы по признакам, которые им выводятся.

В рассматриваемых Абдали примерах признаки выделяются из всей выборки образов, представляющих символы. Следовательно, признаки выражают ту общую часть всех образов, которая содержится в выборке данных. Затем комбинации выделенных признаков используются для восстановления любого образа. Тогда образ описывается не его атрибутами, а составленным из признаков представлением меньшей размерности.

Алгоритм включения атрибутов применяют к данным о молекулярной структуре иначе, чем это делал Абдали в его работе по восстановлению символов. Нас интересует прежде всего классификация образов, а не их восстановление, поэтому формирование множественных признаков мы должны рассматривать скорее как средство разбиения образов на два класса. При этом алгоритм включения атрибутов применяют отдельно к каждому такому классу, а не ко всем вместе как к единой выборке данных. К тому же множественные признаки используют совсем не как альтернативное представление исходных образов меньшей размерности. Напротив, ими пополняют исходные образы, увеличивая размерность пространства образов. По-видимому, «расширенные» образы легче под-

разделить на соответствующие категории, чем исходные образы.

Алгоритм выделения признаков применяют следующим образом. Выбрав то или иное положение m/e и порог интенсивности, берут обучающую выборку молекулярных структур и подразделяют ее на две категории. Соединения, характеризующиеся пиком достаточной интенсивности при этом положении m/e , хранят как один класс образов, называемый первой категорией. Соединения же, для которых в данном положении пик достаточной высоты отсутствует, хранят как второй класс образов, называемый второй категорией. Затем каждый класс образов отдельно предъявляют подпрограмме выделения признаков. Таким путем составляют набор признаков, общий для «пиковых» соединений, и другой набор, общий для «беспиковых» соединений.

Подпрограмма выделения признаков ограничивается 40 бинарными дескрипторами из их перечня. Образы как первой, так и второй категорий, к которым применяют алгоритм включения атрибутов, первоначально состоят из 40 атрибутов, или дескрипторов. Алгоритм начинается с исключения всех дескрипторов, не фигурирующих ни в одном из образов рассматриваемой категории. В процессе формирования признаков любой дескриптор, появляющийся точно в тех же образах, что и другой дескриптор, систематически исключают. Данное состояние известно как состояние равенства. Признаки, сформированные алгоритмом, могут состоять и из одного дескриптора, и из ряда дескрипторов. Признаки, состоящие только из одного дескриптора, обычно бывают следствием появления данной категории во многих образах. Множественные признаки, состоящие более чем из одного атрибута (дескриптора), записывают в конце перечня дескрипторов. Те же признаки, которые состоят только из одного атрибута, в конце перечня не указываются, поскольку они в нем уже фигурируют. Алгоритм используют дважды — по одному разу для каждой категории. Все множественные признаки из обеих категорий дополняют перечень дескрипторов и заносятся в него под порядковыми номерами, начиная с номера 62.

Пример применения подпрограммы выделения признаков к положению m/e 45 приведен в табл. 7.6. Обучающая выборка состоит из 25 соединений с пиками, интенсивность которых превосходит 0,5% полного ионного тока, и из 124 соединений с пиками, интенсивность которых равна или меньше 0,5% полного ионного тока. Обе категории представлены векторами, имеющими по 40 компонент. Значения этих компонент равны либо единице, либо нулю (до нормировки). После ввода 25 «пиковых» соединений в

Таблица 7.6

**Множественные признаки, выделенные для положения *m/e* 45,
при использовании 0,5%-ной пороговой интенсивности**

«Пиковые» объекты обучающей выборки	25
Исходное число атрибутов	40
Число не обнаруженных в объектах атрибутов	24
Окончательное число атрибутов	16
Оставшиеся атрибуты	

26 42 51 10 16 50 49 12 13 43 15 33 9 14 32 23
42-й атрибут исключен по состоянию равенства

<u>Дескриптор</u>	<u>Множественные признаки</u>
62	26 42 43 23
63	51 10 49 13 33 14 23
64	10 49 13 33 14 23
65	16 33 9 14 23
66	50 13 15 33 23
67	49 13 33 23
68	12 15 32 23
69	43 23
70	9 14 23
71	32 23

Число «пиковых» объектов обучающей выборки	124
Исходное число атрибутов	40
Число не обнаруженных в объектах атрибутов	7
Окончательное число атрибутов	33
Оставшиеся атрибуты	

12 40 59 15 39 41 51 55 48 10 11 14 54 36 53 52 20 49
35 42 19 13 26 47 38 17 18 34 16 32 43 33 23

14-й атрибут исключен по состоянию равенства

<u>Дескриптор</u>	<u>Множественные признаки</u>
72	12 15 55 34
73	40 26 23
74	59 39 19 26 17 18
	15
75	39 26
76	41 26

Продолжение табл. 7.6

Дескриптор	Множественные признаки					
77	51	13	23			
	55					
78	48	20	42	26	17	
79	10	14	49	13	23	
80	11	13	23			
81	54	35	19	17	18	
	36					
82	53	52	35	19	17	18
83	52	35	19	17	18	
84	20	42	26			
85	49	13	23			
86	36	19	17	18		
	42					
87	19	17	18			
	26					
	47					
	38					
	17					
	18					
	34					
	16					
88	32	23				
	43					
	33					
	23					

подпрограмму выделения признаков 24 бинарных дескриптора пришлось исключить из рассмотрения по отношению включения, поскольку ни в одном из 25 образов они не появлялись. Один атрибут, 42-й дескриптор, исключили по состоянию равенства. Затем алгоритм вычислил, как показано в таблице, 10 множественных признаков. В этом случае не оказалось ни одного признака, который состоял бы только из одного дескриптора. Когда алгоритм обследовал «беспиковые» объекты обучающей выборки, пришлось исклю-

чить 7 атрибутов, потому что они не появлялись ни в одном из 124 соединений этой выборки. Из оставшихся 33 атрибутов 14-й атрибут пришлось изъять из-за равенства с другим атрибутом. Затем были сформированы 32 признака; из них 16 одиночных признаков пришлось исключить, поскольку они уже фигурировали в перечне из 61 дескриптора. Таким образом, осталось 17 множественных признаков.

10 признаков от «пиковых» соединений обучающей выборки и 16 признаков от «беспиковых» соединений записывают в конце перечня дескрипторов. Тем самым они дополняют векторы образов еще на 26 компонент, доводя число дескрипторов на каждую молекулярную структуру до 88.

Всю совокупность образов — как обучающую, так и контрольную выборки — обследуют на появление множественных признаков. Множественный признак сохраняется в том или ином соединении, если каждый дескриптор признака появляется в молекуле. Например, каждое соединение, будь то «пиковое» или «беспиковое», обследуют на наличие дескриптора 79. Если какое-то соединение содержит кислородную связь, карбонильную группу, эфирную группу, относится к ациклическим соединениям и имеет атом водорода в γ -положении по отношению к кислороду карбонильной группы в соответствии с дескрипторами 10, 14, 49, 13 и 23 множественного признака 79, то значение 79-й компоненты вектора образа считается равным единице. Если же у этого соединения нет какого-то из этих дескрипторов, то 79-я компонента принимает нулевое значение. Нормировочный множитель для множественных признаков полагают равным 5,0, как и для бинарных дескрипторов.

Была предпринята проверка ряда положений m/e по рассмотренному выше алгоритму включения атрибутов. На исследуемых положениях m/e осуществляли простое обучение бинарных классификаторов образов с коррекцией через обратную связь. Предполагалось, что подобный способ выделения признаков повысит прогнозирующую способность бинарных классификаторов образов в результате более совершенного описания молекулярных структур. Неожиданно оказалось, что данный способ позволяет экономить время обучения бинарных классификаторов образов.

В табл. 7.7 сопоставляются данные о скорости сходимости для 8 положений m/e с использованием в одном случае всех 61 дескриптора и во втором — еще и всех множественных дескрипторов, сформированных алгоритмом включения атрибутов. Как обычно, была составлена обучающая выборка из 150 случайно выбранных соединений, а остальные из 600 соединений исходного массива

Таблица 7.7

**Влияние множественных признаков на скорость сходимости
бинарных классификаторов образов**

m/e	Без множественных признаков			С множественными признаками			N_1/N_2
	число дескрипторов	прогнозирующая способность, %	число коррекций N_1	число дескрипторов и множественных признаков	прогнозирующая способность, %	число коррекций N_2	
29	61	88,7	63	87	88,0	57	1,1
	61	88,4	77	87	89,1	51	1,5
31	61	88,9	113	90	90,2	81	1,4
	61	87,1	91	90	88,7	81	1,1
45	61	89,9	130	88	91,7	34	3,8
	61	92,0	74	88	93,3	44	1,7
59	61	82,4	2056	90	83,1	391	5,3
	61	82,8	535	90	82,1	277	1,9
67	61	85,4	1244	88	88,0	844	1,5
	61	84,9	1442	88	81,4	735	2,0
73	61	90,0	664	88	90,5	492	1,4
	61	86,3	379	88	84,5	383	1,0
77	61	91,6	137	89	90,8	119	1,2
	61	92,4	43	89	90,5	33	1,3
104	61	89,2	915	81	88,4	655	1,4
	61	87,1	464	81	87,1	331	1,4

служили контрольной выборкой. До обращения к подпрограмме выделения признаков из обучающей и контрольной выборок исключали все соединения, не совместимые с исследуемым положением m/e . Выше уже объяснялось, почему это приходится делать. Единственная разница между данной и предшествующей работами заключается в том, что подпрограмма выделения признаков увеличи-

вает число дескрипторов для бинарного классификатора образов. В обоих случаях — при использовании множественных признаков и без них — обучающую и контрольную выборки, равно как и пороговую интенсивность для конкретного положения m/e брали одинаковыми.

В первой колонке табл. 7.7 перечислены 8 исследованных положений m/e ; в каждом случае пороговая интенсивность составляла 0,5% полного ионного тока. В третьей колонке указана прогнозирующая способность бинарного классификатора образов при использовании 61 дескриптора (см. вторую колонку табл. 7.7). Верхнее число из пары значений, указанной в третьей колонке, относится к исходному весовому вектору со значением $+1$, а нижнее — к весовому вектору, которому в начальном состоянии было приписано значение -1 . В четвертой колонке приведено число коррекций через обратную связь, необходимых для полного распознавания обучающей выборки. В пятой — полное число дескрипторов, использованных бинарным классификатором образов, включая множественные дескрипторы. Оно равно сумме 61 дескриптора с числом множественных дескрипторов, построенных подпрограммой выделения признаков для конкретного положения m/e . В шестой и седьмой колонках приведены прогнозирующая способность и число коррекций до сходимости при использовании множественных дескрипторов. В восьмой — отношения чисел коррекций, приведенных в четвертой и седьмой колонках. Это отношение характеризует увеличение скорости обучения, обусловленное добавлением к векторам образов множественных дескрипторов.

Для m/e 67 бинарные классификаторы образов способны при использовании 61 дескриптора предсказать появление пика с интенсивностью, превосходящей 0,5% полного ионного тока, с точностью 85,4 и 84,9%. Для обучения весовых векторов с исходными значениями $+1$ и -1 требуется 1244 и 1442 коррекции соответственно. При добавлении 27 множественных признаков, означающем почти 50%-ное увеличение размерности векторов образов, бинарный классификатор образов предсказывает появление пика с точностью 88,0 и 81,4%, для чего ему требуются 844 и 735 коррекций соответственно. Скорость сходимости возрастает в 1,5 и 2 раза по сравнению со случаем, когда множественные дескрипторы не используются.

В среднем скорость сходимости для 8 положений m/e возросла в 1,8 раза, если исходные векторы образов дополняли множественными дескрипторами. Связанное с этим приращение размерности обычно составляет 50%. Поскольку подпрограмма выделения при-

наков работает быстро, время ее выполнения с избытком окупается экономией, даваемой ускорением процесса обучения, а обучение, как известно, — весьма медленный процесс в классификационной методике.

Чтобы проверить, в какой степени использование множественных дескрипторов повышает прогнозирующую способность, бинарные классификаторы образов (табл. 7.7) заставили работать в режиме отбора признаков.

В табл. 7.8 приведены результаты отбора признаков для 8 уже исследованных положений *m/e*. В данном случае брали те же вы-

Таблица 7.8

Влияние множественных признаков на прогнозирующую способность после отбора признаков

<i>m/e</i>	Без множественных признаков		При использовании множественных признаков	
	осталось дескрипторов	прогнозирующая способность, %	число дескрипторов и множественных признаков	прогнозирующая способность, %
29	18	92,0	24	92,9
31	13	89,6	19	90,4
45	15	90,8	17	93,3
59	28	83,3	30	84,7
67	29	86,1	42	85,2
73	18	89,0	21	89,3
77	21	93,4	23	92,6
104	27	90,1	31	91,0

борки — обучающую и контрольную — как и при отборе признаков, так и без него. Для каждого из перечисленных в первой колонке положений *m/e* пороговая интенсивность составляла 0,5% полного ионного тока. Во второй и третьей колонках приведены число дескрипторов и прогнозирующая способность без использования множественных признаков. В четвертой и пятой — число дескрипторов и прогнозирующая способность бинарных классификаторов образов, осуществлявших отбор признаков. Во всех этих случаях приводится прогнозирующая способность, показанная после того, как в процессе отбора признаков были исключены все ненужные дескрипторы.

Без использования множественных дескрипторов и с отбором признаков число дескрипторов для положения *m/e* 59 было доведено до 28 с прогнозирующей способностью бинарного классификатора образов 83,3%; бинарный классификатор образов на 30 дескрипторах и с использованием множественных признаков имел прогнозирующую способность 84,7%.

Отбор признаков повышал прогнозирующую способность в среднем на 0,6%. Среднее число множественных дескрипторов, сохраненных 8 бинарными классификаторами образов после отбора признаков, было равно 5, хотя это обстоятельство не нашло отражения в табл. 7.8. Нет в ней и данных о наблюдавшемся ускорении процесса обучения после последовательных операций по отбору признаков в случаях использования множественных дескрипторов.

Субструктурные дескрипторы

Чтобы усовершенствовать совокупность дескрипторов, к векторам образов добавляют дескрипторы другого типа. Это — субструктурные дескрипторы, характеризующие наряду с самим фактом наличия структурных особенностей и позиционные соотношения.

Была поставлена задача: предсказать масс-спектры молекул углеводов непосредственно по описанию их молекулярных структур. Использовалась выборка из 377 углеводов с молекулярной формулой $C_{3-10}H_{2-22}$. Чтобы приступить к разработке программ, пришлось сформировать две совокупности данных: 1) описания молекулярных структур в необходимом векторном формате; 2) совокупность ответов, т. е. масс-спектров соответствующих структур.

Чтобы вводить описания молекулярных структур в программы обучающихся машин, им пришлось придавать уже упоминавшийся векторный формат. Из множества путей осуществления поставленной задачи остановились на сочетании двух распространенных подходов: кодов фрагментов и субструктурных кодов. Методика кодирования фрагментов состоит из описания соединения в виде составного целого, подразделяющегося на главные структурные фрагменты и связи между ними. Затем полученным признакам были приписаны порядковые номера. Фрагменты, использовавшиеся при описании рассматривавшихся молекулярных структур, перечислены в табл. 7.9. В большинстве случаев они не нуждаются в пояснениях, однако отдельные замечания следует сделать. «Наибольшее кольцо» относится к максимальному числу атомов углерода

Таблица 7.9

Дескрипторы фрагментов^a

1. Молекулярный вес	Ц
2. Наибольшее кольцо	Ц ^б
3. Число атомов углерода	Ц
4. Число атомов водорода	Ц
5. Число колец и двойных связей	Ц
6. Винильная концевая группа	Б
7. Ароматическое соединение	Б
8. Наличие бензольного кольца	Б
9. Одно бензольное кольцо	Б
10. Число связей C=C	Ц
11. Число связей C≡C	Ц
12. Ациклическое соединение	Б
13. Число атомов углерода в точках разветвления цепи	Ц
14. Число <i>n</i> -бутильных групп	Ц
15. Число метильных групп	Ц
16. Число этильных групп	Ц
17. Число <i>n</i> -пропильных групп	Ц
18. $H=2C+2$	Б
19. $H=2C$	Б
20. $H=2C-2$	Б
21. $H=2C-6$	Б
22. $H=2C-4$	Б
23. Число метиленовых групп в цепи	Ц
24. Метил в β-положении по отношению к C=C; C=C—C—CH ₃	Б
25. Изопропильная группа	Б
26. Число колец	Ц
27. Размер цикла в моноциклическом соединении	Ц
28. Наименьшее кольцо	Ц
29. Конденсированные кольца	Б

^a Ц — цифровой, Б — бинарный.^б Немоноциклическое соединение.

при замкнутом однократном обходе кольца (так, у нафталина наибольшее кольцо включает 10 атомов). «Наименьшее кольцо» означает минимальное число атомов при замкнутом однократном обходе кольца (у нафталина, например, наименьшее кольцо включает 6 атомов). Дескрипторы с ненулевыми значениями наибольшего и наименьшего колец могут принадлежать только соединениям, имеющим не менее двух колец. «Одно бензольное кольцо» характеризует принадлежность к моноциклическим молекулам. Для удобства классификации считалось, что бензольное кольцо имеет три двойные связи. «Число атомов углерода в точках разветвления цепи» означает число атомов углерода в соединении, которые связаны непосредственно не менее чем с тремя другими атомами углерода.

Числа метильных, этильных, *n*-пропильных и *n*-бутильных групп указывают число этих функциональных групп, которые образуются при разрыве одинарной связи. Существуют 29 дескрипторов фрагментов.

Дескрипторы принадлежат двум категориям: бинарной и цифровой (в табл. 7.9 они отмечены). Бинарные дескрипторы могут принимать в зависимости от ответов «да» и «нет» только два значения. В векторе образа 1 соответствует наличию фрагмента в кодируемой структуре, тогда как 0 означает его отсутствие. Цифровые дескрипторы могут принимать значения до 142 (молекулярный вес $C_{10}H_{22}$). Из-за изменения значений дескрипторов в очень широком диапазоне их приходится нормировать. Нормировочные множители выбирают с таким расчетом, чтобы сократить разброс дескрипторов в совокупности данных и сузить диапазон их изменения. Так, для числа атомов водорода нормировочный множитель равен 0,5; таким образом, диапазон изменения для этого дескриптора лежит в пределах 0—11. Все бинарные дескрипторы нормировали умножением на 5.

Остальные 26 дескрипторов принадлежат к субструктурным и являются бинарными (табл. 7.10). Массив данных состоит, следовательно, из 377 углеводородов, структуры которых закодированы в виде векторов образов с 55 координатами.

Массив исходных данных был заимствован из сводного каталога масс-спектров в записи на магнитной ленте, имеющегося в распоряжении Центра масс-спектрометрических данных при Управлении атомной энергии Англии. На этой же ленте был записан 2261 спектр из таблиц Американского нефтяного института. 377 масс-спектров углеводородов были переписаны с этой части магнитной ленты. После цифрового преобразования интенсивности в каждом спектре лежали в диапазоне 0,01—100,00%. Интенсивность каждого пика пересчитывали в логарифмический масштаб по формуле

$$I' = 10 \lg (10000 \cdot I),$$

где I — процентная доля полного ионного тока, а I' — пересчитанная интенсивность. Преобразованные интенсивности имеют значения либо нуль, либо находятся в диапазоне 10—60. В логарифмическом масштабе 30 единиц соответствуют 0,1% полного ионного тока, 37—0,5% и 40—1,0%.

Один цикл обучения отдельного порогового элемента проходит следующим образом. Полный массив данных из 377 структур кодируют в 55-мерные векторы и случайным образом подразделяют на две выборки: обучающую — из 200 соединений и контроль-

Субструктурные дескрипторы

30. $\text{CH}_2=\text{CH}-\text{CH}_2-$	41. $\text{CH}_3-\text{CH}-\text{CH}-$ $\begin{array}{c} \\ \text{CH}_3 \end{array}$ $\begin{array}{c} \\ \text{CH}_3 \end{array}$	49. $\begin{array}{c} \text{CH}_3 \\ \\ \text{CH}_2 \\ \\ \text{CH}-\text{CH}- \\ \quad \\ \text{CH}_3 \quad \text{CH}_3 \end{array}$
31. $\text{CH}_2=\text{CH}-\text{CH}_2-\text{CH}_2-$	42. $\begin{array}{c} \text{CH}_3 \\ \\ \text{CH}_3-\text{C}-\text{CH}_2- \\ \\ \text{CH}_3 \end{array}$	50. $\begin{array}{c} \text{CH}_3 \\ \\ -\text{C}- \\ \quad \\ \text{CH}_2 \quad \text{CH}_3 \end{array}$
32. $\text{CH}_3-\text{C}=\text{CH}-$ $\begin{array}{c} \\ \text{CH}_3 \end{array}$	43. $\begin{array}{c} \text{CH}_3-\text{CH}_2 \\ \quad \\ \text{CH}_3-\text{CH}_2-\text{CH}- \\ \quad \\ \text{CH}_3 \quad \text{CH}_3 \end{array}$	51. $\begin{array}{c} \text{CH}_3 \\ \\ -\text{CH}_2-\text{C}-\text{CH}_2- \\ \quad \\ \text{CH}_3 \quad \text{CH}_3 \end{array}$
33. $\text{CH}_2=\text{C}-\text{CH}_2-$ $\begin{array}{c} \\ \text{CH}_3 \end{array}$	44. $\begin{array}{c} \text{CH}_3-\text{CH}_2-\text{C}- \\ \quad \\ \text{CH}_3 \quad \text{CH}_3 \end{array}$	52. $\begin{array}{c} \text{CH}_3 \quad \text{CH}_3 \\ \quad \\ -\text{CH}_2-\text{CH}-\text{CH}_2- \\ \\ \text{CH}_3 \end{array}$
34. $\text{CH}_2=\text{CH}-\text{CH}-$ $\begin{array}{c} \\ \text{CH}_3 \end{array}$	45. $\begin{array}{c} \text{CH}_3 \\ \\ -\text{C}- \\ \\ \text{CH}_3 \end{array}$	53. $\begin{array}{c} \text{CH}_3 \\ \\ -\text{CH}-\text{CH}_2-\text{CH}_2-\text{CH}- \\ \quad \quad \\ \text{CH}_3 \quad \text{CH}_3 \end{array}$
35. $\text{CH}_3-\text{CH}=\text{C}-$ $\begin{array}{c} \\ \text{CH}_3 \end{array}$	46. $\begin{array}{c} \text{CH}_3 \quad \text{CH}_3 \\ \quad \\ -\text{CH}-\text{CH}- \\ \\ \text{CH}_3 \end{array}$	54. $\begin{array}{c} \text{CH}_3 \quad \text{CH}_3 \\ \quad \\ -\text{C}-\text{CH}_2-\text{CH}- \\ \\ \text{CH}_3 \end{array}$
36. $\text{CH}_3-\text{CH}=\text{CH}-\text{CH}_2-$	47. $\begin{array}{c} \text{CH}_3 \quad \text{CH}_3 \\ \quad \\ -\text{CH}-\text{C}- \\ \quad \\ \text{CH}_3 \quad \text{CH}_3 \end{array}$	55. $\begin{array}{c} \text{CH}_3 \\ \\ -\text{CH}-\text{CH}-\text{CH}- \\ \quad \quad \\ \text{CH}_3 \quad \text{CH}_3 \quad \text{CH}_3 \end{array}$
37. $\begin{array}{c} \text{CH}_3 \\ \\ \text{CH}_3-\text{C}- \\ \\ \text{CH}_3 \end{array}$	48. $\begin{array}{c} \text{CH}_3 \quad \text{CH}_3 \\ \quad \\ -\text{CH}-\text{CH}_2-\text{CH}- \\ \\ \text{CH}_3 \end{array}$	
38. $\begin{array}{c} \text{CH}_3 \\ \\ \text{CH}_3-\text{CH}_2-\text{CH}- \\ \\ \text{CH}_3 \end{array}$		
39. $\text{CH}_3-\text{CH}_2-\text{CH}_2-\text{CH}_2-$		
40. $\begin{array}{c} \text{CH}_3 \\ \\ \text{CH}_3-\text{CH}-\text{CH}_2- \\ \\ \text{CH}_3 \end{array}$		

ную — из 177. Затем из каждой выборки удаляют структуры с молекулярным весом меньше значения m/e для того положения, на котором ведут обучение. Составляют выборку правильных ответов. Например, пороговый логический элемент можно обучить выявлению соединений, в масс-спектрах которых при m/e 57 имеется пик с интенсивностью больше пороговой, скажем, 35. Тогда пороговый логический элемент обучают на 200 объектах обучающей выборки для того, чтобы он показывал положительный результат для молекул с таким пиком и отрицательный для молекул без этого пика. После завершения обучения пороговый логический элемент заставляют классифицировать (неизвестные) объекты контрольной выборки. Процентную долю верных классификаций записывают как прогнозирующую способность. Чтобы развить способность принимать решения о величине пиков, нужно обучить несколько пороговых логических элементов с разными значениями пороговой интенсивности.

В табл. 7.11 приведены результаты обучения пороговых логических элементов для разных значений пороговой интенсивности по каждому из 25 положений m/e . В первой колонке перечислены положения m/e , на которых проводилось их обучение. По каждому положению обучали три пороговых логических элемента с такими порогами интенсивности, которые делили обучающую выборку на подмножества со следующими отношениями числа входящих в них объектов: 1:3, 1:1 и 3:1. Для каждого порогового логического элемента регистрировали данные трех типов: 1) интенсивность, соответствующую выбранному порогу; 2) название обучающей выборки; 3) прогнозирующую способность. Исследование проводилось на трех произвольно составленных обучающих выборках, условно обозначенных *A*, *B* и *C*. Если обучение на каждой такой обучающей выборке не завершалось за 2500 коррекций (предел экономичности), то из обучающей выборки исключали пять объектов, которые чаще всего отвергались обучающейся машиной, и предпринималась попытка повторить курс обучения. В случае необходимости цикл обучения можно было повторять. Так, обозначение *B*-10 для выборки, на которой обучали первый пороговый элемент на положении m/e 29, соответствует тому, что в процессе исключения объектов из обучающей выборки было изъято 10 структур, после чего была достигнута сходимость.

Прогнозирующая способность пороговых логических элементов лежала в пределах 70—97% и составила в среднем 87,2%. Средняя прогнозирующая способность пороговых логических элементов с отношениями числа объектов в подмножествах 1:3, 1:1 и 3:1 указана

Таблица 7.11

Результаты обучения пороговых логических элементов

m/e	1:3			1:1			3:1		
	порог	обучающая выборка	прогнозирую- щая способ- ность, %	порог	обучающая выборка	прогнозирую- щая способ- ность, %	порог	обучающая выборка	прогнозирую- щая способ- ность, %
29	41	B-10	94,9	45	B-10	83,6	47	A-10	91,5
41	47	B-5	93,8	49	A-10	69,5	50	B-10	84,8
42	38	B-10	89,3	41	A-10	76,0	44	A-10	81,1
43	35	B-10	93,2	43	A-10	87,4	47	B	96,6
54	31	B	95,4	36	B-10	93,1	40	B-10	84,0
55	40	B-5	91,4	46	A-5	90,3	50	A-5	82,3
56	33	B-5	92,5	42	B-10	86,1	48	A-10	80,8
57	29	B-5	86,8	37	B-5	89,0	46	A	95,4
66	25	A	92,9	31	A-5	82,8	34	B-5	93,5
67	31	B-5	90,5	37	B-10	89,4	42	B	97,1
68	25	B-10	89,4	33	B-10	94,7	39	B-10	88,2
69	31	B-5	90,5	40	A-5	88,8	46	A	84,0
70	28	B-5	95,0	39	B-10	82,9	45	B-10	81,1
71	19	B-10	84,2	31	B-5	82,9	38	B-5	89,6
80	19	B-10	90,6	25	B-10	86,4	30	B	91,3

Продолжение табл. 7.11

m/e	1:3			1:1			3:1		
	порог	обучающая выборка	прогнозирую- щая способ- ность, %	порог	обучающая выборка	прогнозирую- щая способ- ность, %	порог	обучающая выборка	прогнозирую- щая способ- ность, %
81	22	A-5	90,1	29	A	92,1	38	B-5	96,3
82	18	B-10	85,0	27	B-5	91,3	36	B-10	83,7
83	24	B-5	85,6	32	A-10	77,3	41	B-10	87,5
84	21	B-10	86,4	30	B-10	81,6	41	B-10	78,9
85	21	B-5	80,4	27	B-10	72,8	35	A-10	83,9
95	1	B-5	85,1	19	A-5	93,0	31	B-5	87,2
97	19	B-10	79,4	26	B-10	75,9	34	A-10	85,9
98	21	B-5	82,8	28	A-5	78,9	41	A-10	83,7
105				1	B-5	90,8	31	B	92,7
106				1	B-5	88,9	24	B-5	93,2
Среднее			88,9			85,0			87,8

в табл. 7.11. Случайное угадывание соответствует 50%-ной прогнозирующей способности для всех пороговых логических элементов. Предсказание, основывающееся на известных вероятностях для классов 1:3 и 3:1, дает для класса 1:1 величину 50%. Фактически же прогнозирующая способность оказалась значительно выше. Очевидно, что пороговые логические элементы способны делать непосредственно из молекулярных структур некоторые выводы общего характера относительно того, какие механизмы обуславливают появление в масс-спектрах пиков определенной относительной интенсивности.

В табл. 7.12 указаны результаты применения программы отбора признаков к совокупностям данных, сведения о которых были приведены в табл. 7.11. Программа отбора признаков пытается определить, какие дескрипторы молекулярной структуры важны при поиске решения конкретной химической задачи. Делается это следующим образом. Для данного масс-спектрометрического пика и исследуемого порога интенсивности независимо обучают два весовых вектора с разными исходными значениями. Затем в процессе отбора признаков исключаются те дескрипторы, которые ничего или почти ничего не дают для решения задачи. Осуществляется это путем сравнения знаков компонент двух весовых векторов. При этом сохраняют только такие дескрипторы, для которых компоненты обоих весовых векторов имеют одинаковые знаки. Следовательно, данная процедура приводит к сокращению размерности двух сохраняемых весовых векторов. Отбор признаков продолжают до тех пор, пока не останется малозначащих дескрипторов.

В табл. 7.12 приведены результаты отбора признаков для четырех положений m/e . В среднем сохранялось 34 дескриптора. В шестой колонке указано, сколько дескрипторов оставалось для каждого порогового элемента, в седьмой — число оставшихся фрагментов и субструктурных дескрипторов. Сопоставление данных, приведенных в четвертой и восьмой колонках, показывает, что при сокращении числа дескрипторов на обучение потребовалось либо столько же, либо меньше времени, чем при сохранении всей совокупности из 55 дескрипторов. Прогнозирующая способность (последняя колонка таблицы) почти во всех случаях стала выше: средняя прогнозирующая способность для всех 12 классификаций повысилась на 1,1% при сохранении всего ~60% дескрипторов.

Таким образом, информация, относящаяся к задаваемым химическим вопросам, содержится в части компонент исходных векторов образов. Отбрасывая малозначащие дескрипторы, можно улучшить характеристики пороговых логических элементов.

Таблица 7.12

Результаты отбора признаков

m/e	Порог	Обучающая выборка	55 дескрипторов		Образы уменьшенной размерности			
			числа коррекций	средняя прогнозирую- щая способ- ность, %	остаток дескрипторов	число фрагментов/ число субструктур	числа коррекций	средняя прогнозирующая способность, %
29	41	B-10	193/268	94,4	26	17/9	131/285	94,4
	45	B-10	1	71,5	42	23/19	1	78,8
	47	A-10	1946/1981	90,4	38	10/18	1888/2285	91,0
	35	B-10	95/88	92,0	25	20/5	56/60	92,1
	43	A-10	1101/1048	87,1	39	21/18	1290/787	86,3
67	47	B	2297/2276	95,8	38	22/16	2492/2501	95,8
	31	B-5	2197/2504	91,1	42	23/19	2500/2505	91,6
	37	B-10	2005/1783	90,9	46	26/20	1857/1823	91,8
	42	B	2467/1464	92,0	27	22/5	867/384	93,2
	25	B-10	346/570	90,8	27	18/9	349/323	92,0
68	33	B-10	878/731	94,0	33	20/13	689/590	95,0
	39	B-10	309/229	90,5	23	18/5	161/169	91,5

СПИСОК ЛИТЕРАТУРЫ

1. *Schechter J., Jurs P. C.*, Appl. Spectrosc. 27, 30 (1973).
2. *Schechter J., Jurs P. C.*, Appl. Spectrosc., 27, 225 (1973).
3. *Jurs P. C.*, in "Computer Representation and Manipulation of Chemical Information" (W. T. Wipke et al., Eds.), Wiley-Interscience, New York, 1974, p. 265.
4. *Rosenstock H. M., Kraus M.*, in "Mass Spectrometry of Organic Ions" (F. W. McLafferty, Ed.), Academic, New York, 1963.
5. *Vestal M. L.*, in "Fundamental Processes in Radiation Chemistry" (P. Ausloos, Ed.), Interscience, New York, 1968.
6. *Rosenstock H. M.*, in "Advances in Mass Spectrometry" (E. Kendrick, Ed.), Institute of Petroleum, London, 1968, Vol. 4.
7. *Cuchanan B., Sutherland G., Feigenbaum E. A.*, in "Machine Intelligence 4" (B. Meltzer, D. Michie, Eds.), American Elsevier, New York, 1969.
8. *Liddell R. W., III, Jurs P. C.*, Appl. Spectrosc., 27, 371 (1973).
9. *McLafferty F. W.*, Interpretation of Mass Spectra: An Introduction, W. A. Benjamin, New York, 1966.
10. *McLafferty F. W.*, Mass Spectral Correlations, American Chemical Society, Washington, D. C., 1963.
11. *Abdali S. K.*, Pattern Recognition, 3, 3 (1971).

ОБРАЗЕЦ ПРОГРАММЫ ДЛЯ МОДЕЛИРОВАНИЯ ОБУЧАЮЩЕЙСЯ МАШИНЫ

Программа, приведенная ниже в качестве примера, состоит из короткой основной программы, которая вводит массив данных, присваивает начальные значения нескольким параметрам, вызывает обучающую подпрограмму и дважды обращается к прогнозирующей подпрограмме.

DATA — массив данных, содержащий исходные данные или образы. Размерность массива рассчитана на 100 пятикомпонентных образов. Данные, которые предполагаются уже нормированными числами, вводятся с перфокарт. Идентификатор W обозначает изменяемый весовой вектор. LIST содержит признак категории каждого образа, +1 или -1; эта информация также вводится с перфокарт. Идентификатор NUM использован для обозначения общего числа компонент каждого образа. NTRSET — число образов в обучающей выборке. IDTR содержит число образов в контрольной выборке, а IDPR — список таких элементов исходных данных, которые составляют контрольную выборку. Контрольная выборка в приведенной задаче содержит элементы с номерами NTRSET+1, NTRSET+2, ... , NTOT. NPASS содержит число коррекций через обратную связь, разрешенных до самопроизвольно заканчивающегося вследствие отсутствия сходимости процесса обучения. TSHD — размер свободной (мертвой) зоны или порога, с которым обучается бинарный классификатор образов. NCONV сообщает вызывающей программе, была достигнута сходимость (NCONV=0) или нет (NCONV=1). WINIT — значение, присваиваемое перед обучением каждой компоненте весового вектора.

Подпрограмма TRAIN использует процедуру разбиения на подмножества. При первом исполнении эта подпрограмма использует всю обучающую выборку, применяя при необходимости обратную связь. В то же время подпрограмма хранит в NSS список неверно классифицированных элементов, принадлежащих обучающей выборке. При втором исполнении классифицируются только образы, опущенные при первом исполнении, и строится третье подмножество.

Процесс повторяется до тех пор, пока все образы не будут классифицированы верно. Затем классифицируется обучающая выборка полностью, и весь процесс начинается сначала. Обучающая программа выводит на печать число неверно классифицированных образов при последовательном выполнении процедуры разбиения на подмножества.

Ниже приведен пример работы программы для произвольно образованного линейного сепарабельного набора данных.

```

C          BASIC LEARNING MACHINE PROGRAM
C
C          DIMENSION DATA (5,100),W(6),LIST(100),IDTR(100),IDPR(100)
C          NTRSET=80
C          NPRSET=20
C          WINIT=0.1
C          TSHD=0.75
C          NTOT=NTRSET+NPRSET
C          NPASS=1000
C          NUM=5
C          INPUT DATA SET
C          DO 10 I=1,NTOT
10  READ (5,9) LIST(I),(DATA(J,I),J=1,NUM)
C          SET UP TRAINING SET
C          DO 20 I=1,NTRSET
20  IDTR(I)=I
C          SET UP PREDICTION SET
C          DO 30 I=1,NPRSET
30  IDPR(I)=I
C          INITIALIZE WEIGHT VECTOR
C          DO 40 J=1,NUM
40  W(J)=WINIT
C          W(NUM+1)=WINIT
C          CALL TRAIN (DATA,W,LIST,NTRSET,NUM,NPASS,TSHD,NCONV, IDTR)
C          CALL PREDICTION ROUTINE WITH DEADZONE OF 0.75
C          CALL PRED (DATA,LIST,W,NUM,TSHD,NPRSET,IDPR)
C          TSHD=0.0
C          CALL PREDICTION ROUTINE WITH DEADZONE OF 0.0
C          CALL PRED (DATA,LIST,W,NUM,TSHD,NPRSET,IDPR)
9  FORMAT (15,5F10.3)
C          STOP
C          END

```

```

SUBROUTINE TRAIN (DATA,W,LIST,NTRSET,NUM,NPASS,TSHD,NCONV, IDTR)
DIMENSION DATA (5,100),W(6),NSS(100),KPNT(20),LIST(100),IDTR(100)
NCONV=0
WRITE (6,169) NTRSET,NUM,TSHD
NUMM=NUM+1
NF=0
KNK=0
KNV=0
C START OF MAIN LOOP OF TRAINING PROGRAM (RETURN FROM ST 206)
51 KKK=0
IF (KNV) 54,54,53
53 NDSS=KNV
GO TO 65
54 NDSS=NTRSET
DO 60 I=1,NTRSET
60 NSS(I)=IDTR(I)
C THE 200 LOOP CLASSIFIES THE NDSS MEMBERS OF THE CURRENT SUBSET
65 DO 200 IR=1,NDSS
I=NSS(IR)
C THE 70 LOOP CALCULATES THE DOT PRODUCT
S=W(NUMM)
DO 70 J=1,NUM
70 S=S+DATA(I,J)*W(J)
C THE NEXT THREE IF STATEMENTS TEST FOR THE CORRECT ANSWER
IF (LIST(I)) 95,95,96
95 IF (S-TSHD) 200,200,116
96 IF (S-TSHD) 115,115,200
C STATEMENT 115 OR 116 CALCULATES C, THE CORRECTION INCREMENT
115 C=2.0*(TSHD-S)
GO TO 117
116 C=2.0*(-TSHD-S)
117 XX=1.0
DO 120 J=1,NUM
120 XX=XX+DATA(J,I)**2
C=C/XX
C THE 130 LOOP PERFORMS THE FEEDBACK
DO 130 J=1,NUM
130 W(J)=W(J)+C*DATA(J,I)
W(NUMM)=W(NUMM)+C
KKK=KKK+1
NSS(KKK)=I
NF=NF+1
200 CONTINUE
KNV=KKK
KNK=KNK+1
KPNT(KNK)=KNV
IF (KNK-20) 205,203,203
203 WRITE (6,159) KPNT
KNK=0
C STATEMENT 205 TESTS FOR EXCESS NUMBER OF FEEDBACKS
205 IF (NF-NPASS) 206,211,211
C ST 206 TESTS FOR WHETHER CURRENT SUBSET IS ENTIRE TRAINING SET
206 IF (NDSS-NTRSET) 51,207,51
C ST 207 TESTS FOR WHETHER ZERO ERRORS WERE COMMITTED
207 IF (KNV) 51,212,51
211 NCONV=1
C SUMMARY OUTPUT OF TRAINING ROUTINE
212 IF (KNK.GT.0) WRITE (6,159) (KPNT(K),K=1,KNK)

```

```

      WRITE (6,179) (W(J),J=1,NUMM)
      WRITE (6,149) NF
149  FORMAT (1H0,10X,9HFEEDBACKS,I6)
159  FORMAT (1H ,20I4)
169  FORMAT (1H0,10X,8HTRAINING,2I10,F10.2,/)
179  FORMAT ('0',10X,'WEIGHT VECTOR',/,( ' ',F17.3))
      RETURN
      END

```

```

SUBROUTINE PRED (DATA,LIST,W,NUM,TSHD,NPRSET,IDPR)
DIMENSION DATA(5,100),W(6),LIST(100),IDPR(100)
LW1=0
LW2=0
KW=0
NPA=0
NNA=0
DO 120 II=1,NPRSET
  I=IDPR(II)
  S=W(NUM+1)
  DO 50 J=1,NUM
50  S=S+DATA(J,I)*W(J)
    IF (ABS(S)-TSHD) 101,102,102
101  KW=KW+1
    GO TO 120
102  IF (LIST(I)) 103,103,105
103  NNA=NNA+1
    IF (-S-TSHD) 104,104,120
104  LW1=LW1+1
    GO TO 120
105  NPA=NPA+1
    IF (S-TSHD) 106,106,120
106  LW2=LW2+1
120  CONTINUE
      WRITE (6,109) TSHD
      LWT=LW1+LW2
      JW=NPA+NNA
      PW=100.0-FLOAT(LWT)/FLOAT(JW)*100.0
      PW1=100.0-FLOAT(LW1)/FLOAT(NNA)*100.0
      PW2=100.0-FLOAT(LW2)/FLOAT(NPA)*100.0
      WRITE (6,9) JW,KW,LWT
      WRITE (6,119) LWT,JW,PW,LW1,NNA,PW1,LW2,NPA,PW2
9  FORMAT ('0',I10,' NUMBER PREDICTED',/,' ',I10,' NUMBER NOT PREDI
    ICTED',/,' ',I10,' NUMBER PREDICTED INCORRECTLY')
109  FORMAT (1H0,///,' PREDICTION WITH DEAD ZONE',F10.4)
119  FORMAT (1H0,3(I10,1H/,I3,1X ,F6.2,5X))
      RETURN
      END

```

1	9.446	4.652	5.089	6.715	7.659
-1	1.047	7.768	1.232	9.566	9.657
-1	1.008	5.689	2.309	9.151	5.504
1	2.275	7.488	5.846	3.161	6.999
1	8.429	8.758	6.035	4.489	3.155
1	4.399	6.073	7.395	1.627	2.080
1	6.291	8.199	1.430	1.651	5.536
1	6.674	1.946	5.121	6.372	1.194
1	1.749	4.944	8.628	2.311	5.513
-1	8.184	2.320	1.038	9.559	3.555
-1	5.382	1.254	7.743	9.787	3.955
1	3.048	6.231	5.220	9.610	3.046
-1	2.261	4.320	4.996	7.056	9.340
-1	3.579	2.974	1.185	7.190	6.958
-1	4.085	1.623	3.737	7.432	5.965
-1	2.146	4.247	1.190	4.719	9.541
1	2.908	4.805	3.870	5.941	5.363
1	6.922	4.930	1.826	2.480	9.829
1	7.476	3.589	9.231	2.259	5.733
-1	1.177	1.362	1.915	8.731	6.713
-1	1.157	4.540	4.673	7.268	6.750
-1	2.749	1.998	1.341	5.190	8.152
-1	8.958	2.086	2.452	9.245	5.935
-1	2.093	1.533	9.708	8.097	7.790
-1	3.330	3.770	1.455	9.212	8.542
1	5.756	9.958	7.919	7.227	1.535
1	4.175	2.696	3.936	3.557	5.674
1	2.871	9.425	4.766	7.312	9.847
1	8.577	5.058	1.047	7.018	3.870
-1	2.654	3.293	7.049	9.879	3.001
-1	3.726	2.921	4.516	9.968	4.082
-1	3.092	1.436	9.442	9.286	4.120
1	4.348	6.583	5.728	2.982	2.992
1	6.689	8.739	4.492	8.297	3.041
-1	4.378	1.066	2.362	7.397	5.232
-1	8.009	2.085	1.943	9.197	4.167
1	9.455	7.800	2.158	2.379	9.298
1	9.266	9.077	2.034	7.627	6.522
1	5.503	4.550	6.226	4.038	8.867
1	7.345	1.293	3.187	3.292	2.449
-1	1.101	3.529	1.657	5.419	4.061
1	2.352	8.532	1.901	8.235	8.630
1	3.109	7.563	3.125	2.462	8.009
1	4.075	9.100	4.393	7.727	5.511
1	2.555	4.735	5.885	5.698	7.299
1	6.464	3.300	6.113	6.721	3.521
1	1.567	9.808	5.330	9.532	3.487
-1	7.158	1.457	1.034	9.176	9.725
-1	1.317	1.700	2.500	9.372	5.807
-1	2.056	4.475	1.524	6.913	7.857
-1	3.025	1.310	6.264	6.598	5.429
-1	1.737	8.103	2.569	9.755	9.521
-1	3.501	2.109	3.246	9.058	3.312
1	8.783	5.136	6.777	1.782	9.780
1	6.609	4.111	4.837	9.836	3.968
-1	2.221	4.699	3.316	7.243	6.017
1	6.879	5.609	3.993	3.646	3.119
1	9.439	7.733	6.781	4.484	9.830
1	1.654	7.676	7.608	3.614	2.206
-1	2.279	1.586	9.509	9.885	2.946

-1	2.623	1.391	4.690	8.046	8.887
1	3.734	5.677	5.274	7.038	3.660
1	2.135	9.473	8.350	9.809	5.401
1	6.344	3.960	4.158	3.321	5.287
-1	5.128	2.969	1.016	9.746	8.041
-1	2.153	2.851	8.124	9.884	5.044
-1	4.951	3.286	4.621	9.574	3.964
-1	3.533	1.038	9.379	8.235	5.738
-1	1.274	2.449	3.598	7.628	2.649
-1	2.007	3.298	2.301	9.787	1.103
-1	5.083	1.162	3.390	8.351	9.660
1	2.006	8.591	9.868	7.118	9.357
1	1.592	9.357	9.688	3.748	8.289
1	3.594	9.431	9.278	8.999	1.018
-1	2.302	2.236	1.904	8.819	3.910
1	9.445	4.268	3.223	8.712	7.521
1	2.260	4.152	6.866	1.817	9.250
-1	2.200	1.134	2.263	3.299	7.722
-1	1.737	1.964	2.830	5.420	8.458
-1	1.202	4.833	2.748	9.669	6.601
-1	3.478	3.282	1.898	5.178	8.770
-1	6.294	1.508	4.888	9.868	1.885
1	7.725	6.020	4.718	8.456	1.520
-1	1.394	1.421	2.657	6.356	6.314
1	2.125	2.623	2.794	1.176	2.070
-1	2.297	1.351	3.578	8.646	4.493
-1	4.130	2.838	8.205	9.652	6.082
-1	5.227	3.246	1.928	7.250	4.966
1	7.270	8.248	1.466	5.155	6.818
1	2.828	1.552	7.244	5.567	4.739
-1	4.458	1.216	6.749	8.130	5.664
1	5.712	8.347	3.220	9.281	1.201
1	5.004	9.805	4.983	9.737	8.059
-1	4.399	2.167	2.067	5.227	9.735
-1	4.763	1.766	5.146	9.342	9.912
1	9.608	3.267	7.283	9.118	6.363
1	2.467	5.562	4.090	3.670	2.879
1	6.401	1.149	8.228	1.455	4.615
-1	1.562	3.538	3.987	6.758	4.492
1	4.134	2.049	8.256	2.448	7.241

TRAINING															80	5	0.75	
21	10	8	6	4	4	3	3	3	3	3	2	2	2	1	0	12	8	
7	6	4	4	3	3	3	3	3	2	0	6	5	4	4	4	3	3	
3	3	3	3	3	2	0	2	2	2	1	0	6	2	2	1	0	4	3
0	3	2	2	1	0	4	3	3	3	3	2	2	2	2	1	0	0	1

WEIGHT VECTOR

0.605

0.960

0.299

-0.714

-0.410

0.022

FEEDBACKS 254

PREDICTION WITH DEAD ZONE 0.7500

20 NUMBER PREDICTED

0 NUMBER NOT PREDICTED

0 NUMBER PREDICTED INCORRECTLY

0/ 20 100.00

0/ 9 100.00

0/ 11 100.00

PREDICTION WITH DEAD ZONE 0.0

20 NUMBER PREDICTED

0 NUMBER NOT PREDICTED

0 NUMBER PREDICTED INCORRECTLY

0/ 20 100.00

0/ 9 100.00

0/ 11 100.00

СПИСОК МОНОГРАФИЙ

1. *Andrews H. C.*, Computer Techniques in Image Processing, Academic, New York, 1970.
2. *Andrews H. C.*, Introduction to Mathematical Techniques in Pattern Recognition, Wiley-Interscience, New York, 1972.
3. *Bell D. A.*, Intelligent Machines. An Introduction to Cybernetics, Blaisdell, New York, 1962.
4. *Бонгард М.*, Проблема узнавания, «Наука», М., 1967.
5. *Cheng G. C. et al.*, Pictorial Pattern Recognition, Thompson, Washington, D. C., 1968.
6. *Collins N. L., Michie D.* (Eds.), Machine Intelligence 1, American Elsevier, 1967.
7. *Dale E., Michie D.* (Eds.), Machine Intelligence 2, American Elsevier, 1968.
8. *Дуда Р., Харт П.*, Распознавание образов и анализ сцен, «Мир», М., 1976.
9. Вычислительные машины и мышление, под ред. Фейгенбаума Э., Фельдмана Дж., «Мир», М., 1967.
10. *Fogel L. F., Owens A. J., Wals M. J.*, Artificial Intelligence through Simulated Evolution, Wiley, New York, 1966.
11. *Fukunaga K.*, Introduction to Statistical Pattern Recognition, Academic, New York, 1972.
12. *Fu K. S.*, Proceedings of the First International Joint Conference on Pattern Recognition, IEEE, New York, 1973.
13. *Kanal L. N.*, Pattern Recognition, Thompson, Washington, D. C., 1968.
14. Распознавание образов. Исследование живых и автоматических распознающих систем, под ред. Колерс П. А., Иден М., «Мир», М., 1970.
15. *Meisel W. S.*, Computer-Oriented Approaches to Pattern Recognition, Academic, New York, 1972.
16. *Meltzer B., Michie D.* (Eds.), Machine Intelligence 4, American Elsevier, New York, 1969.
17. *Meltzer B., Michie D.* (Eds.), Machine Intelligence 5, American Elsevier, New York, 1970.
18. *Mendel J. M., Fu K. S.* (Eds.), Adaptive Learning and Pattern Recognition Systems, Academic, New York, 1970.
19. *Michie D.* (Ed.), Machine Intelligence 3, American Elsevier, New York, 1968.
20. *Минский П., Пейперт С.*, Перцептроны, «Мир», М., 1971.
21. *Нильсон Н.*, Обучающиеся машины, «Мир», М., 1967.
22. *Нильсон Н.*, Искусственный интеллект. Методы поиска решений, «Мир», М., 1973.
23. *Patrick E. A.*, Fundamentals of Pattern Recognition, Prentice-Hall, Englewood Cliffs, N. J., 1972.
24. *Себестиен Г. С.*, Процессы принятия решений при распознавании образов, «Техника», Киев, 1965.
25. *Slagle J. R.*, Artificial Intelligence. The Heuristic Programming Approach, McGraw-Hill, New York, 1971.

26. *Tou J. T., Wilcox R. H.*, Computer and Information Sciences, Spartan, Washington, D. C., 1964.
27. *Tou J. T.* (Ed.), Computer and Information Sciences II, Academic, New York, 1967.
28. *Uhr L.*, Pattern Recognition, Wiley, New York, 1966.

СПИСОК ДОПОЛНИТЕЛЬНОЙ ЛИТЕРАТУРЫ

1. *Аркадьев А. Г., Браверман Э. М.*, Обучение машины классификации объектов, «Наука», М., 1971.
2. Сб. «Алгоритмы обучения распознаванию образов», под ред. Вапника В. Н., «Советское радио», М., 1973.
3. *Савицкий Е. М., Девингталь Ю. В., Грибуля В. Б.*, ДАН СССР, 178, 1, 79 (1968).
4. *Гладун В. П.*, Кибернетика, 5, 109 (1972).
5. *Айзерман М. А., Браверман Э. М., Розоноэр Л. И.*, Метод потенциальных функций в теории обучения машин, «Наука», М., 1970.
6. *Вапник В. Н.*, Задача обучения распознаванию образов, «Знание», М., 1971.
7. *Киселева Н. Н., Покровский Б. И., Комиссарова Л. Н., Ващенко Н. Д.*, ЖНХ, 22, 4, 883 (1977).

ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

- Автокорреляционная функция (auto-correlation function) 158
Адамара преобразование 159
Адаптация (adaptation) 15
Анализ главных компонент (principal components analysis) 14, 108
АНИ (Американского нефтяного института) каталог спектров 30, 36, 46, 48, 78, 84, 88, 104, 118, 148, 173, 209
Атрибутов включение (attribute inclusion), алгоритм 198
- Байеса* правило 15
Байесова теория решений 47
- Вектор**
нормали (normal vector) 22
образа (pattern vector) 12, 19, 22
Весовой вектор (weight vector) 24, 48, 63, 111
построение по методу наименьших квадратов 75
Висвессера линия 176
Восприятие (perception) 9
Выделение признаков (feature extraction) 29
- Гаусса—Жордана* метод 78
Голосований система (committee machine) 71
- Дескрипторы (descriptors) 107
молекулярной структуры 175
субструктурные 207
фрагментов 208
Динамический диапазон исходных данных 33, 36
Закономерностей распознавание (recognition of regularities) 9
- ИК-спектры** 34, 36
моделирование 125
Информации теория 33
- Карунена—Ловва* преобразование 14
см. также Анализ главных компонент
Квазиравновесия теория 172
Классификатор (classifier) 11, 15, 18
бинарный 19, 45, 67, 87, 178
— ветвящаяся схема 87
— линейный 47
— непараметрический 16
— параллельное соединение 90
кластерный (clustering) 17, 22
кусочно линейный (piecewise-linear) 17, 83, 100
послойный (layered) 17
Классификация 65
ветвящаяся (branching) 89, 96
кусочно линейная 100
множественная 101
на несколько категорий 86

- параллельная 90
по K ближайшим соседям (K -nearest neighbor) 17, 102
— методу наименьших квадратов 103
при помощи бинарного кода 93, 96
сопоставление схем 96, 101
- Кластеры (clusters) 17, 21
- Ковача индекс 102
- Код
бинарный (code binary) 93
трехбитовый 98
- Кодирование фрагментов структуры 174, 207
- Контрольная выборка (prediction set) 31
- Масс-спектры 19, 30, 37, 48, 70, 88, 109, 138, 142, 148
формирование 172
- Матрица
ковариационная (covariance) 13, 15
корреляционная (correlation) 159, 160
наблюдений (observation) 159
расстояний (distance) 126
смежности (adjacency) 138
сходства (similarity) 139
- Мертвая зона (dead zone) 63
- Метрика расстояний (distance metric) 125
- Множественные признаки (multiple features) 155
- Надежность (reliability) 27
- Наименьших квадратов метод 76
- Нормировочные процедуры (normalization procedures) 20
- Обнаружение (detection) 9
- Обучающая выборка (training set) 15, 26
- «Обучающаяся машина» («learning machine») 26
- Обучение (learning phase) 15, 26
без учителя (unsupervised training) 17
- Обучения методы
итерационные по методу наименьших квадратов 75
непараметрические (nonparametric) 15
параметрические (parametric) 15
с исправлением ошибки через обратную связь (error correction feedback) 25
- Отбор признаков (feature selection) 29
по знаку весовых векторов 111
— метрике расстояний 125
- Перекрестные члены (cross terms) 138
формирование 30
- Плотность распределения вероятности (probability density function) 16
- Полярограммы 42, 73, 124
- Пороговый логический элемент (threshold logic unit) 16, 22, 45
двухуровневый 71
кусочно линейный 80
простой 45
свойства 27
с ненулевым порогом 63
- Потенциальных функций (potential function) метод 17
- Преобразователь (transducer) 11
- Препроцессор (preprocessor) 11, 13, 137
- Препроцессорная обработка (preprocessing) исходных данных 29

- Применение распознавания образов 9
- Прогнозирующая способность (prediction) 27
- Программа
DENDRAL 172
PREDICTOR 173
- Пространство образов (pattern space) 12, 29
- Разделяющая функция (discriminant function) 16, 45
 комплексная нелинейная 166
 линейная 23, 45, 137
 непараметрическая 16
- «Распознавание знаков» («character recognition») 10
- Распознавание речи (speech recognition) 11
- Распознающая система (recognition system) 11
- Распознающая способность (recognition) 27
- Решающая функция (decision function) 15
- Решающие поверхности (decision surfaces) 20
- Сходимости скорость (convergence rate) 27, 33, 40
- Тангенс гиперболический 75
- Тейлора ряд 76
- «Текущий взвешенный интеграл» («running weighted integral») 84
- Трансгенерирование (transgeneration) 29
- Тренировка (training) 15 см. также Адаптация
- Уолша преобразование 166
- Факторный анализ (factor analysis) 30, 159
- Фурье-преобразование 14, 30, 147
 быстрое (БПФ-алгоритм) 148
- Хроматография газожидкостная 102
- Хэмминга код 93, 98
- Частичной коррекции правило (fractional correction rule) 82
- ЯМР-спектры 102, 158

СО Д Е Р Ж А Н И Е

Предисловие к русскому изданию	5.
Предисловие к американскому изданию	7
Глава 1. Введение	9
Области применения распознавания образов	9
Общая схема распознающей системы	11
Преобразователь	12
Препроцессор (устройство для выделения признаков)	13
Классификатор	15
Список литературы	18
Глава 2. Введение в теорию бинарных классификаторов образов	19
Векторы образов в гиперпространстве	19
Решающие поверхности	20
Пороговые логические элементы как бинарные классификаторы образов	22
Обучение пороговых логических элементов с исправлением ошибки через обратную связь	24
Свойства пороговых логических элементов	27
Глава 3. Предварительная обработка и преобразования исходных данных	29
Масс-спектры	30
Инфракрасные спектры	34
Спектроскопические данные из многих источников	35
Электрохимические спектры	42
Список литературы	44
Глава 4. Построение разделяющей функции	45
Бинарные классификаторы	45
Простой пороговый логический элемент	45
Пороговые логические элементы с ненулевым порогом	63
Система голосований	71
Применение пороговых логических элементов для классификации электрохимических данных	73
Итерационное обучение по методу наименьших квадратов	75
Кусочно линейные пороговые логические элементы	80
Классификация образов на несколько категорий	86
Дерево бинарных классификаторов образов	87
Параллельное соединение бинарных классификаторов образов	90
Классификация при помощи бинарного кода	93
Кусочно линейная классификация	100

Классификация по K ближайшим соседям	102
Классификация по методу наименьших квадратов	103
Список литературы	106
Глава 5. Отбор признаков	107
Отбор признаков по знаку весовых векторов	111
Отбор признаков при помощи метрики расстояний	125
Карбоновые кислоты	129
Сложные эфиры	131
Первичные амины	133
Список литературы	134
Глава 6. Дополнительные преобразования	136
Генерирование перекрестных членов	138
Преобразование Фурье	147
Факторный анализ	159
Комплексная нелинейная разделяющая функция	166
Список литературы	171
Глава 7. От молекулярной структуры к свойствам	172
Формирование масс-спектров	172
Кодирование фрагментов	174
Множественные признаки	198
Субструктурные дескрипторы	207
Список литературы	216
Приложение. Образец программы для моделирования обучающейся машины	217
Список монографий	224
Список дополнительной литературы	225
Предметный указатель	226

УВАЖАЕМЫЙ ЧИТАТЕЛЬ!

Ваши замечания о содержании книги, ее оформлении, качестве перевода и другие просим присылать по адресу: 129820, Москва, И-110, ГСП, 1-й Рижский пер., д. 2, издательство «Мир».

ИБ 514

П Джурс, Г. Айзензур

РАСПОЗНАВАНИЕ ОБРАЗОВ В ХИМИИ

Редактор Т. Румянцева
Художник С. Брынза
Художественный редактор В. Шаповалов
Технический редактор Н. Толстякова
Корректор Л. Байкова

Сдано в набор 30/III 1977 г. Подписано к печати
26/VII 1977 г. Бумага кн. журн. 60×84¹/₁₆=7,25
бум. л. Усл. печ. л. 13,49. Уч.-изд. л. 13.
Изд. № 3/8806

Цена 1 р. 70 к. Зак. 292.

ИЗДАТЕЛЬСТВО «МИР»

Москва, 1-й Рижский пер., 2

Ярославский полиграфкомбинат Союзполиграфпрома
при Государственном комитете Совета Министров
СССР по делам издательств, полиграфии и книж-
ной торговли. 150014, Ярославль, ул. Свободы, 97

Цена 1 р. 70 к.

