МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ АГЕНТСТВО ПО ОБРАЗОВАНИЮ

МОСКОВСКИЙ ЭНЕРГЕТИЧЕСКИЙ ИНСТИТУТ
(ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ)

В.О. ТОЛЧЕЕВ

СОВРЕМЕННЫЕ МЕТОДЫ ОБРАБОТКИ И АНАЛИЗА ТЕКСТОВОЙ ИНФОРМАЦИИ

Учебное пособие
по курсу
«Интеллектуальные информационные системы»
для студентов, обучающихся по специальности
«Управление и информатика в технических системах»

Москва Издательство МЭИ 2006

УДК
621.398
Т-547

Утверждено учебным управлением МЭИ в качестве учебного пособия для студентов

Подготовлено на кафедре управления и информатики

Рецензенты: докт. техн. наук, профессор Г.Ф. Филаретов (ГОСНИИСИ),
докт. техн. наук, профессор А.Б. Фролов (МЭИ)

Толчеев В.О.

Т-547 Современные методы обработки и анализа текстовой информации: учебное пособие / В.О. Толчеев. М.: Издательство МЭИ, 2006. – 76 с.

110-18

ISBN 5-7046-1285-7

В доступной форме излагаются основные вопросы, связанные с обработкой и анализом многомерных данных применительно к задачам классификации текстовых документов. Рассмотрена методология классификации документальной информации, включая вопросы формирования выборок, выявления информативных признаков и визуализации данных, оценки качества классификации.

Особое внимание уделяется непосредственно самим алгоритмам классификации и их сравнительному анализу. Специальные разделы учебного пособия посвящены методу ближайшего соседа и его модификациям.

Для студентов специальности «Управление и информатика в технических системах» МЭИ (ТУ).

ISBN 5-7046-1285-7

ВВЕДЕНИЕ

В настоящее время существенным образом возросла роль сети Интернет как одного из важнейших источников получения научно-технической информации и средства обмена новыми идеями. Значительно расширились возможности специалистов оперативно получать сведения по интересующим их тематикам, используя виртуальные библиотеки и специализированные порталы, электронные версии статей в профильных журналах и докладов на конференциях. Полученная из электронных источников информация может использоваться при подготовке диссертаций, публикаций, учебных курсов, выполнении НИОКР, существенно увеличивая эффективность научно-исследовательской деятельности.

Наращивание объемов информации, распространяемой в электронном виде, вызывает необходимость перехода от переработки информации на бумажных носителях к новым технологиям обработки больших массивов данных, выдвигает задачу разработки *методов интеллектуального анализа текстовых данных (Text Mining)*, в число самых приоритетных в области информатики. Эти методы основываются на статистических, эвристических подходах и принципах построения систем искусственного интеллекта.

Наиболее эффективным и востребованным на практике механизмом обработки и анализа текстовой неструктурированной информации (электронных файлов с монографиями, газетными и журнальными публикациями, письмами электронной почты, Web-страницами и т.п.) являются методы классификации.

Классификация текстовой информации заключается или в разбиении набора документов на обычно однородные непересекающиеся группы (классы) с целью обеспечения максимальной "близости" элементов одной группы и максимального различия между группами, или в отнесении нового документа к одному из сформированных классов. При этом в качестве *меры "близости"* чаще всего используются метрики расстояния: *евклидова метрика, квадрат евклидовой метрики, манхэттенское расстояние, метрика Чебышева*, а описание класса осуществляется двумя способами: через перечисление всех его элементов или через задание признаков, которые определяют членов данного класса.

Процесс *классификации* можно рассматривать как в *"широком смысле"*, так и *"узком смысле"*. В первом случае предполагается использование целого комплекса моделей и методов, согласно которым производится сбор, обработка, анализ и интерпретация результатов классификации текстовых документов. Во втором случае решается исключительно задача разработки классификаторов с заданными свойствами.

Несмотря на наличие серьезных научно-методических работ по обработке и анализу фактографической информации [1,2], теории распознавания образов [3,4], тем не менее в области разработки принципов, моделей, методов для проведения классификации в "широком смысле" ощущается недостаток специальной литературы, современных учебников и монографий. Долгое

время большая часть исследований по данному направлению основывалась на фундаментальной работе Солтона и его разработке SMART [5]. Однако появление новой парадигмы, связанной с феноменом Интернет, привело по существу к коллапсу традиционной системы обработки и анализа тестовой информации, показало невозможность решать "сегодняшние" проблемы "вчерашними" методами. К числу современных монографий, посвященных данной проблематике, необходимо отнести обзор функциональных возможностей сети Кохонена для группировки текстовых файлов, адаптацию нейросетевых технологий и разработку методов извлечения знаний из специализированных текстовых баз данных [6].

Таким образом, сеть Интернет требует создания новаторских методов выявления имеющихся знаний и усовершенствование, модификацию имеющихся способов обработки, представления и анализа больших массивов текстовой информации. В предлагаемом учебном пособии с системных позиций рассмотрены этапы проведения классификации в "широком смысле". Особое внимание уделено вопросам формирования обучающих выборок и обучения классификаторов, принципам отбора наиболее информативных признаков и визуализации текстовых документов, методам классификации, в частности, семейству методов ближайшего соседа, обладающих высокой точностью при группировании текстовых документов. В данное учебное пособие также включен ряд оригинальных результатов, сосредоточенных главным образом в третьей главе, которые были получены автором в процессе исследований.

Учебное пособие разделено на три главы. В первой главе излагается комплексная методология проведения обработки и анализа больших массивов текстовых данных, а также рассматриваются вопросы представления текстовых документов в виде математических моделей, выбора меры близости, формирования обучающих выборок и обучения классификаторов, определения точности результатов группирования. Вторая глава полностью посвящена различным принципам отбора наиболее информативных признаков и визуализации текстовых данных. Третья глава содержит обзор методов классификации текстовых документов, эффективно используемых на практике, и изложение новых модификаций, созданных на основе метода ближайшего соседа.

1. МЕТОДОЛОГИЯ КЛАССИФИКАЦИИ ТЕКСТОВОЙ ИНФОРМАЦИИ

В данном учебном пособии рассматривается комплексный подход к проблеме классификации текстовой информации в рамках единой методологии, обеспечивающей решение целой совокупности взаимосвязанных проблем: построения модели текстовых документов, формирования обучающих выборок, выявления информативных признаков, выбора метрики расстояния, визуализации многомерных данных, обучения классификаторов, оценки точности результатов классификации.

Таким образом, предлагаемая методология представляет собой набор последовательных шагов, объединенных в общую систему, подчиненных единой задаче и осуществляемых в интегрированной программной среде. На рис.1.1 показаны основные этапы, соответствующие проведению классификации в "широком смысле". Именно они будут подробно рассматриваться в данной работе.

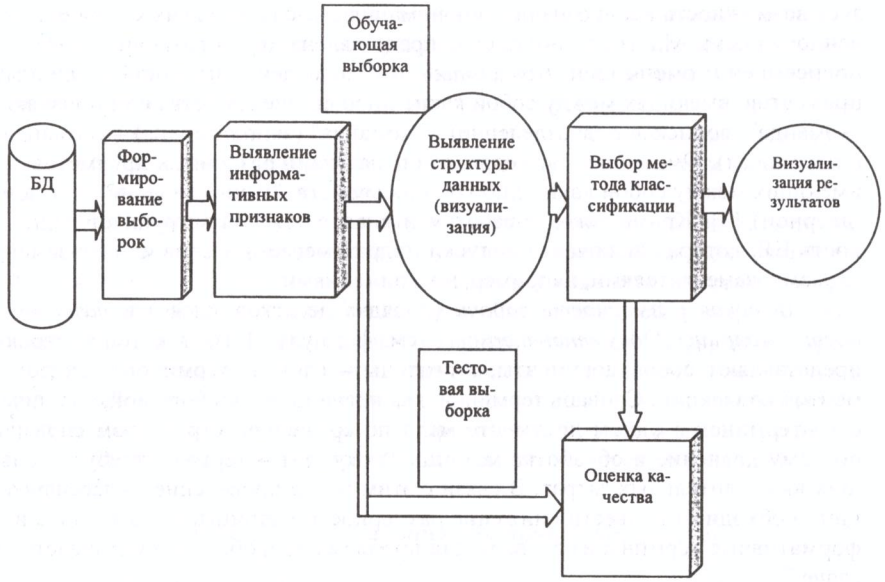

Рис. 1.1

1.1. Особенности классификации текстовой информации

Документы, поступающие на вход системы анализа текстовой информации, записаны на естественном языке (ЕЯ), который обладает существенными недостатками с точки зрения машинной технологии обработки текстовой информации [7].

Многообразие средств передачи смысла. Несмотря на то, что основным средством передачи смысла сообщения является лексика естественного языка, в сообщениях на ЕЯ функцию передачи смысла выполняет и ряд других элементов:

- контекст;
- ссылки на слова (словосочетания, фразы и т. д.), ранее упоминавшиеся в тексте сообщения.

Семантическая неоднозначность. Семантическая неоднозначность возникает в основном из-за синонимии и многозначности слов ЕЯ. *Синонимия* представляет собой тождественность или близость по значению слов, выражающих одно и то же понятие. Синонимами естественного языка являются как отдельные слова, так и словосочетания. *Многозначность* характеризует возможность неоднозначного понимания смысла отдельных слов естественного языка. Многозначность слов представлена двумя разновидностями – полисемией и омонимией. *Полисемия* – это совпадение названий различных предметов, имеющих между собой какие-либо общие свойства или признаки: "команда" (воинское подразделение) – "команда" (экипаж судна) – "команда" (спортивная). *Омонимия* – это совпадение названий различных предметов, не имеющих между собой каких-либо общих свойств: "ключ" (родник) – "ключ" (дверной) [7]. Кроме того, обработку и анализ текстов затрудняет эллипсность ЕЯ, которая включает пропуски подразумеваемых слов или их замену словами–заместителями, например, местоимениями.

Высокая размерность задачи (порядка десятков тысяч) и *разреженность матрицы "документ–термин"* (см. формулу (1.4)), в которой строки представляют собой документы, а столбцы – словарь терминов всех документов коллекции. Словарь терминов, как правило, очень большой, а количество терминов в одном документе мало по сравнению с размером словаря, поэтому хранение и обработка матрицы "документ – термин" требует больших вычислительных затрат. В связи с этим перед проведением классификации необходимо провести снижение размерности матрицы, т.е. выделить информативные термины из словаря (эта проблема подробно рассматривается в главе 2).

Субъективность оценки качества классификации. В каждой классификации имеются элементы как субъективного, так и объективного. Все реальные объекты имеют бесконечное число свойств (часто зависящих друг от друга), и выделение некоторого конечного подмножества этих свойств для описания классов, а также выбор меры близости обычно проводятся субъективно. Качество классификации, казалось бы, определяется объективно по тому, достигается ли поставленная цель, однако выбор цели опять-таки субъективен, и для одной цели данная классификация будет хорошей, а для другой нет [2].

Различная длина документов. Каждый документ состоит из различного числа терминов. Минимальное и максимальное количество терминов на документ внутри даже небольшой коллекции может существенным образом различаться, поэтому термины, встречающиеся в разных документах одинаковое количество раз, будут иметь разный вес. Отчасти эта проблема решается использованием различных мер взвешивания (см. главу 2).

В настоящее время существует два основных подхода к обработке и анализу текстовых документов: на основе *лингвистического анализа (ЛА)* или *статистического анализа (СА).*

Системы ЛА обычно состоят из модели предметной области, содержащей основные тематические термины и их взаимосвязи, а также специализированной базы данных (БД) грамматических конструкций и семантических правил, свойственных конкретному языку. При этом модель предметной области обычно используется для проведения морфологического анализа, а специализированная БД – для синтаксического и семантического анализа [8].

К сожалению, в современных публикациях отсутствует единое мнение по вопросу о том, какие сведения должны включаться в описание предметной области, а используемые на практике процедуры выявления смысла текста не отражают всей сложности языковых явлений и не дают удовлетворительной семантической интерпретации. Чаще всего методы ЛА позволяют лишь выделить определенные связанные группы слов внутри отдельных предложений, однако отношения между отдельными фразами остаются в большинстве случаев неизвестными.

Трудности в определении смысла текста, ориентация на достаточно узкую проблематику, отраженную в модели предметной области, ограничивают возможности использования лингвистического анализа для решения прикладных задач. В этой связи большинство исследователей и разработчиков алгоритмов для обработки больших массивов текстовой информации предпочитают использовать СА. Он заключается в представлении документа в виде набора ключевых слов, вес которых зависит от частоты встречаемости термина в текстовом файле. При этом предполагается, что появление одних и тех же терминов в различных документах свидетельствует об их подобии.

1.2. Модели представления текстовых документов

Под *текстом* принято понимать конечное множество слов (терминов), объединенных лексическими, грамматическими, смысловыми, частотными отношениями и образующих информативное сообщение.

Наиболее часто в качестве модели представления текстовых документов используется так называемый *"мешок слов" (неструктурированная модель "bag of words")*. В этой модели каждый термин рассматривается в качестве независимой случайной величины вне контекста и связи с другими словами текста [5]. При этом вес термина определяется частотой его встречаемости.

В задачах текстовой классификации также широкое применение получили *частично структурированные модели.* В них учитывается дополнительная информация о положении слова в документе (заголовок, аннотация, первый абзац), связь между словами типа *"hypernym"* (*"a knife is a weapon"* – нож – это оружие) или проводится выделение словосочетаний – устойчивых групп слов, которые образуют общее понятие для данной предметной области [5].

Для выявления словосочетаний применяются различные подходы. Так, в [5] используется отношение частот:

$$\omega = \frac{\omega_{kj}}{\omega_k \omega_j}, \qquad (1.1)$$

где ω_{kj} – совместная встречаемость терминов k и j в исследуемой выборке документов; ω_k и ω_j – частота терминов k и j в исследуемой выборке документов; ω – вес словосочетания. При расчете ω_k не обязательно, чтобы оба термина следовали друг за другом, их могут разделять случайные или незначащие термины. В таких случаях вес ω словосочетания корректируется по формуле:

$$\omega = \frac{1}{2^t}\omega_k \omega_j, \qquad (1.2)$$

где ω_k и ω_j – веса терминов k и j в словосочетании, а t – количество незначимых слов между ними.

Третий вид моделей, применяемый для представления текстовых документов, – *полностью структурированные модели*. В таких моделях используются заранее сформированные базы знаний, содержащие ключевые термины, их словосочетания, а также иерархические связи, свойственные какой-либо предметной области. Традиционно для этого разрабатываются *иерархические тезаурусы (hierarchical thesaurus),* на верхних уровнях которых находятся ключевые термины предметной области, уточняемые на более низких уровнях [5].

Другой подход к разработке полностью структурированных моделей реализован в *онтологии (ontology).* Онтология содержит определения понятий и их иерархическую организацию (отношения между понятиями: класс–вид, часть–целое и т.п.). В отличие от обычной базы знаний, онтологии содержат неизменные знания (аксиомы), которые всегда истинны для данной предметной области. Относительно остальных знаний, имеющихся о предметной области, предполагается, что они могут изменяться с течением времени. Для решения задач текстовой классификации наиболее часто используется онтология *WordNet*. На практике в качестве полностью структурированных моделей также используются *ассоциативные семантические сети*, представляющие собой множество понятий – слов и словосочетаний, сила связи между которыми определяется частотой их совместной встречаемости.

Сравнительный анализ *неструктурированных, частично структурированных, полностью структурированных моделей* и их модификаций показал, что вид модели не оказывает значительного влияния на качество классификации различными методами [5]. Несмотря на более высокую сложность, дополнительные вычисления, полностью структурированные модели не всегда позволяют полностью формализовать контекст, разрешить синонимию, ввести адекватную иерархию и показывают для ряда предметных областей более низкую точность, чем неструктурированные и частично структурированные модели.

В качестве математических аппроксимаций в задачах классификации текстовых документов, наибольшее применение получили *векторная* и *вероятностная* модели.

В *векторной* модели любой документ описывается в виде точки в M-мерном пространстве, где M – количество признаков (терминов) [5]:

$$\vec{X}_j = \begin{bmatrix} x_j^{(1)} \\ \vdots \\ x_j^{(i)} \\ \vdots \\ x_j^{(M)} \end{bmatrix}, \qquad (1.3)$$

где $x_j^{(i)}$ – вес термина i в документе j (j=1,...,N; N – количество документов в выборке; i=1,...,M; M – количество признаков).

В качестве весов $x^{(i)}$ в векторной модели могут использоваться не только сами термины, но и последовательности слов или букв (*n-граммы*) [6].

Выборка текстовых документов может быть представлена в виде матрицы:

$$X = \begin{bmatrix} x_1^{(1)} & \vdots & x_1^{(i)} & \vdots & x_1^{(M)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_j^{(1)} & \vdots & x_j^{(i)} & \vdots & x_j^{(M)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_N^{(1)} & \vdots & x_N^{(i)} & \vdots & x_N^{(M)} \end{bmatrix}. \qquad (1.4)$$

Такую матрицу принято называть *матрицей "документ – термин"*, т.к. ее строки представляют собой документы, а столбцы – термины, содержащиеся в этих документах. Для представления текстовых документов также могут быть использована матрица попарных расстояний (близостей):

$$A = \begin{bmatrix} a_{11} & \vdots & a_{1j} & \vdots & a_{1N} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{l1} & \vdots & a_{lj} & \vdots & a_{lN} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{N1} & \vdots & a_{Nj} & \vdots & a_{NN} \end{bmatrix}. \quad (1.5)$$

Элемент a_{lj} определяет результат сопоставления документов $\vec{X}_l$ и $\vec{X}_j$ в смысле некоторого заданного отношения (метрики расстояния или меры близости). Формулы для вычисления метрик расстояния и мер близости приведены в п. 1.3, а методы определения информативных признаков и присвоения весов рассмотрены в главе 2.

В *вероятностной* модели принадлежность документа $\vec{X}$ к классу Q_k определяется как вероятность принадлежности каждого из терминов $x^{(i)}$ к классу Q_k (k=1,...,K; K – количество классов) [1,4]:

$$P(Q_k \mid \vec{X}) = \frac{P(Q_k)P(\vec{X} \mid Q_k)}{P(\vec{X})}, \quad (1.6)$$

где $P(Q_k)$ – априорная вероятность появления документов класса Q_k; $P(\vec{X})$ рассчитывается по формуле полной вероятности появления события $\vec{X}$; $P(\vec{X} \mid Q_k) = \prod_{i=1}^{M} P(x^{(i)} \mid Q_k)$.

Вероятность $P(\vec{X})$ одинакова для различных классов и может быть исключена из дальнейшего рассмотрения, тогда формула (1.6) перепишется в виде

$$P(Q_k \mid \vec{X}) = P(Q_k)\prod_{i=1}^{M} P(x^{(i)} \mid Q_k). \quad (1.7)$$

Согласно вероятностной модели документ принадлежит к тому классу, для которого величина $P(Q_k \mid \vec{X})$ максимальна.

Обратим внимание на то, что далее в учебном пособии термины: "*документ*", "*объект*", "*пример*", "*элемент*" (выборки), (многомерная) "*точка*" используются как *синонимы*, то же самое относится к понятиям "*классификация*", "*группировка*", "*систематизация*", "*распознавание*" и к понятиям "*слово*", "*термин*", "*признак*", "*координата*", "*атрибут*".

1.3. Меры близости и расстояния

Меры расстояния и близости дают численные значения, которые характеризуют расположение различных точек в M-мерном пространстве. Точность классификации с помощью большинства методов существенным образом зависит от вида используемой метрики. Далее приводятся формулы для вычисления наиболее эффективных мер расстояния.

- *Евклидово расстояние (L_2-метрика).* Это наиболее часто используемая метрика, соответствующая простому геометрическому расстоянию в многомерном пространстве. Евклидово расстояние определяется по формуле:

$$d(\vec{X}_j, \vec{X}_l) = \sqrt{\sum_{i=1}^{M} (x_j^{(i)} - x_l^{(i)})^2} \quad (1.8)$$

- *Квадрат евклидова расстояния.* В ряде случаев для придания большего веса различию между признаками евклидово расстояние возводится в квадрат:

$$d(\vec{X}_j, \vec{X}_l) = \sum_{i=1}^{M} (x_j^{(i)} - x_l^{(i)})^2. \quad (1.9)$$

- *Расстояние городских кварталов (манхэттенское расстояние) (L_1-метрика).* Эта метрика представляет собой сумму разностей по координатам. В большинстве случаев она приводит к таким же результатам, как и в случае использования обычного евклидова расстояния, хотя влияние отдельных больших разностей в значении признаков уменьшается. Манхэттенское расстояние вычисляется по формуле:

$$d(\vec{X}_j, \vec{X}_l) = \sum_{i=1}^{M} | x_j^{(i)} - x_l^{(i)} |. \quad (1.10)$$

- *Расстояние Чебышева (L_∞-метрика).* Такое расстояние может оказаться полезным, когда необходимо определить два объекта в разные классы в зависимости от различий между ними по одному наиболее значимому признаку. Расстояние Чебышева вычисляется по формуле:

$$d(\vec{X}_j, \vec{X}_l) = \max\{| x_j^{(i)} - x_l^{(i)} |\}. \quad (1.11)$$

Кроме того, для определения "близости" между текстовыми документами часто используется *косинусоидальная мера близости* (косинус угла между векторами) [8]:

$$d(\vec{X}_j, \vec{X}_l) = \cos(\vec{X}_j, \vec{X}_l) = \frac{\sum_{i=1}^{M} x_j^{(i)} x_l^{(i)}}{\sqrt{\sum_{i=1}^{M} (x_j^{(i)})^2 \sum_{i=1}^{M} (x_j^{(i)})^2}}. \quad (1.12)$$

В отличие от метрик расстояния, которые для наиболее похожих документов будут стремиться к нулю, косинусоидальная мера близости в этом случае будет стремиться к единице.

1.4. Обучение классификаторов на конечных выборках

Существует два различных метода обучения: один из них – объяснение, другой – обучение на примерах. Первый метод предполагает существование достаточно простых правил, которые можно изложить так, чтобы, действуя сообразно этим правилам, каждый раз получать требуемый результат. Однако во многих случаях *"учитель"*, проводящий обучение, не может сформулировать правило, по которому он действует, тогда первый способ обучения неприменим и обучение проводится на примерах (индуктивно).

Разработка моделей, методов и алгоритмов, которые позволяют получить применимые в будущем правила и закономерности исходя из имеющихся в наличие прошлых примеров, проводится в рамках работ по *автоматическому накоплению и формированию знаний (Machine Learning)*. Данный подход особенно эффективен, когда учитель не существует, недостаточно надежен или его услуги очень дороги.

Последовательность примеров с указанием, к какому классу они относятся, называется обучающей выборкой. Целью обучения является выработка правила классификации (решающего правила), позволяющего проводить распознавание так же хорошо, как это делает "учитель". Качество решающего правила определяется количеством несовпадений (ошибкой) при классификации элементов выборки [1,4].

Обычно на практике требуется правильно распознавать как можно больший процент встречающихся, а не всех возможных ситуаций. Дело в том, что некоторые ситуации, описываемые примерами, появляются чаще и именно их важно классифицировать правильно, другие ситуации, хотя и возможны, но встречаются сравнительно редко и их влияние на ошибку меньше. Для математического описания данного факта на множестве всех возможных ситуаций $\mathbf{X}$ вводится функция распределения вероятностей $F(X)$, которая каждой возможной ситуации ставит в соответствие вероятность ее появления среди всех элементов множества. Тогда ошибка на элементе $\vec{X}_j$ ($\vec{X}_j \in \mathrm{X}$) может быть оценена величиной пропорциональной вероятности появления этой ситуации. Для каждого алгоритма обучения можно подсчитать средние потери от всех его ошибок и выбрать тот, который обеспечивает минимальную вероятность ошибок при классификации. При этом функция $F(X)$ по существу является характеристикой той среды, в которой предстоит работать методу классификации после обучения.

Основным условием формирования обучающей последовательности является то, что в нее включаются элементы, которые были случайно и независимо извлечены из генеральной совокупности (гипотетического множества всех возможных объектов каждого класса). Однако при таком случайном подборе элементов обучающей выборки уже нельзя требовать, чтобы обучение было безусловно успешным, так как не исключена вероятность того, что обучающая последовательность будет составлена только из "нетипичных" элементов. Поэтому успех в обучении может быть гарантирован не наверняка и во многом определяется тем, какие элементы попали в обучающую выборку.

1.5. Формирование обучающих выборок

Из вышеизложенного следует, что эффективность обучения методов классификации во многом определяется тем, каким образом были сформированы выборки. Под выборкой понимается последовательность независимых пар наблюдений вида $\{\vec{X}_j, Q_k\}$ (j=1,...,N; k=1,...,K, где N – объем выборки; K – количество классов), в которой определено, к какому классу Q_k относится каждое многомерное наблюдение $\vec{X}_j$.

Представительной (репрезентативной) принято считать обучающую выборку, которая в заданном пространстве признаков и заданном классе решающих функций позволяет построить правило распознавания новых наблюдений, удовлетворяющее принятому критерию качества (см. п.1.7.) [2]. В общем случае возникает задача формирования не только обучающей, но и экзаменационной выборки, т.е. необходимо обучить классификатор минимизировать ошибку как на обучающем множестве, так и ошибку, возникающую при классификации реальных данных.

При одном и том же размере представительной выборки качество обучения может быть разным в зависимости от того, какие части генеральной совокупности в ней представлены. Если распознаются наблюдения, принадлежащие двум различным классам, то не так важно, как выглядит генеральная совокупность во всем пространстве признаков. Гораздо важнее, как она выглядит в районе границы между двумя классами. Поэтому формирование обучающей выборки желательно проводить таким образом, чтобы ее элементы принадлежали той части признакового пространства, где имеется наибольший риск получить ошибочное решение.

Сформированные выборки могут быть представлены в виде нескольких моделей: *"ядерной"* (или центроидной), где все элементы выборки компактно сгруппированы около центра (рис.1.2, а) и *моделью "рассеяния"*, когда существуют элементы, находящиеся в зоне неопределенности около границ классов (рис.1.2, б) (иногда также используется *модель "засорения"*, в которой наряду с компактно расположенными элементами класса присутствуют достаточно далекие выбросы).

а) б)

Рис. 1.2

Для правильного выбора метода классификации и метрики близости необходимо определить, какой из моделей описывается исходная выборка. В этих целях можно эффективно использовать методы визуализации текстовых

данных (см. п. 2.7) или функционалы качества, в которых вычисляются численные показатели, характеризующие расположение классов в многомерном пространстве признаков. Обычно используются следующие показатели качества [1].

1. Средняя сумма внутриклассовой дисперсии:

$$Q_1 = \frac{1}{N_k}\sum_{j=1}^{N_k} d^2(\vec{X}_j, \vec{X}_k) \quad \text{или} \quad Q_1^* = \frac{1}{M}\sum_{k=1}^{M}\frac{1}{N_k}\sum_{j=1}^{N_k} d^2(\vec{X}_j, \vec{\bar{X}}_k), \qquad (1.13)$$

где M – количество классов; N_k – количество примеров в k-м классе; $\vec{X}_k$ – центроид k-го класса; $d^2(\vec{X}_j, \vec{X}_k)$ – квадрат евклидова расстояния между элементами k-го класса и его центроидом.

Этот функционал оценивает, насколько близко к центроиду расположены наблюдения внутри каждого из классов.

2. Средняя сумма квадратов внутриклассовых попарных расстояний:

$$Q_2 = \frac{1}{N_k}\sum_{l=1}^{N_k}\sum_{j=1, j\neq l}^{N_k} d^2(\vec{X}_l, \vec{X}_j) \text{ или } Q_2^* = \frac{1}{M}\sum_{k=1}^{M}\frac{1}{N_k}\sum_{l=1}^{N_k}\sum_{j=1, j\neq l}^{N_k} d^2(\vec{X}_l, \vec{X}_j) \quad (1.14)$$

Чем меньше этот функционал, тем более компактно расположены примеры внутри класса.

3. Средняя сумма квадратов межклассовых попарных расстояний:

$$Q_3 = \frac{1}{N_k N_s}\sum_{l=1}^{N_k}\sum_{j=1, j\neq l}^{N_s} d^2(\vec{X}_l, \vec{X}_j) \quad \text{или}$$

$$Q_3 = \frac{1}{M}\sum_{k=1}^{M}\sum_{s=1, s\neq k}^{M}\frac{1}{N_k N_s}\sum_{l=1}^{N_k}\sum_{j=1, j\neq l}^{N_s} d^2(\vec{X}_l, \vec{X}_j). \qquad (1.15)$$

Чем больше этот функционал, тем лучше разделены между собой примеры разных классов (N_s – количество наблюдений в s-ом классе).

4. Обобщенный функционал: $Q_4 = \frac{Q_3}{Q_2}$. (1.16)

Учитывая смысл функционалов Q_3 и Q_2, очевидно, что чем больше данное выражение, тем более компактно расположены классы в выборке.

На основе расчета значений функционалов качества (а также анализа диаграмм рассеивания, см. п. 2.7) исследователь может провести некоторые изменения в выборке: 1) объединить несколько близких небольших классов в один; 2) удалить "нехарактерные" шумовые элементы, расположенные вдалеке от центра классов (провести фильтрацию модели "выбросов"); 3) заново сформировать выборку, увеличив (уменьшив) количество классов или количество элементов. Вместе с тем при удалении шумовых точек необходимо учитывать, что такие выбросы могут быть не только обособленными "нехарактерными" наблюдениями, но представителями подгрупп, о которых в выборке содержится мало данных, следовательно, анализ на предмет их исключения должен быть особо тщательным.

В ряде случаев процесс обучения классификаторов удается осуществить в два этапа [2]. На первом этапе строится решающее правило с использованием имеющейся обучающей выборки. Затем система распознавания переводится в режим опытной эксплуатации. При появлении ошибки состав обучающей выборки дополняется элементом, вызвавшим ошибку, и решающее правило корректируется. Так продолжается до тех пор, пока частота появления ошибок не снизится до приемлемого уровня.

Сформированные обучающие выборки обладают следующими особенностями:

- любая обучающая выборка конечного размера не является полной, т.е. не содержит необходимого количества элементов для проведения безошибочной классификации;
- элементы обучающей выборки обычно имеют произвольное распределение в пространстве признаков и, как следствие, получаемые решающие правила могут обладать неодинаковой достоверностью в различных областях изменения параметров;
- базы данных текстовых документов, из которых чаще всего составляются обучающие выборки, как правило, содержат шумовые (нерелевантные, не относящиеся к указанным классам) элементы, другую противоречивую или ошибочную информацию, которая так или иначе попадает в обучающую выборку.

Общие рекомендации по выбору количества данных для обучения заключаются в том, что для получения устойчивых результатов необходимо увеличивать объем выборки N с целью уменьшения соотношений M/N (M – количество признаков) и K/N (K – количество классов). Все эти закономерности далеко не всегда справедливы в случае классификации текстовых документов. В первую очередь, это связано с тем, что размерность задачи очень высока и количество информативных признаков может достигать десятков тысяч. При этом увеличение размера обучающей выборки может столкнуться с вычислительными сложностями, так как для многих методов классификации затраты на вычисления растут нелинейно от числа признаков и количества наблюдений.

Приписывание пользователем (исследователем) документа к тому или иному классу может носить субъективный и дискуссионный характер, поэтому необходимо составлять выборки из БД, которые имеют свой встроенный общепризнанный и авторитетный рубрикатор. Такой рубрикатор позволяет заменить индивидуальное мнение пользователя о классе документа на совокупное мнение нескольких независимых экспертов. Однако даже использование экспертных оценок не позволяет полностью избежать ошибок, количество которых зависит как от компетенции самого эксперта, так и ряда "мешающих" случайных факторов (усталость, нехватка времени и т.п.). Если экспертные оценки носят противоречивый характер и не надежны, то возникает задача обучения с "неидеальным учителем" [3].

1.6. Методы минимизации ошибки классификации

Методы определения достоверности получаемых оценок на основе теории математической статистики являются асимптотическими, т.е. предполагают, что с ростом размера выборки можно как угодно близко подойти к оптимальному решению. Это положение, широко используемое для идентификации параметров динамических систем и временных рядов, может также успешно применяться для определения точности методов классификации [9]. Однако на практике исходные выборки имеют ограниченный размер и в классификации (также как в идентификации) возникает задача построения не асимптотически-оптимального алгоритма обучения классификатора, а конечно-оптимального.

Сформулируем проблему обучения в терминах математической статистики. Предположим, что из генеральной совокупности многомерных величин $\vec{X}$ (с распределением вероятностей $F(\vec{X})$) случайно и независимо формируется конечная выборка. Условная вероятность $P(Q_k \mid \vec{X})$ того, что наблюдение $\vec{X}$ относится к классу Q_k, может рассматриваться как *решающее правило (правило классификации)*. Будем считать, что это *"истинное"* правило классификации известно *"учителю"*, который передает свои знания при обучении классификатору, т.е. проводится так называемая *"классификация с учителем"*. Задача обучения классификатора заключается в том, чтобы подобрать наилучшее из множества всех решающих правил (классификаторов) $\{J(\vec{X},\alpha)\}$, где α – настраиваемый параметр, в общем случае таких параметров может оказаться несколько. Схема процедуры обучения представлена на рис. 1.3.

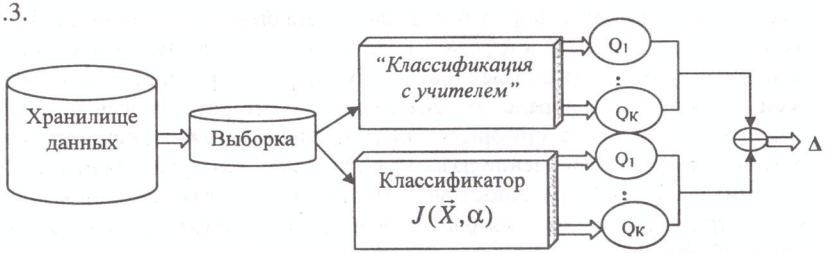

Рис. 1.3

Для каждого $J(\vec{X},\alpha)$ может быть определена точность классификации как вероятность ошибки распознавания на различных элементах выборки [3]:

$$P(\alpha)=\sum_{k=1}^{K}\sum_{j=1}^{N}\left(Q_k-J(\vec{X}_j,\alpha)\right)^2 P(\vec{X}_j)P(Q_k\mid\vec{X}_j), \qquad (1.17)$$

где j=1,...,N; k=1,...,K; N – число документов в выборке и K – число классов.

Среди всех функций $J(\vec{X},\alpha)$ есть такая $J(\vec{X},\alpha_0)$, которая минимизирует вероятность ошибки $P(\alpha)$. Именно ее необходимо найти по обучающей последовательности фиксированной длины N, т.е. необходимо выяснить при каких значениях параметра α *средняя величина потерь* Δ (или *величина среднего риска*) будет минимальной. Определим квадратичную функцию потерь:

$$L_1(Q,X,\alpha)=L_1(z,\alpha)=\left(Q_k-J(\vec{X}_j,\alpha)\right)^2. \qquad (1.18)$$

Здесь введена новая переменная $z=\{\vec{X}_j,Q_k\}$, которая имеет размерность (1+M), M – число признаков. Функция потерь характеризует степень "недоученности" алгоритма и несовершенства решающего правила.

В теории статистического оценивания используются также другие функции потерь [4]:

$$L_2(z,\alpha)=\left|Q_k-J(\vec{X}_j,\alpha)\right|; \qquad (1.19)$$

$$L_3(z,\alpha)=\begin{cases}0, & Q_k=J(\vec{X}_j,\alpha)\\ 1, & Q_k\neq J(\vec{X}_j,\alpha)\end{cases}. \qquad (1.20)$$

В [3] рассматривается несколько основных способов минимизации функции потерь. Первый подход связан с восстановлением функции распределения вероятностей. В общем случае это крайне трудоемкая процедура, однако вычислительная сложность может быть сокращена при использовании некоторых предположений: 1) о независимости признаков между собой $P(\vec{X})$ = $P(x^{(1)})*P(x^{(2)})*\ldots*P(x^{(M)})$ (задача сводится к так называемому наивному байесовскому методу, см. п.3.3); 2) о нормальном распределении элементов по классам с одинаковыми ковариационными матрицами (далее задача решается с помощью дискриминантного анализа, см. [10]).

Второй подход предусматривает проведение минимизации функционала $R(\alpha)=\int L(z,\alpha)dP(z)$, например, с помощью градиентных процедур или метода стохастической аппроксимации [3].

Третий подход связан с минимизацией, вместо функционала $R(\alpha)$, *функции эмпирического риска*. В случае квадратичной функции потерь минимизация будет проводиться по формуле

$$\hat{R}(\alpha)=\frac{1}{N}\sum_{j=1}^{N}L_1(\vec{X}_j,\alpha)=\frac{1}{N}\sum_{j=1}^{N}\left(Q_k-J(\vec{X}_j,\alpha)\right)^2. \qquad (1.21)$$

Для каждого фиксированного параметра α формула (1.21) определяет среднюю величину потерь при их оценке по экзаменационной выборке, проведении *"скользящего контроля"* или *статистического моделирования (Bootstrap)* (см. п. 1.7).

Таким образом, ошибка классификатора зависит от структуры и размера выборки, способа определения ошибок, а также множества решающих правил, среди которого проводится поиск наилучшего.

1.7. Критерии качества классификации

В качестве критериев качества классификации принято использовать *ошибку обучения*, оценка которой определяется по обучающей выборке, и *ошибку обобщения*, оценка которой рассчитывается по экзаменационной выборке. Величина ошибки обучения может служить мерой корректности использования конкретного классифицирующего правила для анализа данной выборки, т.е. представляет собой ошибку *ad hoc* (на этот случай). Однако ее малое значение не гарантирует малости ошибки обобщения, т.е. количества ошибок, которые классификатор допускает на примерах, не использовавшихся при обучении. По существу ошибка обобщения показывает мощность классифицирующего правила – на какое количество новых наблюдений удается распространить закономерности и эвристики, полученные во время обучения [11].

На практике используются три основных подхода для получения оценки ошибки обобщения.

1. *Оценка по экзаменационной выборке.* Исследуемую выборку делят на две части. По первой части (обучающая выборка) производится оценка параметров и построение решающего правила (обучение метода), по второй части (экзаменационная выборка) – определение оценки вероятности ошибок. К недостаткам этого способа следует отнести уменьшение объема выборки, используемой при классификации (соответственно уменьшается и возможное число оцениваемых параметров в алгоритме). К достоинствам этого способа относятся его несмещенность и универсальность.

2. *Оценка с помощью скользящего контроля (экзамена).* Для оценки параметров используется вся выборка кроме одного элемента, а оставшийся элемент используется для проверки. Затем этот элемент включается в общую выборку, а для контроля извлекается другой элемент. Данная процедура повторяется для всех членов исходной выборки. Таким образом, объем обучающей выборки все время равен N–1 элементу, где N – объем всей выборки. Недостатком скользящего контроля является существенных рост количества необходимых вычислений, достоинством – несмещенность получаемых оценок.

3. *Метод статистического моделирования (Bootstrap)* . Предложен в рамках разработки *методов коллективной классификации (boosting and bagging)* (см. п.3.5) для формирования обучающих и экзаменационных выборок в условиях ограниченного количества наблюдений с целью проведения многократного обучения и тестирования. При этом обучающая выборка принимается за генеральную совокупность и из нее случайным образом производится составление обучающих и экзаменационных подвыборок [1].

1.7.1. Разложение ошибки классификатора на смещение и дисперсию

В последнее время в ряде работ, содержащих анализ точностных свойств различных методов классификации (прежде всего *методов коллективной классификации)* вводится, по аналогии с регрессионными методами, понятия *смещения и дисперсии классификатора*. При этом предложено несколько различных способов декомпозиции ошибки классификатора, которые дают отличающиеся друг от друга определения смещения и дисперсии (в целом данная проблема еще находится в стадии обсуждения специалистами).

Как известно, задача идентификации параметров $\vec{\theta}$ динамических систем (или временных рядов) заключается в определении оценок $\hat{\vec{\theta}}$, которые являются наиболее точными приближениями к истинным значениям $\vec{\theta}$ (рис. 1.4) [9,12].

Рис.1.4

Средний квадрат ошибки для параметра θ может быть разложен на квадрат смещения и дисперсию [9]:

$$\Delta^2 = M\left[\left(\hat{\theta}-\theta\right)^2\right] = \textit{смещение}^2 + \textit{дисперсия} \ ,$$

где $M\left[\left(m(\hat{\theta})-\theta\right)^2\right]$ – квадрат смещения; $M\left[\left(\hat{\theta}-m(\hat{\theta})\right)^2\right]$ – дисперсия; θ – значение истинного параметра; $\hat{\theta}$ – оценка параметра; M[] – означает знак математического ожидания.

Определения, которые справедливы в теории статистического оценивания, нельзя полностью применить к задаче классификации данных, однако возможно, сохраняя их общий смысл, провести аналогию. Сформулируем задачу классификации по имеющейся выборке фиксированного размера $\{\vec{X}_j, Q_k\}$ (j=1,...,N): необходимо найти наилучшее решающее правило отнесения многомерного вектора $\vec{X}_j$ к классу Q_k, обеспечивающее наименьшее количество ошибок. В отличие от теории статистического оценивания, в которой представляется возможным определить степень соответствия выхода процесса и модели (например, в процентах), в задачах классификации ошибка принимает только два значения: “0” – классификация правильная и “1” – классификация неправильная. Поэтому в качестве функции потерь чаще все-

го используется не квадратичная функция, а функция потерь *"ноль или единица"* (см. формулу (1.20)).

Итак, задача заключается в определении такого решающего правила $J(\vec{X},\alpha_0)$, которое минимизирует количество ошибок на данной обучающей выборке. Введем понятие *оптимального классификатора (optimal classifier).* Оптимальным называется такой классификатор, в котором каждое наблюдение $\vec{X}_j$ наилучшим образом (с минимальным количеством ошибок) относится к правильному классу Q_k. Как известно [1,4], наименьшее количество ошибок распознавания обеспечивает байесовский классификатор $J(\vec{X},\alpha_0)$, выбирающий тот класс, который имеет наибольшую условную вероятность $P(Q_k \mid \vec{X})$.

Однако настраиваемый параметр α_0 будет изменять свое значение на различных выборках. Поэтому наряду с оценкой значения α на конкретной выборке имеет смысл рассматривать средние (ожидаемые) значения этого параметра, полученные путем усреднения по всем выборкам данного объема. Для анализа точности группировки на множестве различных выборок введем понятие обобщенного (главного) классификатора *(aggregated (main) classifiers).* Обобщенным называется такой классификатор $J(\vec{X},\alpha^*)$, который имеет на всех выборках фиксированной длины наименьшую ошибку. Для случая квадратичной функции потерь обобщенным классификатором является среднее от результатов использования всех классификаторов (решающих правил), для функции, в которой потери рассчитываются по модулю, – медиана, а для функции потерь "ноль или единица" – мода, т.е. тот класс, который наиболее часто был предсказан всеми классификаторами $\{J(\vec{X},\alpha)\}$. Например, если проводилось обучение на S выборках и 0,6S классификаторов предсказывают k-ый класс, а 0,4S классификаторов – q-й класс, то обобщенный классификатор в случае функции потерь "ноль или единица" будет предсказывать k-й класс.

Дадим определения смещению и дисперсии в задачах классификации. Под смещением понимается погрешность обобщенного классификатора в определении класса по отношению к предсказанию оптимального классификатора:

$$\Delta_0 = Q^{\text{обобщ}} - Q^{\text{оптим}}. \qquad (1.22)$$

Здесь Δ_0 – смещение; $Q^{\text{обобщ}}$ – результат классификации (предсказанный класс) с помощью обобщенного классификатора; $Q^{\text{оптим}}$ – результат классификации с помощью оптимального (байесовского) классификатора.

Классификатор является несмещенным, если $Q^{\text{обобщ}} = Q^{\text{оптим}}$ и на различных выборках длины N он предсказывает правильные классы. Смещение обуславливается систематическими погрешностями классификатора, которые не зависят от выборочных данных, и при использовании функции потерь "ноль или единица" всегда равняется единице. В то же время смещение зависит от вида решающего правила, настраиваемых параметров, информативных признаков, используемых для классификации.

Под дисперсией понимается отклонение в предсказании класса классификатором $\hat{Q}$ по отношению к результату, полученному обобщенным классификатором:

$$\sigma^2 = \hat{Q} - Q^{\text{обобщ}}. \qquad (1.23)$$

Дисперсия отражает зависимость классификатора от выборочных колебаний и существенным образом зависит от наличия в выборке нерелевантных шумовых элементов. Дисперсия равна нулю, если решающее правило делает одно и то же предсказание вне зависимости от используемой выборки.

Таким образом, в случае использования функции потери "ноль или единица" наиболее корректная декомпозиция ошибки классификатора $J(\vec{X},\alpha)$, построенного по выборке длины N, имеет вид

$$\Delta_N^{\alpha} = \Delta_0 + \sigma^2 + \Delta_1. \qquad (1.24)$$

Здесь Δ_0 – смещение; σ^2 – дисперсия; Δ_1 – составляющая, которая характеризует "неустранимую" ошибку классификации (например, отнесение "учителем" объекта к неправильному классу или невозможность корректного определения принадлежности шумового наблюдения к какому-либо классу, представленного в выборке и т.п.). Расчет смещения и дисперсии предполагает формирование множества выборок фиксированной длины с известным законом распределения. Для этого обычно используется статистическое моделирование (*Bootstrap*) [1].

1.7.2. Использование меры полнота–точность для анализа точности классификатора

Массив анализируемых документов часто бывает неоднородным: содержит нерелевантные примеры, а точность отнесения примеров к различным классам зависит от их размера и структуры (в первую очередь, вида зоны неопределенности в области границ классов). Часто возникает необходимость оценить долю документов одного класса, которые были правильно распознаны по отношению к их общему количеству в выборке, или определить, как много посторонних элементов при группировке были включены в класс Q_k, хотя ему и не принадлежат, т.е. ввести меры, которые позволили бы проводить дополнительный анализ точности классификации ("чувствительности" решающего правила).

Для этого в задачах текстовой классификации рассчитываются коэффициенты полноты и точности аналогично тому, как это делается в теории информационного поиска [8]. При этом составляется таблица сопряженности (табл. 1.1), в которую включаются:

- a – количество документов, принадлежащих классу Q_k и отнесенных к классу Q_k;

- b – количество документов, не принадлежащих классу Q_k, но отнесенных к классу Q_k;
- c – количество документов, принадлежащих классу Q_k, но не отнесенных к классу Q_k;
- d – количество документов, не принадлежащих классу Q_k и не отнесенных к классу Q_k.

Таблица 1.1

Принадлежность документа	Документ принадлежит к классу Q_k	Документ не принадлежит к классу Q_k
Документ отнесен к классу Q_k	a	b
Документ не отнесен к классу Q_k	c	d

Введем следующие понятия.

Коэффициент полноты (Recall) характеризует долю правильно отнесенных к классу Q_k документов к общему числу документов выборки, принадлежащих классу Q_k:

$$R = \frac{a}{a+c}. \tag{1.25}$$

Коэффициент точности (Precission) характеризует отношение числа документов, принадлежащих классу Q_k, к числу документов, приписанных к классу Q_k при классификации:

$$P = \frac{a}{a+b}. \tag{1.26}$$

Коэффициент шума (Noise) характеризует число документов, неверно отнесенных к классу Q_k, среди всех документов массива, отнесенных к классу Q_k при классификации:

$$N = \frac{b}{a+b}. \tag{1.27}$$

F_β – *мера* представляет собой комбинацию мер полноты и точности:

$$F_\beta = \frac{(\beta^2+1)PR}{\beta^2 P + R}, \tag{1.28}$$

где P – полнота; R – точность; β – эмпирически определяемый параметр, позволяющий задавать полноте и точности разные веса. Часто на практике коэффициенту β присваивают значение 1 и мера называется F_1. Для удобства перечисленные показатели измеряют в %, т. е. в указанных формулах появляется дополнительный сомножитель 100 %.

В ряде работ проводится аналогия между коэффициентами полноты–точности и ошибками первого и второго рода в теории статистической проверки гипотез [1].

При проведении классификации документов используют или микроточность (микрополноту), рассчитываемую только для отдельных классов, или макроточность (макрополноту), получаемую путем усреднения значений микроточности и микрополноты для всех классов выборки.

Контрольные вопросы

1. Дайте определение классификации в “широком” и “узком” смыслах?
2. Каковы модели представления текстовых документов?
3. Чем отличаются меры близости и расстояния?
4. В чем заключается обучение классификаторов на выборках фиксированного размера?
5. Каковы принципы формирования обучающих выборок?
6. Для чего используются функционалы качества?
7. Какие существуют методы оценки точности классификации?

2. МЕТОДЫ ВЫЯВЛЕНИЯ ИНФОРМАТИВНЫХ ПРИЗНАКОВ В ЗАДАЧЕ ТЕКСТОВОЙ КЛАССИФИКАЦИИ

Одной из основных проблем, возникающих при классификации текстовых документов является высокая размерность признакового пространства. Количество слов даже в небольшой выборке может составлять сотни тысяч, из-за чего многие стандартные методы классификации становятся малоэффективными. В этой связи особую важность приобретает задача выявления *информативных* (классообразующих) признаков, известная также как задача снижения размерности.

Процедура снижения размерности заключается в отборе из P исходных признаков M наиболее информативных, обладающих наилучшими разделяющими свойствами. К сожалению, вопросам снижения размерности при обработке текстовой информации в современной литературе уделено значительно меньше внимания, чем разработке самих методов классификации. Более того, в ряде авторитетных публикаций отмечается, что информативность признаков не имеет достаточно глубокого самостоятельного теоретического значения, т. к. она существенным образом зависит от используемого метода классификации, конкретной обучающей выборки, самого алгоритма выявления информативных признаков [3], и на практике снижение размерности может даже заметно ухудшить результирующую точность классификации [1].

Нельзя не согласиться, что информативность признаков – понятие относительное: один и тот же набор признаков может оказаться информативным при классификации на данной конкретной выборке и абсолютно бесполезным и неинформативным на другой. Однако необходимо отметить очевидные преимущества выбора наиболее информативной подсистемы признаков – это позволяет устранить дублирование информации, существенно уменьшить вычислительные затраты и исключить из процесса классификации те неинформативные признаки, которые никак не связаны с правилом классификации, но, в силу ограниченности выборки, получили достаточно высокий вес на стадии обучения.

В настоящее время для выявления классообразующих признаков используется два подхода. Первый подход предполагает их отбор на стадии предварительной обработки в независимости от метода, который планируется использовать для классификации. Результирующая подсистема признаков вряд ли будет “оптимальной” для всех методов классификации (например, точность метода ближайшего соседа в значительной степени зависит от эффективной фильтрации неинформативных признаков, а наивный байесовский метод толерантен к наличию малоинформативных слов).

Второй подход, получивший название *wrapper*, рассматривает отбор информативных признаков и их взвешивание как часть процедуры обучения классификатора, проводя “тонкую” настройку под специфику конкретного решающего правила. В данной работе рассматриваются универсальные методы, осуществляющие выбор информативных признаков без подстройки под какой-либо определенный алгоритм классификации.

На рис. 2.1 показано, что все термины документа могут быть разделены на три группы: *информативные, слабо информативные* и *неинформативные признаки*. Неинформативные признаки удаляются на стадии предварительной обработки текстовых документов. Для этого обычно используется словарь служебных слов, который включает местоимения, предлоги, союзы и т.д. В некоторых случаях в целях дополнительного сокращения размерности задачи проводится выделение корня слова (*stemming*) [5].

После проведения предварительной обработки задача выявления классобразующих терминов по существу сводится к разделению их на две группы, состоящие из информативных и слабоинформативных признаков.

Для выявления информативных признаков в задаче классификации текстовых документов в настоящее время наиболее эффективно используются несколько теоретических подходов:

1) взвешивание терминов;
2) переход к новой системе признаков (факторный и компонентный анализ);
3) статистический подход (χ^2-статистика);
4) теоретико-информационный подход.

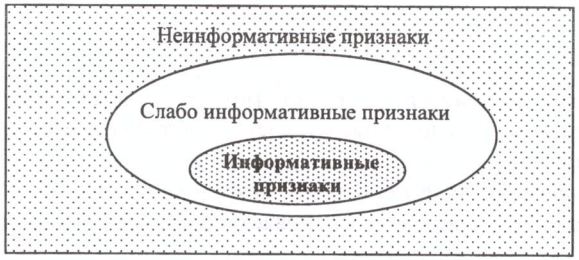

Рис. 2.1.

2.1. Методы взвешивания терминов

Данный подход основывается на предположении, что смысловая составляющая любого документа может быть представлена в виде совокупности терминов, которые с разной частотой встречаются в тексте [8]. При этом используются следующие эмпирические наблюдения (рис.2.2):

- чем чаще слово встречается в документе, тем в большей степени оно отражает тематику документа;
- чем чаще слово встречается во всей выборке документов, тем меньшей выделительной (дискриминирующей) способностью оно обладает.

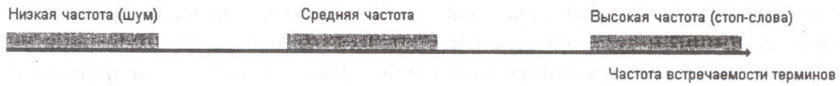

Рис. 2.2

Таким образом, для проведения классификации желательно отбирать среднечастотные термины, которые лучше всего описывают документ заданной тематики.

Формализуем процедуру взвешивания терминов следующим образом: пусть f_{ij} – частота слова i в документе j; N – число документов в выборке; M – число слов в выборке после удаления служебных слов и выделения корня слова; N_i – общее количество документов, содержащих слово i. Далее рассматриваются наиболее распространенные методы определения веса термина $x_j^{(i)}$.

1. *Логическое взвешивание.* Самый простой подход состоит в том, чтобы присвоить весу слова i значение "1", если оно встречается в документе и "0" – в противном случае:

$$x_j^{(i)} = \begin{cases} 1, & f_{ij} > 0 \\ 0, & f_{ij} = 0 \end{cases}. \qquad (2.1)$$

Основным достоинством данного метода является простота реализации, а недостатком – игнорирование частоты встречаемости термина в различных документах.

2. *Взвешивание частотой слова (term frequencies).* Другим простым подходом для определения веса $x_j^{(i)}$ является вычисление частоты слова i в документе j:

$$x_j^{(i)} = f_{ij} . \qquad (2.2)$$

В [8] показано, что использование частоты слова дает примерно 25 % увеличение эффективности классификации по сравнению с логическим взвешиванием.

3. *tf–idf - взвешивание (term frequencies – inverse document frequencies).* Предыдущие два метода не учитывают частоту встречаемости термина во всех документах выборки, его дискриминирующую способность. Для устранения этого недостатка предложено использовать так называемое *tf–idf* - взвешивание [5], присваивающее вес слову i в документе j пропорционально числу вхождений слова в данный документ и обратно пропорционально общему числу документов в выборке, в которых также содержится это слово:

$$x_j^{(i)} = f_{ij} \log\left(\frac{N}{N_i}\right). \qquad (2.3)$$

4. *tfc - взвешивание.* В *tf–idf* - взвешивании не принимается во внимание тот факт, что документы могут быть различной длины. В результате чего веса терминов в "коротких" и "длинных" документах могут существенно различаться. В *tfc* - взвешивании формула (2.3) модифицируется путем проведения нормализации длин документов [13]:

$$x_j^{(i)} = \frac{f_{ij} \log\left(\frac{N}{N_i}\right)}{\sqrt{\sum_{i=1}^{M}\left[f_{ij} \log\left(\frac{N}{N_i}\right)\right]^2}}. \qquad (2.4)$$

Суммирование в знаменателе дроби проводится по всем терминам, встречающимся в j-ом документе).

5. *ltc - взвешивание.* Данный подход заключается в использовании логарифма частоты слова вместо f_{ij}. Это позволяет сократить характерный для большинства текстовых документов существенный разброс в частотах различных терминов. Формула *ltc* - взвешивания имеет вид [13]:

$$x_j^{(i)} = \frac{\log(f_{ij}+1)\log\left(\frac{N}{N_i}\right)}{\sqrt{\sum_{i=1}^{M}\left[\log(f_{ij}+1)\log\left(\frac{N}{N_i}\right)\right]^2}}. \qquad (2.5)$$

6. *atc - взвешивание.* При таком взвешивании веса будут изменяться от 0,5 до 1, что в ряде случаев приводит к улучшению качества классификации, позволяя учесть значимые термины, имеющие редкую встречаемость в конкретной выборке:

$$x_j^{(i)} = \frac{(0{,}5+0{,}5\frac{f_{ij}}{\max f})\log\left(\frac{N}{N_i}\right)}{\sqrt{\sum_{k=1}^{Mk}\left[(0{,}5+0{,}5\frac{f_{ij}}{\max f})\log\left(\frac{N}{N_k}\right)\right]^2}}. \qquad (2.6)$$

Здесь $\max f$ – частота термина, наиболее часто появлявшегося в j-ом документе.

2.2. Факторный и компонентный анализ

Подобно тому, как формулируется задача группировки похожих по смыслу документов (строк матрицы "документ–термин") [5], которые, как предполагается, близко расположены в признаковом пространстве, может быть поставлена задача объединения терминов (столбцов матрицы "доку-

мент–термин"), которые одинаково проявляют себя на документах выборки, принадлежащих определенному классу [1,14]. Таким образом, классификация терминов заключается в группировке родственных, синонимичных, семантически связанных терминов в общий класс. В результате группа выявленных терминов будет часто встречаться в одном из классов документов и редко во всех остальных. Построенные классы терминов по существу будут отражать специфику терминологии, используемой в данной конкретной предметной области, ее внутреннюю структуру.

Методы классификации, которые при этом используются, аналогичны методам классификации текстовых документов (например, иерархический кластерный анализ, факторный и компонентный анализ, самоорганизующиеся карты Кохонена и др.) [10,15,16,6].

В данной работе более подробно рассматривается возможность использования факторного и компонентного анализа для снижения размерности в задачах классификации текстовых документов. В основу *факторного анализа (ФА)*, как метода снижения размерности, положено предположение о том, что в реальных условиях непосредственно наблюдаемые признаки обычно лишь косвенно отражают существо явления. Задача минимизации описания в факторном анализе осуществляется путем конструирования новых переменных – факторов, позволяющих в ряде случаев вскрыть логическую структуру исследуемой выборки [17,18]. *Метод главных компонент (МГК)* также предназначен для перехода к новой системе переменных, причем каждая из них определяется как линейная комбинация исходных [17].

Проведение снижения размерности с помощью ФА и МГК особенно эффективно для визуализации данных, т.е. отображения документов в трехмерное пространство или на плоскость. В ряде работ также предложено использовать близкий к факторному и компонентному анализу – *латентный семантический анализ – ЛСА (Latent Semantic Analysis)* для выявления информативных признаков и классификации текстовых документов.

В ЛСА выборка исследуемых документов представляется в виде матрицы $\mathbf{X}$ *"термин–документ"*, в которой каждый уникальный термин, встречающийся в данной выборке, задает строку матрицы, а каждый документ задает ее столбец. Несмотря на то, что данная матрица будет являться транспонированной по отношению к ранее определенной матрице "документ–термин" (см. формулу (1.4)), при дальнейшем изложении материала в данной статье знак транспонирования использоваться не будет, так как изначально выборка документов может быть описана как матрицей "документ–термин", так и матрицей "термин–документ". При этом элементы матрицы $x_{ji} = x_j^{(i)}$ (i=1,...,M; j=1,...,N) представляют собой веса, определенные, например, с использованием формул (2.1) – (2.6), отражая степень важности и значимости каждого термина в конкретном документе.

Далее в ЛСА применяется декомпозиция матрицы $\mathbf{X}$ с помощью вычисления *сингулярных (собственных) значений (SVD – Singular Value Decomposition)*. Известно, что любая прямоугольная матрица может быть разложена следующим образом:

$$\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^{\mathbf{T}}. \qquad (2.7)$$

Здесь $\mathbf{U}[M \times R]$ и $\mathbf{V}[R \times N]$ – ортогональные матрицы, в которых столбцы являются собственными векторами матрицы $\mathbf{X}$; $\Sigma[R \times R]$ – матрица, содержащая на главной диагонали неотрицательные действительные *собственные числа*; $R \le \min(M, N)$ – ранг матрицы $\mathbf{X}$.

Если собственные числа матрицы $\mathbf{X}$ расположены по убыванию, то можно перейти к редуцированным матрицам, которые будут соответствовать F наиболее значимым значениям собственных чисел:

$$\mathbf{X}_{\mathbf{F}} = \mathbf{U}_{\mathbf{F}}\Sigma_{\mathbf{F}}\mathbf{V}_{\mathbf{F}}^{\mathbf{T}} \qquad (2.8)$$

где $\Sigma_{\mathbf{F}}(F \times F)$ – матрица, полученная вычеркиванием из Σ $(R - F)$ строк и столбцов; $\mathbf{U}_{\mathbf{F}}(M \times F)$ и $\mathbf{V}_{\mathbf{F}}^{\mathbf{T}}$ $(R \times F)$ получаются путем отбрасывания собственных векторов, соответствующих незначащим собственным числам.

Таким образом, исходная матрица $\mathbf{X}$ аппроксимируется редуцированной матрицей $\mathbf{X}_{\mathbf{F}}$ (обычно в задачах текстовой классификации F =70÷300, т.е. $F << M$). При этом значение F должно быть достаточно велико для корректного отображения структуры и ассоциативных зависимостей в исходной выборке, но в то же время достаточно мало, чтобы исключить из анализа случайные и маловажные зависимости.

Тематическая близость пары терминов характеризуется тем, насколько часто они используются в документах одной и той же тематики. Для вычисления близости между всеми парами терминов достаточно рассчитать матрицу $\mathbf{X}\mathbf{X}^{\mathbf{T}}$. При этом в качестве меры близости как между документами $\vec{X}_j$ и $\vec{X}_l$, так и между терминами может использоваться *косинусоидальная метрика* (см. формулу (1.12)).

Далее в ЛСА для матрицы $\mathbf{X}\mathbf{X}^{\mathbf{T}}$ проводятся процедуры декомпозиции и аппроксимации (см. формулы (2.7) и (2.8)), после чего получаем:

$$\mathbf{X}_{\mathbf{F}}\mathbf{X}_{\mathbf{F}}^{\mathbf{T}} = \mathbf{U}_{\mathbf{F}}\Sigma_{\mathbf{F}}\mathbf{V}_{\mathbf{F}}^{\mathbf{T}}(\mathbf{U}_{\mathbf{F}}\Sigma_{\mathbf{F}}\mathbf{V}_{\mathbf{F}}^{\mathbf{T}})^{\mathbf{T}}.$$

В силу ортонормированности матрицы $\mathbf{V}_{\mathbf{F}}\mathbf{V}_{\mathbf{F}}^{\mathbf{T}} = \mathbf{I}$ ($\mathbf{I}$ – единичная матрица) окончательно получаем формулу, аналогичную методу главных компонент:

$$\mathbf{X}_{\mathbf{F}}\mathbf{X}_{\mathbf{F}}^{\mathbf{T}} = \mathbf{U}_{\mathbf{F}}^{\mathbf{T}}\Sigma_{\mathbf{F}}^{2}\mathbf{U}_{\mathbf{F}}. \qquad (2.9)$$

Применение новых F линейно независимых признаков (компонент) вместо исходных приведет к тому, что два термина (например, *классификация* и *кластеризация*), часто используемые в разных документах в близких по смыслу контекстах, будут расположены рядом друг с другом в простран-

стве F признаков, т.е. при использовании ЛСА представляется возможным учесть синонимию.

Обычно в случае использования ЛСА при поступлении в коллекцию нового документа заново проводятся процедуры декомпозиции и аппроксимации. Вместе с тем в ряде работ предложены формулы пересчета редуцированных матриц без проведения полного цикла вычислительных операций.

2.3. Статистический подход для выявления информативных признаков

Для выявления зависимости между термином $x^{(i)}$ и классом Q_k в математической статистике часто используется χ^2 - *критерий* [11]. При этом анализируется *таблица сопряженности* (табл. 2.1), в которой число строк соответствует числу признаков $(i=1,...,M)$, а число столбцов – числу классов $(k=1,...,K)$. Таким образом, клеточная частота n_{ik} характеризует частоту совместной встречаемости i-го признака и k-го класса.

Таблица 2.1

X \ Q	Q_1	Q_2	...	Q_K	$\sum_{k=1}^{K} n_{ik} = n_{i*}$
$x^{(1)}$	n_{11}	n_{12}	...	n_{1K}	n_{1*}
$x^{(2)}$	n_{21}	n_{22}	...	n_{2K}	n_{2*}
...	...	...	...	...	...
$x^{(M)}$	n_{M1}	n_{M2}	...	n_{MK}	n_{M*}
$\sum_{i=1}^{M} n_{ik} = n_{*k}$	n_{*1}	n_{*2}	...	n_{*K}	$n_{**} = N$

В табл. 2.1 использованы следующие обозначения: n_{ik} – клеточная частота, т.е. число объектов в выборке, обладающих данным сочетанием переменных $\left\{x^{(i)}, Q_k\right\}$; $n_{*k} = \sum_{i=1}^{M} n_{ik}$ – сумма клеточных частот по k-му столбцу таблицы, $n_{i*} = \sum_{k=1}^{K} n_{ik}$ – сумма клеточных частот для i-й строки таблицы.

Проверяется гипотеза о статистической независимости переменных $x^{(i)}$ и Q_j, т.е. H_0: $n_{ik} - \hat{n}_{ik} = 0$ (n_{ik} и $\hat{n}_{ik}$ – соответственно эмпирические и теоретические частоты). Классическим тестом, используемым для установления факта наличия связи, является χ^2 - критерий [19]. Величина χ^2 обычно определяется как сумма квадратов разностей между эмпирическими и теоретическими частотами двумерного распределения, деленная на $\hat{n}_{ik}$:

$$\chi^2 = \sum_{i=1}^{M}\sum_{k=1}^{K} \frac{(n_{ik} - \hat{n}_{ik})^2}{\hat{n}_{ik}}. \qquad (2.10)$$

Здесь n_{ik} – эмпирические (полученные из экспериментов) частоты; $\hat{n}_{ik}$ – теоретически ожидаемые частоты, которые рассчитаны в предположении о независимости $x^{(i)}$ и Q_k по формуле:

$$\hat{n}_{ik} = P(x^{(i)}, Q_k) = P(x^{(i)})P(Q_k) = N\frac{n_{i*}}{N}\cdot\frac{n_{*k}}{N} = \frac{n_{i*}n_{*k}}{N}.$$

Выбирая уровень значимости α, можно определить соответствующее критическое значение теста χ^2 с числом степеней свободы $S=(M-1)(K-1)$. Гипотеза о независимости отвергается с уровнем значимости α, если рассчитанная величина χ^2 превышает критическое значение $\chi^2_{\alpha,S}$.

В задачах выявления информативных признаков обычно используется частный случай χ^2-критерия, в котором анализируются отдельные фрагменты исходной таблицы, а именно *таблицы сопряженности размером 2x2* [11].

Таблица 2.2

X \ Q_k	Принадлежность классу Q_k	Непринадлежность классу Q_k	$\sum$
Наличие признака $x^{(i)}$	A	B	$A+B$
Отсутствие признака $x^{(i)}$	C	D	$C+D$
$\sum$	$A+C$	$B+D$	N

В таблице использованы следующие обозначения: A – число раз, когда $x^{(i)}$ и Q_k встречаются вместе; B – число раз, когда $x^{(i)}$ встречается без Q_k; C – число раз, когда Q_k встречается без $x^{(i)}$; D – число раз, когда ни Q_k, ни $x^{(i)}$ не встречались; N – общее количество документов в выборке.

При этом формула (2.10) перепишется в виде

$$\chi^2(x^{(i)}, Q_k) = N\frac{(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1*}n_{2*}n_{*1}n_{2*}} = N\frac{(AD-CB)^2}{(A+B)(C+D)(A+C)(B+D)}. \qquad (2.11)$$

Далее для класса Q_k вычисляется χ^2-статистика для определения значимости связи между ним и каждым признаком $x^{(i)}$ $(i=1,...,M)$. Общая величина χ^2-теста может быть определена по трем различным формулам:

$$a)\ \chi^2_{\text{сред}}(x^{(i)}) = \sum_{k=1}^{K} P(Q_k)\chi^2(x^{(i)},Q_k); \qquad (2.12)$$

$$b)\ \chi^2_{\max 1}(x^{(i)}) = \max_{k=1}^{K} P(Q_k)\chi^2(x^{(i)},Q_k); \qquad (2.13)$$

$$c)\ \chi^2_{\max 2}(x^{(i)}) = \max_{k=1}^{K} \chi^2(x^{(i)},Q_k). \qquad (2.14)$$

Вероятность $P(Q_k)$ может быть рассчитана по выборке

$P(Q_k) = \frac{N_k}{N}$ (N_k – количество элементов класса Q_k в выборке N).

Значения χ^2 – статистики, вычисленные по формулам (2.12)–(2.14), могут достаточно сильно различаться, в частности, формулы (2.12) и (2.13) существенно зависят от размера классов.

Если $x^{(i)}$ и Q_k – независимы, то χ^2 – статистика должна равняться нулю. Гипотеза отклоняется, если статистика, вычисленная по какой-либо из формул (2.12)–(2.14) больше табличного значения χ^2_α, т.е. $\chi^2(x^{(i)}) > \chi^2_\alpha$. Таким образом, для выявления информативности признаков используется предположение, что чем сильнее связаны признак $x^{(i)}$ и класс Q_k, часто появляясь совместно, тем большую важность имеет признак $x^{(i)}$ для определения класса Q_k.

Недостатком χ^2 – критерия является его вычислительная сложность (количество необходимых операций пропорционально N^2) и невысокая точность для редко встречающихся терминов, которые имеют малые частоты в таблице сопряженности признаков.

2.4. Определение информативных признаков по критерию взаимной информации

В теории информации часто возникает задача выявления количества информации, содержащейся в одном ансамбле сообщений *{X}* относительно другого зависящего от него ансамбля *{Q}*. Для этого вычисляется *взаимная информация* как *среднее количество информации*, содержащейся в *X* относительно *Q*:

$$I(X,Q) = H(X) - H(X \mid Q), \qquad (2.15)$$

где $H(X), H(X \mid Q)$ – соответственно энтропия и условная энтропия.

Используя определения энтропии и условной энтропии, формулу (2.15) можно переписать следующим образом [20]:

$$I(X,Q) = -\sum_{i=1}^{M} P(x^{(i)}) \log P(x^{(i)}) + \sum_{i=1}^{M}\sum_{k=1}^{K} P(x^{(i)},Q_k) \log \frac{P(x^{(i)},Q_k)}{P(Q_k)}. \qquad (2.16)$$

Из теории вероятности известно, что если событие $x^{(i)}$ может произойти только при выполнении одного из событий $Q_1,...,Q_K$, то вероятность $P(x^{(i)})$ вычисляется по формуле полной вероятности:

$$P(x_i) = P(Q_1)P(x^{(i)} \mid Q_1) + P(Q_2)P(x^{(i)} \mid Q_2) + ... + P(Q_K)P(x^{(i)} \mid Q_K),$$

что эквивалентно формуле

$$P(x^{(i)}) = \sum_{k=1}^{K} P(x^{(i)},Q_k). \qquad (2.17)$$

Подставляя (2.17) в (2.16) окончательно получаем:

$$I(X,Q) = -\sum_{i=1}^{M} P(x^{(i)}) \log P(x^{(i)}) + \sum_{i=1}^{M}\sum_{k=1}^{K} P(x^{(i)},Q_k) \log \frac{P(x^{(i)},Q_k)}{P(Q_k)} = \sum_{i=1}^{M}\sum_{k=1}^{K} P(x^{(i)},Q_k) \log \frac{P(x^{(i)},Q_k)}{P(Q_k)P(x^{(i)})}. \qquad (2.18)$$

В ряде публикаций по классификации текстовых документов предложено определять информативность признака путем расчета величины взаимной информации между термином $x^{(i)}$ и классом Q_k по имеющейся выборке [21]. Аналогично тому, как это делалось для χ^2 – критерия, введем обозначения: A – число раз, когда $x^{(i)}$ и Q_k встречаются вместе; B – число раз, когда $x^{(i)}$ появляется без Q_k; C – число раз, когда Q_k встречается без $x^{(i)}$; N – общее количество документов в выборке. Тогда оценки вероятностей в формуле (2.18) могут быть вычислены следующим образом:

$$P(x^{(i)},Q_k) = \frac{A}{N};\ \ P(x^{(i)}) = \frac{A+B}{N};\ \ P(Q_k) = \frac{A+C}{N}.$$

Величина взаимной информации между термином $x^{(i)}$ и классом Q_k рассчитывается по формуле

$$MI(x^{(i)},Q_k) = \log_2 \frac{AN}{(A+B)(A+C)}. \qquad (2.19)$$

Общее количество информации может быть определено одним из следующих способов:

$$1)\ I_{\text{сред}}(X,Q) = \sum_{k=1}^{K} P(Q_k) MI(x^{(i)},Q_k); \qquad (2.20)$$

$$2)\ I_{\max}(X,Q)=\max_{k=1}^{K}\{MI(x^{(i)},Q_k)\}. \qquad (2.21)$$

Количество взаимной информации вычисляется для каждого слова из обучающей выборки, и слова, чье количество взаимной информации меньше, чем некоторый устанавливаемый экспериментально порог, удаляются [21].

2.5. Использование критерия прироста информации для выявления информативных признаков

Данный критерий во многом аналогичен предыдущему и основывается на анализе количества информации, которое обеспечивает признак $x^{(i)}$ для предсказания класса Q_k.

Формула для расчета *прироста информации (IG – Information Gain)* имеет вид [13],[21]:

$$IG(x^{(i)},Q_k)=-\sum_{k=1}^{K}P(Q_k)\log P(Q_k)+P(x^{(i)})\sum_{k=1}^{K}P(Q_k\mid x^{(i)})\log P(Q_k\mid x^{(i)})+$$
$$+P(\tilde{x}^{(i)})\sum_{k=1}^{K}P(Q_k\mid \tilde{x}^{(i)})\log P(Q_k\mid \tilde{x}^{(i)}) . \qquad (2.22)$$

Здесь вероятность $P(Q_k)$ рассчитывается как число документов выборки, которые принадлежат классу Q_k; вероятности $P(x^{(i)})$ и $P(\tilde{x}^{(i)})$ рассчитываются соответственно как число документов выборки, в которых слово $x^{(i)}$ встречается хотя бы однажды или не встречается ни разу; условные вероятности $P(Q_k\mid x^{(i)})$ и $P(Q_k\mid \tilde{x}^{(i)})$ рассчитываются соответственно как число документов класса Q_k, которые имеют, по крайней мере, одно вхождение слова $x^{(i)}$, и число документов класса Q_k, которые не содержат слова $x^{(i)}$.

Количество полученной информации вычисляется для каждого слова из обучающей выборки, и слова, чье количество полученной информации меньше, чем некоторый заранее заданный порог, удаляются.

2.6. Разведочный анализ данных. Методы многомерного шкалирования

Разведочный анализ данных РАД (*Exploratory Data Analysis*) употребляется, когда априорная информация о природе данных отсутствует [1]. В этой ситуации РАД может оказать помощь в получении компактного и понятного исследователю описания структуры массива (например, в форме визуального представления), отталкиваясь от которого можно "прицельно" исследовать данные, обоснованно выбирая тот или иной метод многомерного статистического анализа, выявляя закономерности, присутствующие в данных.

Одним из наиболее эффективных инструментов РАД являются методы многомерного шкалирования (ММШ). ММШ относятся к нелинейным методам снижения размерности и визуализации данных. Как известно, линейные методы снижения размерности (компонентный и факторный анализ) достоверно передают структуру лишь в том случае, когда множество точек имеет большой разброс по одним направлениям и совсем небольшой – по другим. Когда же в данных имеются существенные нелинейности, при ортогональном проектировании сильно различающиеся точки могут накладываться друг на друга, искажая отображение исходной структуры [22].

Для методов многомерного шкалирования исходными данными служит информация о сходствах или различиях между объектами, т.е. рассматриваются матрицы мер сходства или различия (см. формулу (1.5)):

$$D=\begin{pmatrix} d_{11} & d_{12} & \dots & d_{1N} \\ d_{21} & d_{22} & \dots & d_{2N} \\ \dots & \dots & \dots & \dots \\ d_{N1} & d_{N2} & \dots & d_{NN} \end{pmatrix}.$$

В дальнейшем будем предполагать, что в качестве меры близости используется евклидово расстояние (см. формулу (1.8)).

Задача многомерного метрического шкалирования заключена в построении конфигурации точек возможно меньшей размерности q (желательно $q\leq 3$), которая порождала (объясняла) бы матрицу мер близости **D**. В случае когда $q\leq 3$, исследователь может заранее получить некоторую информацию о структуре данных, увидеть, на какие группы они распадаются [23].

Рассмотрим более подробно алгоритм метода многомерного шкалирования. В исходном M-мерном пространстве рассчитываются расстояния между парами точек d_{ij}. Необходимо найти такое расположение точек в пространстве невысокой размерности q, при котором новые расстояния d^*_{ij} (d^*_{ij} – расстояние между точками в пространстве невысокой размерности q) наилучшим образом аппроксимируют расстояния в исходном многомерном пространстве.

В зависимости от поставленной цели можно использовать тот или иной критерий отображения. Чаще всего он основан на непосредственном сравнении двух матриц близости D и D^* (содержащих соответственно расстояния d_{ij} и d^*_{ij}).

Наиболее простой и естественный критерий оптимизации имеет следующий вид [1]:

$$1)\ S_1=\min\sum_{i=1}^{N}\sum_{j=1}^{N}(d_{ij}-d^*_{ij})^2 \text{ для } \forall x^{(i)}_j . \qquad (2.23)$$

Широкое использование получили также три других критерия [24]:

$$2)\ S_2 = \frac{\sum_{l=1}^{N}\sum_{j=1}^{N}(d_{lj} - d^*_{lj})^2}{\sum_{l=1}^{N}\sum_{j=1}^{N} d_{lj}}; \qquad (2.24)$$

$$3)\ S_3 = \sum_{l=1}^{N}\sum_{j=1}^{N}(d_{lj} - d^*_{lj})^2 d_{lj}; \qquad (2.25)$$

$$4)\ S_4 = \frac{\sum_{l=1}^{N}\sum_{j=1}^{N}(d_{lj} - d^*_{lj})^2 d^a_{lj}}{\sum_{l=1}^{N}\sum_{j=1}^{N} d_{lj}},\ (l \neq j). \qquad (2.26)$$

Рассмотрим последний критерий более подробно в зависимости от параметра a. Если $a>0$, то искажение расстояния (разность $(d_{lj} - d^*_{lj})^2$ оказывает на величину критерия тем большее влияние, чем значительнее величина расстояния между точками в исходном пространстве. Поэтому при использовании критерия с параметром $a>0$ следует ожидать, что чем больше расстояние между точками в исходном пространстве, тем меньше будет искажено взаимное положение этих двух точек в пространстве невысокой размерности. По аналогичным соображениям, если $a<0$, то в меньшей степени будут искажаться взаимные положения точек, находящихся на малом расстоянии в исходном пространстве.

Более гибким является подход, основанный на использовании параметра a, значение которого меняется в зависимости от величины разности расстояний $(d_{lj} - d^*_{lj})$:

$$a = \begin{cases} -a_1, & d_{lj} \le d^*_{lj} \\ a_2, & d_{lj} > d^*_{ll} \end{cases}.$$

Рассмотрим случай, когда $a_1>0$; $a_2>0$ и пусть $d_{lj} < d^*_{lj}$ (величина d_{lj} велика). Тогда вес $\left(1/d^{a1}_{lj}\right)$ искажения расстояний $(d_{lj} - d^*_{lj})$ будет незначителен.

С другой стороны, пусть $d_{lj} > d^*_{lj}$ (величина d_{lj} велика), тогда вес искажений расстояний $(d_{lj} - d^*_{lj})^2$ будет большим $\left(d^{a2}_{lj}\right)$. Таким образом, наблюдается тенденция искажения больших расстояний в сторону увеличения, а малых – в сторону уменьшения. Естественно ожидать, что при таком характере искажений отображение исходной конфигурации в пространство невысокой размерности будет весьма удачным. Более того, при выделении кластеров подобный тип искажений может быть даже полезен, поскольку расстояния между точками, относящимися к одному кластеру, имеют тенденцию быть малыми, а расстояния между точками из разных кластеров – большими, и, следовательно, искажения рассмотренного типа в значительной степени увеличивают "контрастность" картины при визуализации.

Далее рассмотрим алгоритм минимизации функционалов S_1–S_4 на примере функционала S_4. Данная задача сводится к поиску минимума функций $(q \times N)$ – переменных с помощью итерационных схем, в частности, градиентного метода [24].

2.6.1. Градиентные методы

Градиентом функции $f(\vec{X})$ называют вектор, величина которого определяет скорость изменения функции $f(\vec{X})$, а направление совпадает с направлением наибольшего возрастания этой функции. Если $\vec{X}$ – многомерная величина (см. формулу (1.3)), то градиент представляет собой вектор–столбец вида: $\text{grad}\ f(\vec{X}) = \left(\frac{\partial f(\vec{X})}{\partial x^{(1)}}, \ldots, \frac{\partial f(\vec{X})}{\partial x^{(M)}}\right)^T$.

Условием достижения экстремума является равенство нулю градиента $\text{grad}\ f(\vec{X}) = 0$. Разложим функцию $f(\vec{X})$ в ряд Тейлора в окрестности точки $\vec{X}_0$ и ограничимся лишь линейными членами:

$$f(\vec{X}) \approx f(\vec{X}_0) + \sum_{i=1}^{M}\frac{\partial f}{\partial x^{(i)}}\Delta x^{(i)} + \frac{1}{2}\sum_{i=1}^{M}\sum_{s=1}^{M}\frac{\partial^2 f}{\partial x^{(i)}\partial x^{(s)}}\Delta x^{(i)}\Delta x^{(s)} + \ldots \approx$$

$$\approx f(\vec{X}_0) + \frac{\partial f}{\partial x^{(i)}}\Big|_{\vec{X}_0}(x^{(1)} - x_0^{(1)}) + \ldots + \frac{\partial f}{\partial x^{(M)}}\Big|_{\vec{X}_0}(x^{(M)} - x_0^{(M)}).$$

Стратегия, называемая *градиентным методом*, представляет собой последовательность следующих шагов:

1) выбор начальной точки $\vec{X}_0$ и определение направления градиента в ней;

2) осуществление одного шага движения в найденном направлении $x_1^{(i)} = x_0^{(i)} + \lambda$ (для k-ой итерации $x_k^{(i)} = x_{k-1}^{(i)} + k\lambda$, где λ – величина шага).

Шаги 1 и 2 повторяются до тех пор, пока все координаты градиента не окажутся весьма близкими к нулю, что может служить признаком того, что достигнута некоторая окрестность точки экстремума.

Градиентный метод обладает рядом недостатков: большое количество вычислений; слабая чувствительность к точкам экстремумов, находящихся в "овраге" или на "гребне" (градиентный метод может вызвать "прыжки" через овраг, так что траектория хотя и достигает точки экстремума, но ценой многих неэффективных шагов) [25]. Так как степень свободы для перемещения

точек в пространстве высокой размерности больше, чем в пространстве низкой размерности, то проблема локальных экстремумов там возникает реже.

В ряде случаев представляется целесообразным использовать модификацию градиентного метода – *метод сопряженных градиентов*, в котором направление движения из данной точки определяется с учетом значения градиента в этой точке и градиентом (направлением движения) на предыдущем шаге. При использовании метода сопряженных градиентов для минимизации функции $f(\vec{X})$ на каждой итерации все точки передвигаются в направлении:

$$p_k = -\operatorname{grad}_k f(\vec{X}) + \beta_{k-1} p_{k-1},$$

где $p_{k-1} = \dfrac{\operatorname{grad}_{k-1}}{|\operatorname{grad}_{k-1}|} \dfrac{\vec{f(X)}}{f(\vec{X})|}$; $|\operatorname{grad}_k f(\vec{X})| = \sqrt{\left(\dfrac{df}{dx^{(i)}}\right)^2 + \ldots + \left(\dfrac{df}{dx^{(i)}}\right)^2}$; p_{k-1} – значение градиента на предыдущем (k–1) шаге; $\operatorname{grad}_k f(\vec{X})$ – градиент $f(\vec{X})$ на k-ом шаге; $\beta_{k-1} = \dfrac{(\operatorname{grad}^2}{(\operatorname{grad}^2} \dfrac{f(\vec{X}))_k}{f(\vec{X}))_{k-1}}$ – отношение квадратов градиентов на k и k–1 шагах.

Наименее формализованным моментом в итерационных методах является выбор шага λ и начального приближения $\vec{X}_0$. Шаг λ выбирается на различных этапах алгоритма по-разному (он может быть переменным или постоянным) и должен обеспечивать сходимость алгоритма к оптимальному решению. Выбор шага во многом произволен, изменяя его можно получить целую совокупность возможных направлений движений. Единственное, что при этом не изменяется, – это знак каждой из компонент градиента, указывающий, в какую сторону следует проводить изменения. Удачный выбор начального приближения помогает избежать большого количества дополнительных вычислений при движении к глобальному экстремуму. Таким образом, поиск минимума функционала S_4 может проводиться с помощью градиентного метода или метода сопряженных градиентов. При этом вычисление производных осуществляется с использованием разностной схемы, получаемой при поочередном изменении одного из признаков $x^{(i)}$, т.е. расчет направления градиента состоит из p последовательных шагов расчета его координат. Итак, поиск точки минимума S_4 в $q \times N$-мерном пространстве проводится согласно итерационной процедуре:

$$d^{*(i)}_{j(k-1)} = d^{*(i)}_{j(k)} - \frac{1}{2\sum_{l=1}^{N}\sum_{j=1}^{N} d^a_{lj}} \left[\frac{\partial S_4}{\partial d^{*(i)}_j}\right]_{(k)}; l \neq j; \qquad (2.27)$$

$$\frac{\partial S_4}{\partial d^{*(i)}_j} = 2\sum_{l=1}^{N}\sum_{j=1}^{N} d^a_{lj}\left[\sum_{l=1}^{N}\sum_{j=1}^{N}\left(1 - \frac{d_{lj}}{d^*_{lj}}\right)(d^{*(i)}_l - d^{*(i)}_j)\right], \qquad (2.28)$$

где $i = 1,\ldots,q$; k – номер шага итерации; $d^{*(i)}_j$ – i-ая координата образа j-го объекта в q-мерном пространстве; $\dfrac{1}{2\sum_{l=1}^{N}\sum_{j=1}^{N} d^a_{lj}}$ – шаг градиентного метода (при a=0 шаг $\lambda = \dfrac{1}{2}N$). Доказано [1], что данный итерационный процесс сходится вне зависимости от задания начальных условий.

В результате нелинейного шкалирования центр тяжести новой конфигурации, полученной в результате отображения, совпадает с центром тяжести первоначальной конфигурации в пространстве большей размерности.

Оценка размерности конфигурации может определяться несколькими способами.

1. Строится центрированная матрица скалярных произведений между объектами, находятся ее собственные значения и упорядочиваются в порядке убывания. Проверяется, сколько необходимо взять собственных векторов, чтобы соответствующие им собственные значения объясняли большую часть дисперсии (75–95%). Это число принимается за размерность признакового пространства.

2. Подбор адекватной размерности можно осуществить, постепенно добавляя по одной оси и анализируя полученные представления. После того как минимально допустимая размерность будет достигнута, добавление каждой новой оси, соответствующей новому фактору, будет лишь незначительно уточнять предыдущее решение. При проведении визуализации данных выбор, по существу, должен проводиться между q=2 и q=3. В тех практических ситуациях, когда существуют факторы, по которым объекты различаются наиболее сильно, можно дать интерпретацию полученным осям (оси будут соответствовать выявленным факторам).

2.7. Сеть Кохонена

Обычно *самоорганизующиеся карты Кохонена (Kohonen's Self Organizing Maps)* или *сеть Кохонена (Kohonen's Neural Network)* рассматриваются в курсе по системам искусственного интеллекта в качестве одной из существующих нейросетевых парадигм. Однако, на наш взгляд, включение этого раздела в данное учебное пособие, содержащее в основном методы многомерного статистического анализа, закономерно и логично. Сам автор первоначально относил самоорганизующиеся карты к методам кластеризации, подробно рассматриваемым в главе 3 [6]. Кроме того, сеть Кохонена является мощным инструментом визуализации данных большой размерности на плоскости, что особенно ценно при анализе массивов текстовой информации.

Самоорганизующиеся карты Кохонена всегда состоят из двух слоев: первый – входной слой, содержащий нейроны для каждого признака входного вектора (обозначены черными заштрихованными кружочками на рис. 2.3), второй – выходной или решетка из m нейронов (на рис. 2.4. представлены прямоугольная и шестиугольная решетки), связанных со всеми нейронами входного слоя. Именно настройка синоптических весов нейронов выходного слоя является основной задачей алгоритма. При этом к важному отличию сети Кохонена от многослойной нейросети (НС) относится то, что используется метод обучения без учителя [6,7].

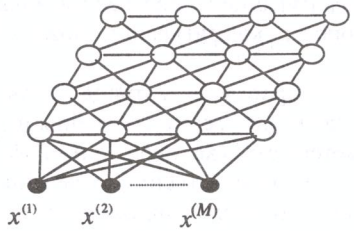

Рис. 2.3

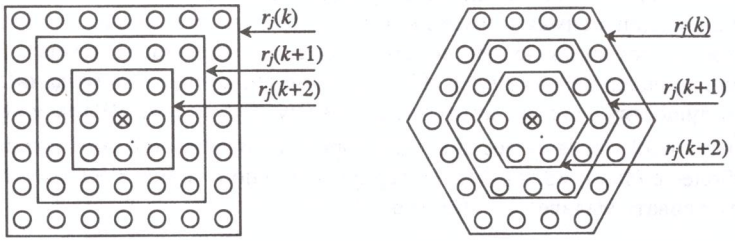

Рис 2.4

В сети Кохонена предполагается наличие взаимодействия между нейронами, сила которого определяется расстоянием между ними. На рис. 2.4 зона соседства $r_j(k)$ объединяет множество нейронов, которые считаются соседями j-го нейрона в момент времени k процесса обучения. Зоны соседства уменьшаются с течением времени. Количество нейронов в сетке определяет степень детализации результата работы алгоритма, и, в конечном счете, от этого зависит обобщающая способность сети Кохонена.

Алгоритм настройки весов в сети Кохонена.

1. Инициализация весовых коэффициентов нейронов может проводиться следующими способами:

- инициализация случайными значениями, когда всем весам присваиваются малые случайные величины;

- инициализация примерами, когда в качестве начальных весов задаются значения признаков случайно выбранных примеров из обучающей выборки;

- все веса принимаются равными одной и той же величине $\frac{1}{\sqrt{M}}$, где M – число координат входного вектора. Общее число настраиваемых синоптических весов равно Mm. Начальная зона соседства – $r_j(k)$.

2. Предъявление сети нового входного объекта $\vec{X}_{N+1}$, который выбирается случайно из обучающего множества.

3. Вычисление расстояния до всех нейронов решетки. Расстояние d_j от входного объекта до каждого нейрона j определяется по формуле

$$d_j = \sum_{i=0}^{M} (x^{(i)} - W_j^{(i)}(k))^2, \qquad (2.29)$$

где $x^{(i)}$ – i-ый признак $\vec{X}_{N+1}$; $W_j^{(i)}(k)$ – вес связи от i-го входного сигнала к j-му нейрону при обучении в момент времени k.

4. Выбор нейрона с наименьшим расстоянием, т.е. выбирается нейрон j^* ($j^*=1,\ldots, m$), для которого расстояние d_j наименьшее.

5. Настройка весов нейрона j^* и его соседей. Производится подстройка весов для нейрона j^* и всех нейронов из его зоны соседства r_j по формуле

$$W_j^{(i)}(k+1) = W_j^{(i)}(k) + r_j(k)(x^{(i)} - W_j^{(i)}(k)), \qquad (2.30)$$

где $r_j(k)$ – зона соседства в момент времени k, уменьшающаяся в процессе обучения (это позволяет делать большие шаги для быстрого грубого обучения и меньшие шаги при подходе к окончательной величине).

6. Возвращение к шагу 2.

Таким образом, в результате применения сети Кохонена происходит отображение большого числа входных объектов в карту кластеров меньшей размерности, при этом близким точкам на карте отвечают близкие друг к другу входные объекты в исходном пространстве. Время, необходимое для обучения сети Кохонена, во много раз меньше, чем обучение алгоритма обратного распространения для многослойных нейросетей, однако при этом достигается худшая точность.

Контрольные вопросы

1. Проведите сравнение методов взвешивания терминов.
2. Как можно использовать латентный семантический анализ для выявления информативных признаков и классификации текстовых документов?
3. В чем заключается использование χ^2 – критерия для выявления информативных признаков?
4. В чем заключается использование теоретико-информационного подхода для выявления информативных признаков?
5. Проанализируйте точность визуализации текстовой информации с помощью метода многомерного шкалирования и сети Кохонена.

3. МЕТОДЫ КЛАССИФИКАЦИИ ТЕКСТОВЫХ ДОКУМЕНТОВ

Классификация объектов – область науки, имеющая древнюю историю, ей посвящено огромное количество публикаций. Одним из первых, кто постарался разделить окружающие объекты на категории ("роды", "виды"), был Аристотель. Демокрит в "Письме ученому соседу" писал: "Если тебе, дорогой друг, нужно разобраться в сложном нагромождении фактов или вещей, ты сначала разложи их на небольшое число куч по похожести. Картина прояснится, и ты поймешь природу этих вещей" [2].

Основные положения теории классификации были заложены и развиты в течение XX века, при этом можно выделить два основных этапа становления данного научного направления. Первый этап охватывает 60-годы прошлого века, когда большая часть исходных идей, методов и моделей теории классификации развивалась в рамках работ по распознаванию образов, машинному обучению, многомерному статистическому анализу, нейросетям. Несмотря на весьма бурное и успешное развитие, классификация на этом этапе представляла из себя область, обладающую яркими идеями нескольких искусных людей, некий "мешок с фокусами", а не единую теорию, которая предлагала бы надежный инструментарий для решения практических задач.

Второй этап в развитии теории классификации был инициирован компьютерным и информационным бумом. Быстродействующие компьютеры, сеть Интернет породили новую парадигму – информация стала общедоступна, и на первый план в области информатики вышли проблемы ее поиска, обработки и анализа. Для решения этих проблем стали успешно применяться методы теории классификации как разработанные на первом этапе ее развития, так и новые подходы и модификации известных методов, появившиеся совсем недавно.

Прежде чем приступить к детальному рассмотрению алгоритмов классификации текстовых документов, введем несколько правил, которые желательно соблюдать при их разработке.

1. Результаты классификации не должны зависеть от порядка обработки наблюдений – любая перестановка наблюдений не должна изменять получаемые результаты.
2. Классификации необходимо быть устойчивой к шуму, т.е. нерелевантные шумовые наблюдения, которые не могут быть отнесены ни к одному из представленных в выборке классов, вызывают лишь незначительные изменения в точности классификации.
3. Классификация должна быть независимой от масштаба. Это означает, что умножение на константу значений признаков, идентифицирующих наблюдение, не должно влиять на результаты классификации.

Методы классификации принято различать по следующим характеристикам:

- по количеству используемой априорной информации (параметрические и непараметрические);

- по структуре организации классов (иерархические и неиерархические);
- по способу построения классов (пересекающиеся и непересекающиеся);
- по наличию или отсутствию обучения.

Таким образом, главная цель классификации – нахождение групп похожих объектов в анализируемой выборке данных. К числу основных характеристик таких групп необходимо отнести *плотность* (скопление точек в пространстве признаков относительно плотное по сравнению с другими областями этого пространства), *дисперсию* (степень рассеяния точек в пространстве признаков относительно центра класса), *размер* (количество объектов в группе), *структуру* (расположение точек в пространстве). Итак, под классом будем понимать "густонаселенные" области признакового пространства, отделенные от других таких же областей "разряженными" участками с относительно низкой плотностью точек.

Постановка задачи классификации приведена на рис. 3.1. Далее рассматриваются непараметрические методы классификации (и кластеризации) текстовых файлов, получивших, на наш взгляд, наибольшее применение для решения практических задач.

Рис. 3.1

3.1. Центроидные методы

Одними из первых в теории классификации были разработаны *центроидные методы*, в которых по выборке для каждого класса k (k=1,...,K) вычисляется центроид $\vec{\mathbf{C}}_k$ – вектор со средними значениями весов терминов документов данного класса:

$$\vec{\mathbf{C}}_k = \frac{1}{N_k}\sum_{j=1}^{N_k}\vec{X}_j\ , \qquad (3.1)$$

где N_k – количество документов, принадлежащих классу Q_k. Для классификации нового документа $\vec{X}_{N+1}$ определяется расстояние, например по формуле (1.8), между ним и центроидами всех классов; $\vec{X}_{N+1}$ относится к классу с наиболее близким центроидом.

В *методе Роккио* для каждого центроида проводится увеличение значений тех признаков, которые наиболее часто встречаются в данном классе, и уменьшение значений тех признаков, которые относятся к другим классам. Центроидный вектор $\vec{\mathbf{C}}_k$ для класса Q_k вычисляется по формуле [5]:

$$\vec{C}_k = \alpha\frac{1}{N_k}\sum_{j=1}^{N_k}\cos(\vec{C}_k,\vec{X}_j) - \frac{\beta}{N-N_k}\sum_{l=1}^{N-N_k}\cos(\vec{C}_k,\vec{X}_l)\,, \qquad (3.2)$$

где α и β – настраиваемые коэффициенты (обычно $\alpha > \beta$).

3.2. Метод потенциальных функций

В *методе потенциальных функций* для классификации $\vec{X}_{N+1}$ определяются "относительные потенциалы" (аналогично тому, как это делается в электротехнике), которые наводятся в новой точке признакового пространства объектами, уже распределенными по классам, и $\vec{X}_{N+1}$ относится к классу, чей наведенный совокупный потенциал выше [26].

Относительный потенциал в $\vec{X}_{N+1}$, который создается объектами k-го класса, рассчитывается по формуле:

$$\Phi_k(\vec{X}_{N+1}) = \sum_{j=1}^{N_k}\varphi(\rho(\vec{X}_{N+1},\vec{X}_j))/N_k\,. \qquad (3.3)$$

Здесь $\varphi(\rho)$ – некоторая известная положительная функция от метрики расстояния ρ, стремящаяся к 0, при $\rho\to\infty$. Обычно $\varphi(\rho) = e^{-\alpha\rho^{\beta}}$ или $\varphi(\rho) = (1+\alpha\rho^{\beta})^{-1}$, $\alpha > 0$, $\beta > 0$ [26],[27].

Таким образом, согласно методу потенциальных функций новый объект $\vec{X}_{N+1}$ относится к тому классу, который имеет наибольший относительный потенциал:

$\Phi_k(\vec{X}_{N+1}) > \Phi_g(\vec{X}_{N+1}) \Rightarrow$ новый объект относится к k-му классу (g,k=1,...,K, $g\neq k$).

3.3. Наивный байесовский классификатор

В *наивном байесовском классификаторе* (*Naive Bayes Classificator*) используется вероятностная модель определения класса документов (см. формулу (1.6)), и оценка $\hat{P}(Q_k)$ для $P(Q_k)$ рассчитывается по документам обучающей выборки [13]: $\hat{P}(Q_k) = \frac{N_k}{N}$, (3.4)

где N_k – число документов обучающей выборки, принадлежащих классу Q_k.

Оценка $\hat{P}(x^{(i)} \mid Q_k)$ для $P(x^{(i)} \mid Q_k)$ также может быть рассчитана по документам обучающей выборки:

$$\hat{P}(x^{(i)} \mid Q_k) = \frac{N_{ik}}{N_k}, \tag{3.5}$$

где N_{ik} – частота встречаемости слова i в документах класса Q_k в обучающей выборке; N_k – общее количество терминов в документах класса Q_k. Более часто используется уточненная формула:

$$\hat{P}(x^{(i)} \mid Q_k) = \frac{1 + N_{ik}}{M + N_k}, \tag{3.6}$$

где M – общее количество терминов во всех документах выборки.

3.4. Метод деревьев решений

В методе *деревьев решений* [28] проводится последовательное разделение множества документов на основе значений выбранного признака $x^{(i)}$, в результате чего строится дерево, содержащее *нетерминальные узлы* (узлы проверок), в которых происходит разбиение по выбранному атрибуту, и *терминальные узлы* (узлы ответа), в которых должны находиться элементы одного класса.

Популярность деревьев решений обуславливается быстротой их построения, наглядностью представления результатов и возможностью преобразования в наборы символьных правил, описывающих каждый из классов.

Для выбора наиболее информативного признака, по которому проводится разбиение, в методе деревьев решений чаще всего используется *теоретико-информационный (энтропийный) подход*.

Согласно теоретико-информационному подходу среднее количество информации (энтропия), необходимое для определения класса примера из обучающей выборки T, может быть определено по формуле

$$I(T) = \sum_{k=1}^{K} P_k \log_2 \frac{1}{P_k} = -\sum_{k=1}^{K} \frac{N_k}{N} \log_2 \frac{N_k}{N}, \tag{3.7}$$

где P_k – вероятность того, что случайно выбранный пример из множества T принадлежит k-му классу; N_k – количество примеров, содержащихся в k-м классе; N – размер обучающей выборки T; K – количество классов.

После разбиения множества T по i-му признаку $x^{(i)}$, который может принимать S значений, среднее количество информации, необходимое для идентификации класса примера в каждом подмножестве, определяется следующим образом [29]:

$$I(x^{(s)}, T) = \sum_{i=1}^{S} \frac{N_s}{N} I(T_s) = \sum_{s=1}^{S} \frac{N_s}{N} \left(-\sum_{k=1}^{K} \frac{N_{ks}}{N_s} \log_2 \frac{N_{ks}}{N_s} \right), \tag{3.8}$$

где S – количество узлов; N_s – количество элементов в s-м узле после разбиения исходного множества по признаку $x^{(i)}$; N_{ks} – количество элементов s-го узла, которые соответствуют k-му классу.

Величина прироста информации (information gain) вычисляется по формуле $Gain(x^{(s)}, T) = I(T) - I(x^{(s)}, T)$ и характеризует информативность признака $x^{(i)}$. В алгоритме деревьев решений важно выбрать такой признак, чтобы при разбиении по нему один из классов имел наибольшую вероятность появления, т.е. формировался терминальный узел. Это возможно в том случае, если $I(x^{(s)}, T)$ будет иметь минимальное значение и, соответственно, значение $Gain(x^{(s)}, T)$ достигнет своего максимума.

По существу деревья решений являются одним из методов, в которых проводится *вывод логических закономерностей (индукция правил – rule induction)*. Другим таким подходом, популярным при решении задач текстовой классификации, является *метод RIPPER*. Он состоит из двух этапов. На первом этапе по обучающему множеству документов для каждого класса строится набор правил – на шаге q к правилу r_q добавляется новое условие согласно модифицированной формуле прироста информации [30]:

$$Gain(r_{q+1}, r_q) = T_{q+1}^{+} \left(-\log_2 \frac{T_q^{+}}{T_q^{+} + T_q^{-}} + \log_2 \frac{T_{q+1}^{+}}{T_{q+1}^{+} + T_{q+1}^{-}} \right), \tag{3.9}$$

где $T_q^{+}(T_q^{-})$ – количество документов в обучающем множестве, удовлетворяющее (не удовлетворяющее) выявленному правилу. На втором этапе проводится минимизация выявленных правил (pruning).

3.5. Коллективная классификация

Для улучшения точности классификации часто проводят объединение различных классификаторов в единый коллектив (ансамбль) (например, методы *bagging* и *boosting*), в котором решение об отнесении наблюдения к тому или иному классу принимается путем голосования [13],[31]. Так как ошибки каждого классификатора в ансамбле независимы, то при увеличении числа методов, используемых для коллективного распознавания, общая ошибка будет стремиться к нулю.

При объединении классификаторов методом *bagging* каждый из них проходит обучение на примерах, случайным образом изъятых из одной и той же выборки, т.е. какие-то примеры могут использоваться для обучения нескольких классификаторов, а какие-то не использоваться вообще. На рис. 3.2 схематично представлен процесс коллективной классификации.

При объединении классификаторов методом *boosting* обучающая выборка формируется специальным образом. Цель заключается в том, чтобы каждый последующий классификатор "исправлял" ошибки предыдущего и точно группировал те примеры, которые ранее были ошибочно отнесены к неправильному классу. Поэтому примерам, которые были ошибочно распознаны, присваивается больший вес с тем, чтобы они чаще попадали в обучающую выборку для следующего классификатора (вероятность такого попадания тем выше, чем чаще пример был неправильно распознан предшествующими классификаторами).

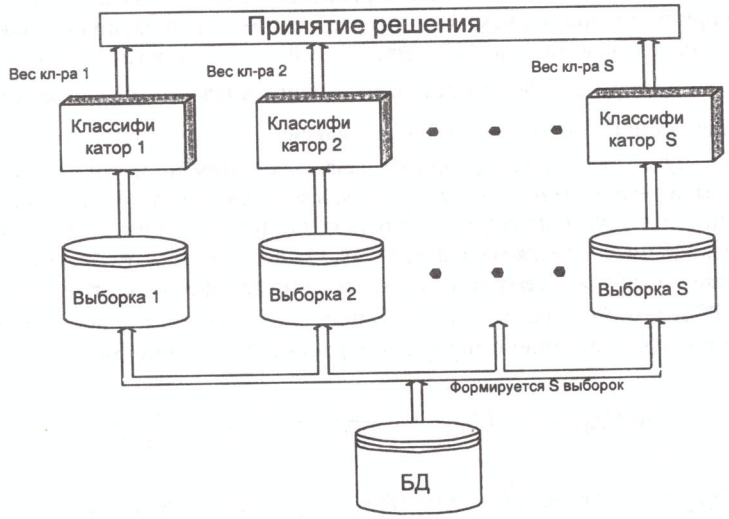

Рис. 3.2

Обычно предполагается, что при коллективной классификации используются классификаторы, имеющие различный "принцип действия". Каждому из таких классификаторов в зависимости от демонстрируемой точности может присваиваться весовой коэффициент, который отражает степень участия конкретного метода в коллективной выработке окончательного решения о принадлежности нового наблюдения тому или иному классу. Исследования показали, что *boosting* обычно более эффективен, чем *bagging*, однако является особо чувствительным к шумам (именно шумовые примеры чаще всего попадают в обучающую выборку).

3.6. Обзор методов текстовой классификации

Метод опорных векторов (*Support Vector Machines*) строит по обучающей выборке гиперплоскость, разделяющую классы [3]. В простейшем

случае двух классов задача заключается в определении такого вектора $\vec{\varphi}$ и такого числа C, чтобы для векторов $\vec{X}_l$ и $\vec{X}_j$, относящихся к разным классам, были справедливы неравенства:

$$(\vec{X}_l, \vec{\varphi}) > C, \quad \vec{X}_l \in Q_1$$
$$(\vec{X}_j, \vec{\varphi}) < C, \quad \vec{X}_j \in Q_2 \qquad (3.10)$$

Построение гиперплоскости может быть сведено к оптимизационной задаче: необходимо найти вектор $\vec{\varphi}$, который доставляет максимум специальной функции: $Margin = C_1(\vec{\varphi}) - C_2(\vec{\varphi})$, т.е. расстояние (*Margin* – зазор) между проекциями множеств Q_1 и Q_2 на направление нормали $\vec{\varphi}$ должно быть максимально (рис.3.3).

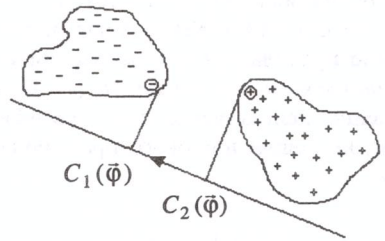

Рис. 3.3

Особенность метода заключается в том, что с его помощью строится оптимальная разделяющая гиперплоскость, которая полностью определяется крайними векторами, находящимися в граничной области между классами и получившими *название опорных векторов – support vectors* (на рис.3.3 опорные вектора взяты в кружок).

Еще один эффективный подход к классификации реализован в *регрессионном* методе *(Linear Least Squares Fit)* [21]. В нем обучающая выборка представляется в виде *"входных–выходных"* данных, где в качестве входных используются текстовые документы, задаваемые своими признаками, а в качестве выходных – вектор классов (см. рис. 3.4).

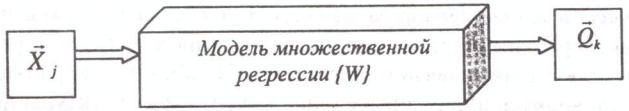

Рис. 3.4

Задача заключается в нахождении матрицы коэффициентов *W*, которые позволяют корректно проводить преобразование входных векторов в выходные. Для определения этих коэффициентов обычно используется метод наименьших квадратов.

В настоящее время для решения задачи классификации текстовых документов активно используются *нейросети (НС)*: *многослойные персептроны (МП), RBF-сети, вероятностные (байесовские) сети, самоорганизующиеся карты Кохонена* [6],[7]. НС отличаются друг от друга размерами и структурой, методами отбора данных для анализа, временем и алгоритмами обучения, точностью, способностью к обобщению и т.п. Как представляется, их использование наиболее оправдано, если выборка состоит из линейно неразделимых классов и "классические" методы оказываются мало эффективными. Применение различных нейросетевых подходов для решения задачи текстовой классификации выходит за рамки данного учебного пособия и является темой отдельного исследования.

Необходимо также отметить ряд других подходов, которые достаточно успешно используются для классификации текстовых документов: *методы графов* [8], *линейный дискриминантный анализ* [13], *классификаторы с использованием нечеткой логики* [22],[32], методы, использующие лингвистические и семантические правила для формализации содержания текстовых документов [5], *автоматические рубрикаторы* [7]. В то же время *параметрические методы*, успешно используемые для классификации фактографической информации [33], не нашли широкого применения для группирования текстовых документов.

3.7. Иерархические методы для классификации текстовой информации

Иерархические процедуры по сравнению с другими кластерными процедурами дают более полный и тонкий анализ структуры исследуемого множества наблюдений. Привлекательной стороной подобных алгоритмов является возможность наглядной интерпретации проведенного анализа. К недостаткам иерархических процедур следует отнести громоздкость их вычислительной реализации. Соответствующие алгоритмы на каждом шаге требуют вычисления всей матрицы расстояний, поэтому реализация таких алгоритмов при значительном числе наблюдений крайне затруднительна и нецелесообразна. Эффективность иерархической классификации существенно зависит от внутренней структуры исследуемого множества объектов, от пороговых значений и выбранной меры близости между классами [16],[32].

Иерархической классификацией данного множества объектов называется построение иерархии, отражающей наличие однородных в определенном смысле классов и взаимосвязи между ними. На практике иерархическая кластеризация проводится или для выявления внутренней структуры данных, или для сокращения количества кластеров до заданного.

Существует два подхода, реализующих иерархическую кластеризацию, – агломеративный и дивизимный. Под *агломеративным* подходом подразумевается результат процедуры последовательного формирования сначала кластеров, состоящих из групп наиболее тесно связанных объектов, а затем кластеров следующего уровня путем присоединения "тяготеющих" к ним объектов. Такая процедура при движении от вершин к корню итерируется,

пока кластеризация не охватит всё множество объектов. При *дивизимной* (делящей) иерархической процедуре на 1-м уровне иерархии находится один класс, включающий все объекты (корень дерева), а на наивысшем уровне число классов равно числу классифицируемых объектов. Агломеративная и дивизимная процедуры соответственно представлены на рис. 3.5 а и б.

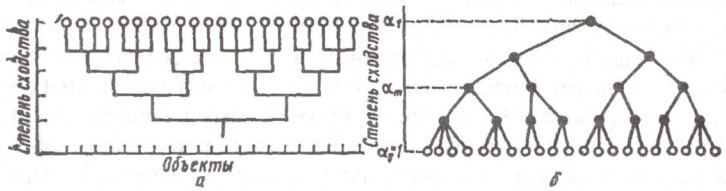

Рис. 3.5

Иерархическая структура объектов геометрически представляется в виде *дендрограммы* (иерархия с численными уровнями) либо в виде дерева классов. Такое представление результата работы алгоритма наглядно иллюстрирует, как *N* объектов распределились по группам, находящимся на последовательных уровнях иерархии (каждый уровень соответствует числовому порогу, указывающему степень сходства объектов в группах этого уровня).

После выполнения очередного шага агломеративной процедуры проверяется, достигнуто ли желательное разбиение. Существуют различные методы определения критерия остановки процедуры:

- получено определенное заранее количество кластеров;
- все кластеры содержат более определенного числа элементов;
- кластеры обладают требуемым соотношением внутренней однородности и разнородности между собой.

При объединении кластеров в иерархических процедурах можно использовать различные меры близости между ними [34].

- *Одиночная связь (метод ближайшего соседа).* В этом методе расстояние между двумя кластерами определяется расстоянием между двумя наиболее близкими объектами (ближайшими соседями) в различных кластерах.
- *Полная связь (метод наиболее удаленных соседей).* В этом методе расстояния между кластерами определяются наибольшим расстоянием между двумя объектами (наиболее удаленными соседями) в различных кластерах.
- *Невзвешенное попарное среднее.* В этом методе расстояние между двумя различными кластерами вычисляется как среднее расстояние между всеми парами объектов в них.
- *Взвешенное попарное среднее.* Метод идентичен методу невзвешенного попарного среднего за исключением того, что вычисляется размер соответствующих кластеров (т.е. число объектов, содержащихся в них) и эта величина используется в качестве весового коэффициента. Предлагаемый

метод обычно используется, когда предполагаются неравные размеры кластеров.

- *Невзвешенный центроидный метод.* В этом методе расстояние между двумя кластерами определяется как расстояние между их центрами тяжести.
- *Взвешенный центроидный метод.* Данный метод идентичен предыдущему за исключением того, что при вычислениях используются веса для учёта разницы между размерами кластеров (т.е. количеством объектов в них).
- *Метод Варда.* Этот метод отличается от всех других методов, поскольку он использует методы дисперсионного анализа для оценки расстояний между кластерами. Метод минимизирует сумму квадратов расстояний для любых двух (гипотетических) кластеров, которые могут быть сформированы на каждом шаге. В целом метод является весьма эффективным, однако он стремится создавать кластеры малого размера.

3.8. Правило ближайшего соседа

Идея метода ближайшего соседа заключается в том, чтобы принимать решение об отнесении нового наблюдения $\vec{X}_{N+1}$ к тому или иному классу не по всей выборке, а лишь по наблюдению, которое находится в непосредственной близости от классифицируемого объекта.

Сформулируем *правило ближайшего соседа (БС)*. Пусть к моменту появления нового наблюдения $\vec{X}_{N+1}$ все предыдущие N наблюдений уже распределены по K классам, т.е. имеется множество пар $\{\vec{X}_j, Q_k\}$, в которых $\vec{X}_j$ является вектором в M-мерном признаковом пространстве с метрикой d и принадлежит одному из k классов Q_k (j=1,…,N, k=1,…,K). После появления нового наблюдения $\vec{X}_{N+1}$ возникает вопрос, к какому классу его отнести? Будем называть $\vec{X}_j^*$ ближайшим соседом к новому наблюдению $\vec{X}_{N+1}$, если расстояние между ними наименьшее [35]:

$$d\left(\vec{X}_j^*, \vec{X}_{N+1}\right) = \min d\left(\vec{X}_j, \vec{X}_{N+1}\right), \text{ для } \forall\, j=1,\ldots,N. \qquad (3.11)$$

Согласно правилу ближайшего соседа новое наблюдение $\vec{X}_{N+1}$ относится к тому же классу Q_k (k=1,…,K), к которому принадлежит его ближайший сосед $\vec{X}_j^*$. Таким образом, МБС относится к *"локальным"* методам классификации, принимающим решение на основе свидетельства лишь того наблюдения, которое является наиболее близким к классифицируемому.

МБС находит широкое практическое применение благодаря тому, что его реализация на ЭВМ очень проста, результаты не зависят от порядка следования наблюдений и легко интерпретируемы. Как справедливо отмечается в [3], "исследуя явления природы, неразумно искать объясняющий их закон в классе сложных функций, это и бесполезно, потому что не хватит экспериментального материала. Принятие гипотезы о "простом мире" позволяет, отбросив подавляющую часть всех функций – сложные функции, искать решения в сравнительно малочисленном классе простых функций. Не это ли имел в виду Эйнштейн, когда заметил, что бог изощрен, но не злонамерен".

Несмотря на то, что метод ближайшего соседа является эвристическим, для него существуют теоретически рассчитанные точностные характеристики: *асимптотическая вероятность ошибки* ($R_{БС}$) для МБС не превышает вероятность ошибки правила Байеса (R^*), наименьшей из возможных, более чем в два раза ($R^* < R_{БС} < 2R^*$) [4].

Однако необходимо отметить, что МБС имеет ряд значительных недостатков. Так, при принятии решения по правилу ближайшего соседа по существу игнорируются остальные (N–1) наблюдений, в МБС может также происходить значительное ухудшение качества классификации в случае наличия в выборке шумовых (нерелевантных) наблюдений, которые не принадлежат ни одному из классов Q_k (k=1,…,K), и использования неинформативных признаков $x_j^{(i)}$. Кроме того, возникает вычислительная проблема, т.к. приходится рассчитывать все расстояния между новым наблюдением $\vec{X}_{N+1}$ и N уже имеющимися наблюдениями, что существенно ограничивает скорость классификации.

Для решения данных проблем предложено целое семейство методов, улучшающих характеристики метода ближайшего соседа (рис. 3.6) с целью компенсации какого-либо из его недостатков. Метод *к-ближайших соседей* (*к-БС*) расширяет число представителей выборки, участвующих в принятии решения о классификации объекта, *адаптивные методы* позволяют уменьшать негативное влияние неинформативных признаков на точность МБС, *модифицированные методы* проводят сортировку элементов исходной выборки с целью сокращения вычислительных операций без увеличения ошибки распознавания, *редуцированные методы* удаляют из выборки шумовые нехарактерные наблюдения, увеличивая точность и быстродействие. Однако необходимо отметить, что во всех вариациях МБС принятие решения о классификации нового наблюдения по-прежнему осуществляется на основе правила ближайшего соседа (или правила *к-ближайших соседей*, которое будет рассмотрено ниже).

Рис. 3.6

3.9. Метод *к*-ближайших соседей

В *методе к-ближайших соседей* аналогично МБС для классификации нового наблюдения $\vec{X}_{N+1}$ проводится упорядочивание исходных элементов выборки по какой-либо метрике (например, евклидову расстоянию). При этом определяется не один ближайший сосед, а группа соседей, наиболее близких к новому наблюдению. Число соседей *к* является настраиваемым на стадии обучения (или задаваемым экспертно) параметром метода. Решение об отнесении $\vec{X}_{N+1}$ к классу Q_k (k=1,…,K) принимается путем голосования его *к*-ближайших соседей с помощью простого подсчета голосов. Если более половины *к*-БС принадлежат классу Q_k, то $\vec{X}_{N+1}$ также относится к этому классу. Таким образом, в методе *к*-БС устраняется один из недостатков МБС, так как решение принимается на основе голосования не одного, а нескольких элементов исходной выборки.

Во *взвешенном методе к-ближайших соседей (Вк-БС)* те соседи, которые являются наиболее близкими к новому наблюдению имеют больший вес при голосовании. Если $\vec{X}_{N+1}$– новое наблюдение, которое имеет *к*-ближайших соседей j=1,…, *к*, пронумерованных от самого близкого до самого дальнего в соответствии с расстоянием d_j, то вес каждого соседа может быть рассчитан по формуле [35]

$$\omega_j = \begin{cases} \dfrac{d_k - d_j}{d_k - d_1} & ,d_j \neq d_1 \\ 1 & ,d_j = d_1. \end{cases} \qquad (3.12)$$

Такая весовая функция изменяется от максимума, равного единице, который соответствует ближайшему соседу, до минимума, равного нулю, который соответствует наиболее отдаленному *к*-му соседу (рис. 3.7,б). Новое наблюдение $\vec{X}_{N+1}$ относится к тому классу, который набирает наибольший вес при голосовании *к*-БС.

В ряде случаев представляется целесообразным использовать *гауссово взвешивание* $\omega_j = \omega_k^{d_j^2/d_k^2}$ или *экспоненциальное взвешивание* $\omega_j = \omega_k^{d_j/d_k}$, где ω_k – вес последнего *к*-го соседа (см. соответственно рис. 3.7,в и 3.7,г). Выбор ω_k для большинства задач представляется трудно формализуемой проблемой, требующей отдельных исследований и ограничивающей возможности применения гауссова и экспоненциального взвешивания.

Если количество примеров N велико по сравнению с количеством *к*-БС и они расположены компактно в "густонаселенных" областях признакового пространства, то качество классификации простым и каким-либо из взвешенных методов *к*-БС приблизительно одинаковы. Однако метод Вк-БС часто бывает более точен в случае малой выборки и неравномерного распределения по классам.

Согласно формуле (3.12) *к*-ый сосед имеет нулевой вес и не участвует в голосовании. Представляется целесообразным изменить формулу (3.12) и расчет весов производить следующим образом:

$$\omega_j = \begin{cases} \dfrac{(d_k - d_j) + \rho(d_k - d_1)}{(1+\rho)(d_k - d_1)} & ,d_j \neq d_1 \\ 1 & ,d_j = d_1. \end{cases} \qquad (3.13)$$

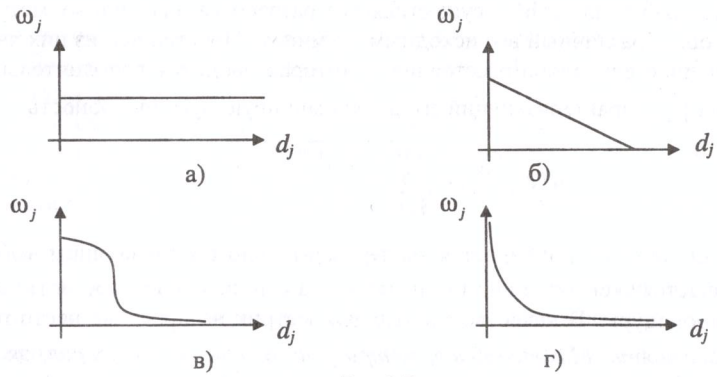

Рис. 3.7

При этом *к*-ый сосед имеет вес $\omega_k = \dfrac{\rho}{(1+\rho)}$. Варьирование коэффициента ρ позволяет корректировать веса различных соседей, по существу он аналогичен настраиваемому параметру ω_k в формулах гауссова и экспоненциального взвешиваний. Значение ρ можно выбирать в зависимости от числа соседей, используемых в алгоритме: $\rho = \dfrac{1}{к}$; при увеличении числа *к-БС* параметр ρ будет стремиться к нулю, а результаты взвешивания приближаться к тем, которые получаются по формуле (3.12). В некоторых случаях для придания большего веса *к*-му соседу можно использовать формулу расчета параметра ρ в зависимости от числа классов в исходной выборке: $\rho = \dfrac{K}{к}$.

Весовую функцию в методе *к*-БС можно сформировать и другими способами, сделав ее функцией не от расстояния, а от ранга соседа: $\omega_j = к - j + 1$ (такая функция легче вычисляется, однако обеспечивает худшее качество классификации) или применив нечеткое взвешивание с помощью специальным образом подобранной функции принадлежности.

3.10. Адаптивные методы ближайшего соседа

Как отмечалось выше, важным фактором, влияющим на качество классификации с помощью метода ближайшего соседа, является используемая метрика расстояния. Данная проблема становится особенно значимой в условиях высокой размерности задачи, когда слабо информативные признаки могут существенным образом влиять на положение точки в многомерном пространстве и определять то, какими будут ее соседи.

В ряде работ для МБС осуществляется разработка адаптивных метрик, присваивающих различный вес исходным терминам. Простейшей из них является взвешенное евклидово расстояние, в которое вводится дополнительный вес термина ω_i, характеризующий его дискриминирующую способность:

$$d(\vec{X}_j, \vec{X}_l) = \sqrt{\sum_{i=1}^{M} \omega_i (x_j^{(i)} - x_l^{(i)})^2}. \qquad (3.14)$$

Вес ω_i обычно подбирается экспериментально по обучающим выборкам, назначается экспертно или вычисляется на основе какой-либо оптимизационной процедуры. В качестве адаптивных метрик на практике часто применяют *расстояние Махаланобиса, метрику на основе* χ^2*-критерия, экспоненциальную метрику* и так называемые локальные метрики, вид которых может изменяться в зависимости от положения классифицируемого наблюдения в признаковом пространстве [1], [15], [10].

Рассмотрим *адаптивный метод ближайшего соседа (АМБС)* на основе *модели библиографического документа*, в которой учитывается дискриминирующая способность термина в зависимости от его местоположения в документе.

Предполагается, как это часто делается в информационно-поисковых системах, что любой библиографический документ состоит из семантически неравнозначных полей – названия, аннотации, ключевых слов, т.е. разделяющая способность термина зависит от того, в какой части документа он находится.

Для описания библиографического документа предложена линейная модель следующего вида:

$$x_j^{(i)} = \alpha t_{ij} + \beta a_{ij} + \gamma k_{ij}\ , \qquad (3.15)$$

где $x_j^{(i)}$ – результирующий вес термина i в документе j; t_{ij} – вес термина i в названии документа j; a_{ij} – вес термина i в аннотации документа j; k_{ij} – вес термина i в ключевых словах документа j; α, β, γ – соответствующие весовые коэффициенты, настраиваемые экспериментально.

Для нахождения значений α, β, γ могут быть применены две стратегии. Согласно первой стратегии коэффициенты рассчитываются как отношение ошибок, получаемых в процессе классификации текстовых документов при

использовании только терминов из названий или аннотаций или ключевых слов. Проведенные исследования показали, что наименьшая точность получается при классификации с применением терминов из аннотаций, весовой коэффициент для слов, имеющих такое местоположение, был принят за единицу ($\beta = 1$). Другие коэффициенты вычислялись по формулам $\alpha = \frac{\Delta_a}{\Delta_t}$, $\gamma = \frac{\Delta_a}{\Delta_k}$. Здесь $\Delta_a, \Delta_t, \Delta_k$ – ошибки классификации при раздельном использовании терминов из аннотаций, названий и ключевых слов.

Для определения коэффициентов по второй стратегии рассчитываются отношения количества терминов (без учета частоты их появления), которые встречаются в названиях, аннотациях и ключевых словах. При этом веса терминов в аннотациях по-прежнему не изменялись: $\beta = 1$, а α, γ находились по формулам $\alpha = \frac{M_a}{M_t}$, $\gamma = \frac{M_a}{M_k}$, где M_a, M_t, M_k – количество терминов в аннотациях, названиях и ключевых словах.

При проведении исследований на различных выборках первая стратегия обычно не дает устойчивых результатов (коэффициенты α, γ изменяются в широких пределах). Вторая стратегия, как представляется, является более эффективной и позволяет определить интервалы изменения коэффициентов α и γ ($\alpha \in [2{,}725; 3{,}094]$, $\gamma \in [4{,}166; 4{,}981]$), т.е. в формуле (3.15) в качестве значений коэффициентов целесообразно использовать средние значения из экспериментально определенного интервала: $\alpha = 2{,}9$; $\gamma = 4{,}57$; $\beta = 1$.

3.11. Модифицированные методы ближайшего соседа

В модифицированных методах или *методах ускоренного поиска ближайшего соседа* (*Fast Nearest Neighbor Searching*) производится упорядочивание исходной выборки с использованием различных эвристик с целью более быстрого нахождения БС при классификации нового наблюдения $\vec{X}_{N+1}$. В модифицированных методах принцип принятия решения об отнесении объекта к классу остается прежним – применяется правило БС (или k-БС). Однако для этого используется не последовательное вычисление всех расстояний с целью определения ближайшего соседа, а поиск БС в модифицированной, специально организованной, упорядоченной структуре, полученной на основе индексирования обучающей выборки.

Для упорядочивания объектов в исходной выборке используются несколько подходов: *построение деревьев решений* и *извлечение правил* для определения узла дерева, в котором может находиться БС; *проецирование многомерных наблюдений в двухмерное или трехмерное пространство*, например с помощью метода главных компонент, и поиск БС по двум или трем координатам; *построение диаграмм Вороного* и определение, в какую ячейку

диаграммы попадает классифицируемое наблюдение [36]. Эффективность этих подходов сильно зависит от внутренней структуры и размера выборки и не обеспечивает достаточную точность в случае большого количества признаков, что характерно для задач текстовой классификации. В этой связи более обоснованным представляется использование различных эвристик с целью проведения сортировки элементов обучающей выборки и нахождения такой организационной структуры, которая позволит значительно сократить количество вычислительных операций при классификации нового наблюдения без потери точности. При этом точностные характеристики и быстродействие модифицированного метода будут определяться, в первую очередь, обобщающей силой найденной эвристики.

3.11.1. Разработка модифицированного метода ближайшего соседа – ММБС

Данный алгоритм предусматривает наличие стадии обучения и модифицирует МБС так, чтобы существенным образом сократить количество вычислительных операций, необходимых для проведения классификации, и тем самым увеличить быстродействие.

Целью алгоритма является определение области в M-мерном пространстве, в которую попадает новое наблюдение $\vec{X}_{N+1}$, и использование для классификации только тех $\vec{X}_l (l = 1,...,N_l, N_l << N)$, которые принадлежат выявленной области.

Для этого предлагается выбрать *опорные точки* $P_1,...,P_S$, расположенные на достаточном расстоянии друг от друга, например являющиеся центрами различных классов. Предполагается, что количество опорных точек может изменяться от минимального значения, равного трем ($S_{\min}$=3), до максимального значения, равного количеству классов ($S_{\max} = K$). Дальнейшее изложение алгоритма проводится для простейшего случая трех опорных точек.

Для классификации нового наблюдения $\vec{X}_{N+1}$ строится шесть гиперсфер с центрами в точках P_1, P_2, P_3 и радиусами $R_1 \pm \Delta R, R_2 \pm \Delta R, R_3 \pm \Delta R$ (R_j – расстояние от опорной точки до $\vec{X}_{N+1}$. Рекомендации по выбору приращений ΔR_j будут даны ниже). После этого ведется поиск общей многомерной области пересечения трех гиперколец в пространстве (для двухмерного случая – заштрихованный участок на рис. 3.8).

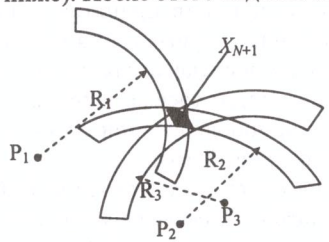

Рис. 3.8

Решающее правило в ММБС имеет вид: новое наблюдение $\vec{X}_{N+1}$ относится к тому классу, к которому принадлежит ближайшее к нему наблюдение $\vec{X}_l$ $(l = 1,...,N_l)$, расположенное в области пересечения трех гиперколец (если таких точек окажется несколько, то решение принимается с использованием "голосования большинства" аналогично методу *к- ближайших соседей*).

Алгоритм ММБС состоит из последовательного выполнения следующих шагов.

1. Определение центров классов, расчет евклидовых расстояний между ними, выбор опорных точек P_1, P_2, P_3.

2. Вычисление расстояний:

- расчет расстояний от всех $\vec{X}_l (l = 1,...,N)$ обучающей выборки до трех опорных точек, получение N-мерных векторов расстояний:

$$\vec{d}_1 = \begin{bmatrix} d_1^{(1)} \\ d_1^{(2)} \\ \vdots \\ d_1^{(N)} \end{bmatrix}; \quad \vec{d}_2 = \begin{bmatrix} d_2^{(1)} \\ d_2^{(2)} \\ \vdots \\ d_2^{(N)} \end{bmatrix}; \quad \vec{d}_3 = \begin{bmatrix} d_3^{(1)} \\ d_3^{(2)} \\ \vdots \\ d_3^{(N)} \end{bmatrix}; \tag{3.16}$$

- проведение сортировки внутри векторов $\vec{d}_1, \vec{d}_2, \vec{d}_3$ так, чтобы элементы $d_1^{(l)}, d_2^{(l)}, d_3^{(l)}$ располагались по возрастанию расстояния до опорных точек (от самых близких к самым дальним), и расширение векторов $\vec{d}_1, \vec{d}_2, \vec{d}_3$ до матриц $\mathbf{D}_1, \mathbf{D}_2, \mathbf{D}_3$ размерностью $[N \times 2]$. Добавленный столбец содержит целочисленные значения, соответствующие исходному (до сортировки) номеру элемента:

$$\vec{d}_1 = \begin{bmatrix} d_1^{(1)} \\ d_1^{(2)} \\ \vdots \\ d_1^{(i)} \\ \vdots \\ d_1^{(N)} \end{bmatrix} \Rightarrow D_1 = \left[\begin{array}{c|c} d_1^{(j)} & j \\ d_1^{(r)} & r \\ \vdots & \vdots \\ d_1^{(s)} & s \\ \vdots & \vdots \\ d_1^{(m)} & m \end{array}\right], \tag{3.17}$$

где j,r,s,m,i – порядковые номера примеров в исходной выборке.

3. Расчет расстояний от нового наблюдения $\vec{X}_{N+1}$ до трех опорных точек $d_1^{(N+1)}, d_2^{(N+1)}, d_3^{(N+1)}$. Определение $d_j^{(l)}$ (l=1,..,N; j=1,2,3) из первого столбца упорядоченных матриц $\mathbf{D}_1, \mathbf{D}_2, \mathbf{D}_3$, наиболее близких к

$d_1^{(N+1)}, d_2^{(N+1)}, d_3^{(N+1)}$. Расстояния $d_j^{(l)} < d_j^{(N+1)} < d_j^{(l+1)}$ будут использоваться в качестве радиусов:

$$R_j = d_j^{(l)};\ \Delta R_j = d_j^{(l+1)} - d_j^{(l)}. \qquad (3.18)$$

4. Определение номеров точек из второго столбца упорядоченных матриц $\mathbf{D}_1, \mathbf{D}_2, \mathbf{D}_3$, соответствующих найденным на предыдущем шаге расстояниям $d_j^{(l)}$. Поиск общих точек, находящихся в области пересечения трех гиперсфер. Для этого анализируются на совпадение точки, соответствующие $d_1^{(l)}$ и $d_1^{(l+1)}$, $d_2^{(l)}$ и $d_2^{(l+1)}$, $d_3^{(l)}$ и $d_3^{(l+1)}$.

5. В случае если на предыдущем шаге обнаружить общие точки не удалось, увеличивается ΔR_j ($\Delta R_j = d_j^{(l+2)} - d_j^{(l)}$ или $\Delta R_j = d_j^{(l+1)} - d_j^{(l-1)}$). Теперь поиск общих точек проводится среди расширенного числа многомерных наблюдений $d_1^{(l)}, d_1^{(l+1)}, d_1^{(l+2)}; d_2^{(l)}, d_2^{(l+1)}, d_2^{(l+2)}; d_3^{(l)}, d_3^{(l+1)}, d_3^{(l+2)}$. Увеличение ΔR_j проводится аналогичным образом до тех пор, пока не обнаружатся общие точки.

Для ММБС, как и для большинства других алгоритмов классификации, точность существенным образом зависит от пространственной ориентации точек в *M*-мерном пространстве. При этом возможно несколько типовых случаев:

- новое наблюдение $\vec{X}_{N+1}$ попадает в скопление точек и находится:
  - близко от центра класса;
  - на границе классов;
- новое наблюдение $\vec{X}_{N+1}$ расположено на периферии (далеко от скопления основного числа точек класса);
- новое наблюдение $\vec{X}_{N+1}$ находится вне пределов признакового пространства, определенного элементами исследуемой выборки (т.е. является шумовым элементом).

При классификации периферийного наблюдения $\vec{X}_{N+1}$ отличие от основной процедуры заключается в том, что в матрицах $\mathbf{D}_1, \mathbf{D}_2, \mathbf{D}_3$ может отсутствовать точка, для которой будет справедливо $d_j^{(N+1)} < d_j^{(l+1)}$, т.е. $\vec{X}_{N+1}$ является самым удаленным наблюдением от одной из опорных точек. В этом случае параметр ΔR_j может быть рассчитан по формуле $\Delta R_j = d_j^{(l)} - d_j^{(l-1)}$.

Для отсечения шумовых элементов вводится значение $L_{\max} = \frac{t}{3}(d_1^* + d_2^* + d_3^*)$, где d_1^*, d_2^*, d_3^* – самые удаленные элементы упорядоченных по расстоянию матриц $\mathbf{D}_j$ (j=1,2,3), а t – экспериментально устанавливаемый коэффициент. В наших исследованиях мы использовали t =2 и считали, что если расстояние хотя бы от одной из опорных точек до нового наблюдения $\vec{X}_{N+1}$ превосходит значение $L_{\max}$, то происходит *"отказ от классификации"*, т.е. считается, что $\vec{X}_{N+1}$ является шумовым и нехарактерным для исследуемой выборки наблюдением.

В *ММБС* от удачного выбора опорных точек во многом зависит эффективность классификации. Предлагается опорные точки выбирать на основе выполнения следующих требований:

- они должны принадлежать различным классам, желательно наиболее удаленным друг от друга;
- они должны находиться в областях скопления многомерных наблюдений и не являться граничными или краевыми точками.

Первый подход реализован в следующем *алгоритме выбора опорных точек*.

1 шаг. Для имеющихся в обучающей выборке *K* классов вычисляются центроиды.

2 шаг. Определяются расстояния ρ_{jl} между центроидами (l,j = 1,...,*K*, $l \neq j$).

3 шаг. Центры классов, которые имеют наибольшее ρ_{jl}, становятся опорными точками P_1, P_2, в качестве третьей опорной точки выбирается центр того класса, который является равноудаленным от P_1, P_2. Следующая опорная точка также должна быть по возможности равноудаленной от уже найденных (максимальное количество опорных точек в данном подходе равняется количеству классов).

Однако определять опорные точки можно не только на основе расчета расстояний между центроидами, но и в зависимости от "населенности" (плотности точек в классе), его компактности (близости всех точек к центру класса), однородности и размера. В этих целях можно эффективно использовать *метод потенциальных функций,* определяя "потенциал", который наводится в центроиде всеми элементами класса (см. п. 3.2). Для определения опорных точек с помощью метода потенциальных функций, так же как и в предыдущем алгоритме, первоначально производится расчет центроидов по обучающей выборке, а затем, вместо расстояний между ними, вычисляются "потенциалы", создаваемые в каждом центре класса. В качестве опорных точек выбираются центроиды с наибольшим "потенциалом".

Таким образом, в разработанной модификации МБС предусмотрена настройка нескольких параметров: приращения радиуса ΔR_j и количества опорных точек *S*. В наших исследованиях при их выборе мы исходили из нескольких критериев: *точности классификации*, получаемой на экзаменационной выборке, *быстродействия* (скорости распознавания новых примеров) и *времени обучения*. Настройка параметров осуществлялась от "простого к сложному", т.е. вначале проводился поиск значения приращений ΔR_j для минимального количества опорных точек S_1 (S_1 =3), затем, если желаемая

точность не достигалась, подбирались значения ΔR_j для большего количества опорных точек:

$$\left\{\begin{matrix}\Delta R\\ S_1\end{matrix}\right\}\cdots\left\{\begin{matrix}\Delta R\\ S_K\end{matrix}\right\};\ S_1 \subset \ldots \subset S_K, \text{ где } S_K = K.$$

При этом из нескольких решений с приблизительно одинаковой точностью выбиралось то, для которого обеспечивалось более высокое быстродействие и затрачивалось меньшее время на обучение.

Обсуждение свойств разработанного метода (ММБС)

В рамках обсуждения предложенного модифицированного метода ближайшего соседа, на наш взгляд, необходимо рассмотреть несколько вопросов:

- как можно оценить плотность распределения вероятностей в многомерной области пересечения гиперколец Ω;
- соответствуют ли те κ-БС, которые попадают в Ω, реальным ближайшим соседям в выборке до проведения ее сортировки.

Далее данные вопросы обсуждаются более подробно.

Вероятность попадания любой точки выборки в многомерную область Ω, образованную в результате пересечения гиперколец, есть некоторое положительное число P, которое зависит от функции плотности распределения случайной величины $\vec{X}$ – $f(\vec{X})$:

$$P = \int\limits_{\vec{X}\in\Omega} f(\vec{X})d\vec{X}. \qquad (3.19)$$

Предположим, что все N элементов выборки извлекаются из генеральной совокупности независимо и в соответствии с вероятностным законом $f(\vec{X})$. Вероятность попадания κ элементов из выборки размером N в многомерную область Ω может быть задана биноминальным законом:

$$P_\kappa = \binom{N}{\kappa} P^\kappa (1-P)^{N-\kappa}. \qquad (3.20)$$

При этом для биноминального закона отношение частот $\frac{\kappa}{N}$ будет хорошей оценкой вероятности P, т.е.

$$P = \frac{\kappa}{N}. \qquad (3.21)$$

Предположим, что изменения $f(\vec{X})$ в пределах многомерной области Ω незначительны, тогда можно записать [4]:

$$\int\limits_{\vec{X}\in\Omega} f(\vec{X})d\vec{X} \approx f(\vec{X})V. \qquad (3.22)$$

Здесь $\vec{X}$ – точка внутри многомерной области Ω, имеющей объем V. Из уравнений (3.21) и (3.22) можно получить:

$$f(\vec{X}) \approx \frac{P}{V} = \frac{\kappa/N}{V}. \qquad (3.23)$$

Образуем последовательность многомерных областей $\Omega_1,\ldots,\Omega_s$, содержащих $\vec{X}$ и имеющих объемы $V_1,\ldots,V_s$. Данные объемы увеличиваются, пока в них не попадет определенной количество κ-БС. Рассмотрим, например, Ω_s с объемом V_s и κ_s – количеством точек, попадающим в Ω_s. Тогда оценку функции плотности распределения $f(\vec{X})$ можно определить по формуле

$$\hat{f}_s(\vec{X}) = \frac{\kappa_s/N}{V_s}. \qquad (3.24)$$

Из анализа формулы (3.24) можно сделать вывод: для того чтобы $\hat{f}_s(\vec{X})$ при $N \to \infty$ сходилась к $f(\vec{X})$, необходимо выполнение трех условий:

1) $\lim\limits_{N\to\infty} V_s = 0$;

2) $\lim\limits_{N\to\infty} \kappa_s = \infty$;

3) $\lim\limits_{N\to\infty} \frac{\kappa_s}{N} = 0$.

Три полученные условия сходимости можно объединить в единой формулировке: если размер выборки возрастает неограниченно и при этом объем многомерной области уменьшается до нуля (хотя все еще содержит очень большое число элементов, которые, однако, составляют лишь незначительно малую часть от всей бесконечной выборки), обеспечивается сходимость $\hat{f}(\vec{X}) \to f(\vec{X})$. Таким образом, в малой окрестности новой точки $\vec{X}_{N+1}$, подлежащей классификации, действительно окажутся ее ближайшие соседи.

К сожалению, эти теоретические выводы ничего не говорят о том, как конкретно надо выбирать V_s и κ_s, чтобы получить хорошие результаты на конечных выборках. Если V_s слишком мал, то оценка $\hat{f}_s(\vec{X})$ может быть некорректной из-за малой населенности области Ω_s, если слишком большой, то, наоборот, большая населенность приведет к нежелательным отклонениям оценки от действительных значений $f(\vec{X})$. Однако в любом случае размер V_s определяется структурой данных, т. к. если плотность распределения вблизи $\vec{X}$ высокая, то объем будет иметь малую величину, если плотность распределения малая, то объем будет увеличиваться, и рост области Ω_s при-

остановится только после ее вступления в область более высокой плотности распределения.

3.12. Редуцированные методы ближайшего соседа

При классификации с помощью *модифицированного* МБС удается провести упорядочивание исходной выборки и увеличить быстродействие метода за счет сокращения операций при поиске ближайшего соседа. Однако на практике точность МБС и его модификаций часто уменьшается из-за наличия в выборке нерелевантных шумовых элементов. С целью повышения робастности семейства методов ближайшего соседа, обеспечения устойчивости их результатов при наличии шумовых примеров разрабатываются *редуцированные методы ближайшего соседа (РМБС).*

К методам редукции предъявляются следующие требования:

- уменьшение количества примеров, используемых для обучения классификатора;
- точность распознавания не должна ухудшаться в результате проведения редукции (она может даже улучшиться, если алгоритм редукции способен удалить из исходной выборки шумовые примеры);
- скорость обучения должна быть приемлемой.

Для проведения редукции используется два основных подхода: объединение нескольких элементов одного класса, *т.е. переход от первоначальной выборки к сокращенной путем слияния исходных примеров;* непосредственное удаление наблюдений *при условии, что это не увеличивает ошибку классификации.*

3.12.1 Метод нахождения прототипов

Первый подход реализован в *методе прототипов*, в котором ближайшие друг к другу элементы обучающей выборки, принадлежащие одному классу, объединяются в прототип $\vec{P}$, если их слияние не нарушает классификацию (т.е. все $\vec{X}_j (j = 1,..., N-2)$ по-прежнему относятся к правильным классам). Элементы нового вектора $\vec{P}$ могут быть найдены как средневзвешенные значения признаков объединяемых векторов. Слияние примеров завершается при увеличении ошибки классификации [37].

Рассмотрим данный алгоритм более подробно. Допустим, что имеется обучающая выборка $T = \{\vec{X}_j, Q_k\}$. Ближайшие друг к другу элементы выборки $\vec{X}_j$ и $\vec{X}_l$ объединяются в прототип $\vec{P}$, если их слияние не нарушает классификацию (т.е. все $\vec{X}_j$ по-прежнему относятся к правильным классам). Элементы нового вектора $\vec{P}$ могут быть найдены как, например, средние взвешенные значения векторов $\vec{X}_j$ и $\vec{X}_l$. Слияние примеров завершается, как только ошибка классификации начинает увеличиваться.

Итак, алгоритм заключается в следующем: дана обучающая выборка. На каждом этапе все примеры $\vec{X}_j$ могут принадлежать двум множествам ***A*** и ***B***.

1. $A=0, B=T$.
2. Выбирается произвольная точка $\vec{X}_l = \vec{P} \in B$ и помещается в ***A***.
3. Находится такая точка $\vec{X}_j \in B$, что расстояние $d(\vec{X}_j, \vec{X}_l)$=min.
4. Осуществляется объединение $\vec{X}_l \cup \vec{X}_j = \vec{P}^*$, если они имеют один и тот же класс; $\vec{X}_l$ и $\vec{X}_j$ заменяются на $\vec{P}^*$. Если при этом распознавание элементов в множестве ***T*** не ухудшилось, то объединение признается успешным, $\vec{X}_l$ и $\vec{X}_j$ удаляются из ***A*** и ***B*** соответственно, а $\vec{P}^*$ помещается в ***A***. Повторяется шаг 2.
5. В случае, если $\vec{X}_l$ и $\vec{X}_j$ нельзя объединить, т.к. они относятся к разным классам, $\vec{X}_j$ перемещается из ***B*** в ***A***, и процедура повторяется с шага 2, пока ***B*** не станет пустым.

Для того чтобы формализовать процедуру объединения $\vec{X}_l$ и $\vec{X}_j$, вводят два расстояния: ω_j – расстояние между $\vec{X}_j$ и ближайшим примером того же класса; b_j – расстояние между $\vec{X}_j$ и ближайшим противником (примером, принадлежащим другому классу).

Чтобы все примеры из **T** классифицировались правильно с помощью правила ближайшего соседа, необходимо, чтобы для любого $\vec{X}_j$ выполнялось неравенство: $\omega_j < b_j$. Если до слияния $\vec{X}_l$ и $\vec{X}_j$ соответствовали ω_j и b_j, а после слияния – ω_j' и b_j' и при этом $\omega_j < b_j$ и $\omega_j' \geq b_j'$, то это означает, что пример $\vec{X}_j$ правильно классифицировался до слияния и ошибочно после, т.е. слияние $\vec{X}_l$ и $\vec{X}_j$ невозможно.

3.12.2. Декрементные методы редукции

Второй подход заключается в поиске эвристики, которая позволяет удалить из исходной выборки шумовые и неинформативные (дублирующие друг друга) примеры. Такие алгоритмы бывают двух видов: инкрементные и декрементные. *Инкрементные алгоритмы* функционируют следующим образом: из обучающей выборки отбираются примеры, удовлетворяющие определенному критерию включения в новую редуцированную выборку. *Декрементные алгоритмы* основаны на удалении из обучающей выборки примеров, которые удовлетворяют выбранному условию исключения. Для обоих подходов особую важность имеет очередность предъявления примеров. Декрементные алгоритмы обычно более ресурсозатратные, однако они обеспечивают лучшую точность. Основным достоинством инкрементных алгоритмов является возможность включения в редуцированную выборку новых примеров на этапе дообучения (в режиме *on-line*). Далее в работе будут рассмотрены наиболее известные декрементные алгоритмы.

В *сжатом методе ближайшего соседа (Condensed NN)* предлагается случайным образом формировать редуцированное множество ***S***, выбирая из

исходной обучающей выборки (множества X) фиксированное количество наблюдений каждого класса, после чего остающиеся в множестве X примеры проходят классификацию. Если они правильно распознаются, значит, случайный выбор наблюдений, помещенных в S, был удачен. В случае неправильного определения класса пример, который был неправильно распознан, добавляется в S. Обучение продолжается до тех пор, пока в X не останется наблюдений, которые были бы обработаны неправильно. Однако отметим, что данный алгоритм особенно чувствителен к шумовым примерам, потому что именно они чаще всего ошибочно классифицируются и включаются в редуцированную выборку. В *редактируемом методе ближайшего соседа (Edited NN)* проводится фильтрация шума, получившая название *"редактирования"*. При этом пример из обучающей выборки **X** удаляется, если его класс определяется неверно голосованием *к*-БС. Часто подобная фильтрация приводит к удалению слишком большого числа элементов из исходной выборки. В одной из модификаций *редактируемого метода ближайшего соседа* пример не включается в новое множество, если все его *к*-ближайшие соседи принадлежат одному классу, т.е. можно предположить, что он находится внутри однородной области признакового пространства, например около центроида.

В *выборочном методе ближайшего соседа (Selective Nearest Neighbor)* вводится определение *"связанного" примера (партнера)*: $\vec{X}_j$ – "связанный" пример для $\vec{X}_l$, если $\vec{X}_l$ является одним из *к*-БС для $\vec{X}_j$. "Связанный" пример $\vec{X}_j$ может принадлежать тому же классу, что и $\vec{X}_l$, т.е. являться положительно "связанным" примером, или $\vec{X}_j$ и $\vec{X}_l$ могут принадлежать разным классам, тогда $\vec{X}_j$ – отрицательно "связанный" пример. Важной характеристикой примера $\vec{X}_l$ является то, как много он имеет "связанных" соседей, насколько данный пример активно участвует в процессе обучения классификатора.

Алгоритм предусматривает нахождение партнеров для каждого $\vec{X}_l$. Такие партнеры могут быть описаны двоичной матрицей **A** размера [N×N], при этом единицы в столбце l обозначают, что соответствующее наблюдение является партнером для $\vec{X}_l$ (l=1,...,N), а единицы в строке j показывают, для каких наблюдений $\vec{X}_j$ (j=1,...,N) является партнером. Задача заключается в нахождении наименьшего количества таких наблюдений, которые смогут "представлять" своих партнеров в новом, редуцированном множестве **S**.

Таким образом, редуцированные методы позволяют сократить число примеров, которые в дальнейшем используются для классификации новых наблюдений, формируя сокращенное множество **S** (**S**∈ **X**) (рис. 3.9). В результате удается уменьшить время вычисления и увеличить скорость классификации.

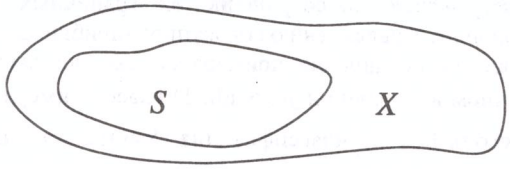

Рис.3.9

3.12.3. Разработка процедуры фильтрации шума

На практике часто встречаются задачи классификации, в которых границы между классами имеют сложную конфигурацию. В этих случаях можно использовать *две стратегии удаления примеров*.

1. Из исходной выборки прежде всего удаляются примеры, находящиеся около границы группы, т.е. самые близкие к примерам другого класса.

2. Из исходной выборки прежде всего удаляются примеры, находящиеся в центре группы, т.е. расположенные дальше всего от примеров другого класса.

Преимущество *первой стратегии* в том, что вблизи границы обычно располагаются шумовые примеры, удаление которых способно значительно улучшить точность классификации. Однако при этом может возникнуть ситуация, когда удаление граничных нешумовых элементов приведет к искажению границ классов, а значит, к увеличению ошибки распознавания.

Вторая стратегия основана на предположении, что "внутренние" (близкие к центру класса) точки оказывают значительно меньшее влияние на принятие решения о классификации нового объекта, чем "граничные" (именно в области границ классов провести классификацию наиболее сложно, см. рис. 1.2). Однако, если большинство "граничных" точек – шумовые, вторая стратегия может привести к удалению из групп наиболее характерных, типовых "внутренних" точек, что способно крайне негативно сказаться на качестве классификации.

Предлагаемый алгоритм основывается на двух известных в литературе редукционных подходах:

- проведения "редактирования" среди исходных данных;
- "мягкого" удаления примеров из выборки, т.е. исключение происходит только в том случае, если это не ухудшает точность распознавания примеров, остающихся в выборке.

Как отмечалось выше, использование только первого подхода для ряда выборок может привести к агрессивной фильтрации, когда из выборки исключается слишком большое количество примеров и происходит искажение границ между классами. Чтобы не допустить ухудшения точности классификации после *"редактирования"*, применяется второй подход, позволяющий проводить "мягкое" сглаживание границ классов, нацеленное, в отличие от про-

цедуры *"редактирования"*, на сохранение тех "граничных" примеров, удаление которых приводит к увеличению совокупной ошибки классификации.

Таким образом, исключение примера из исходной выборки происходит при одновременном выполнении условий: 1) класс примера $\vec{X}_j$ неправильно определяется его *к*-БС, 2) классификация "связанных" примеров $\vec{X}_j$ не ухудшается в его отсутствии (для проверки этого условия используется *к*+1 пример).

Алгоритм РМБС состоит из следующих шагов.

1. Строится квадратная *матрица "связанности–соседства"* **X** (см. табл. 3.1). Наличие единицы в позиции x_{jl} означает, что $\vec{X}_j$ является "связанным" примером для $\vec{X}_l$, т.е. $\vec{X}_l$ – один из *к*-БС для $\vec{X}_j$. Полученная матрица не будет симметричной, т.к. из того, что $\vec{X}_l$ – сосед $\vec{X}_j$, не следует, что $\vec{X}_j$ также является соседом $\vec{X}_l$. В матрицу **X** вводится дополнительная строка, содержащая коэффициент η (равный сумме всех единиц в столбце *l*), который характеризует, как часто $\vec{X}_l$ принимает участие в классификации "связанных" примеров.

Таблица 3.1

$\vec{X}$	$\vec{X}_1$	…	$\vec{X}_l$	…	$\vec{X}_N$
$\vec{X}_1$	0	…	1	…	1
⋮		…		…	⋮
$\vec{X}_j$	0	…	0	…	0
⋮		…		…	⋮
$\vec{X}_N$	1	…	1	…	0
r	r_1		r_l		r_N

2. Для каждого элемента выборки $\vec{X}_j$ проводится упорядочивание его *к*-БС по убыванию расстояния, кроме того, определяется *n* примеров, которые могут быть дополнительно включены в список *к*-БС в случае исключения одного или нескольких примеров (т.е. по существу составляется список из (*к*+*n*) соседей, обычно *n*=3).

3. Проводится классификация всех элементов выборки с помощью *взвешенного метода к-ближайших соседей* (см. формулу (3.13)), хотя в зависимости от структуры выборки в ряде случаев может быть целесообразным использование простого метода *к*-БС без взвешивания. Определяются примеры $\vec{X}_j$, которые были распознаны неверно своими *к*-БС.

4. Каждый пример $\vec{X}_j$, неправильно распознанный на предыдущем шаге, далее рассматривается не как "связанный" пример, а как пример $\vec{X}_l$ (*l*=*j*),

участвующий в классификации в качестве одного из *к*-БС. Такие элементы выборки будут обозначаться $\vec{X}_{j(j)}$ (обратим внимание на то, что каждый пример выборки одновременно является "связанным" примером, задавая *j*-ую строку матрицы **X**, в то же время он участвует в классификации как чей-то *к*-БС, задавая *l*–ый столбец (*l*=*j*) в матрице **X**). Проводится дополнительная проверка на возможность исключения $\vec{X}_{j(j)}$: определяется μ_1 – количество "связанных" примеров $\vec{X}_j$, которые классифицируются корректно с $\vec{X}_{j(j)}$ в качестве одного из *к* соседей; определяется μ_2 – количество "связанных" примеров, которые были проклассифицированы корректно без $\vec{X}_{j(j)}$, используя *к*+*1* соседа.

5. Шаг 4 повторяется для всех ошибочно сгруппированных на шаге 3 примеров. Очередность на исключение определяется исходя из значений параметра r_l (*l*=*j*), показывающего, как часто пример $\vec{X}_{j(j)}$ являлся чьим-либо ближайшим соседом и использовался в процессе классификации. Первыми проходят проверку на возможность удаления наименее "активные" примеры, чаще всего представляющие периферийные или нерелевантные наблюдения.

Для эффективного функционирования РМБС на стадии обучения, используя ряд выборок фиксированной длины, необходимо провести настройку параметров метода: *к* – количества ближайших соседей; ρ – коэффициента взвешивания; θ – порога "редактирования"; $\mu = \mu_1 - \mu_2$ – параметра, характеризующего "ценность" примера для классификации "партнеров".

Отметим, что при изложении алгоритма предполагалось, что все классы имеют одинаковый размер. В случае, если классы существенно отличаются по количеству содержащихся в них примеров, алгоритм фильтрации изменяется так, чтобы отбор кандидатов на проверку проводился бы преимущественно среди элементов наиболее "населенных" классов.

По существу нами предлагается частный случай РМБС, а именно фильтрация нерелевантных элементов выборки. Такой этап предварительной обработки многомерных данных способен существенно повысить точность метода ближайшего соседа и его модификаций за счет удаления "выбросов", которые не согласуются с другими наблюдениями. Отметим также, что этот подход может быть легко расширен для проведения полномасштабной редукции, т.е. удаления из исходной выборки не только нерелевантных шумовых, но и дублирующих друг друга примеров. С этой целью на шаге 4 изложенного выше алгоритма необходимо организовать проверку на исключение не только для примеров, которые были неправильно сгруппированы своими *к*-БС на этапе "редактирования", но и для всех остальных элементов выборки аналогично тому, как это делается в ряде известных методов. Однако такая агрессивная редукция, как отмечалось ранее, может привести к значительному возрастанию ошибки обобщения, получаемой на экзаменационной выборке.

3.13. Обобщенный метод ближайшего соседа

Предложенное выше семейство методов ближайшего соседа *(взвешенный метод к-БС – Вк-БС* (на основе формулы (3.13)), *адаптивный МБС – АМБС* (на основе модели библиографического документа), *модифицированный МБС – ММБС, редуцированный – РМБС)* позволяет скомпенсировать отдельные недостатки классического МБС, обеспечивая при этом существенное улучшение его точностных свойств. Однако в ряде практических случаев возникает необходимость обработки и анализа текстовых выборок, имеющих сложную внутреннюю структуру. Для получения заданной точности в таких случаях целесообразно использовать комплексный подход, объединяющий преимущества каждой из модификаций (Вк-БС, АМБС, ММБС, РМБС). Этот подход реализован в *обобщенном методе ближайшего соседа (ОМБС)*. Алгоритм ОМБС состоит из трех основных шагов.

1. Выявление информативных признаков с помощью классических критериев (например, *взаимной информативности*, χ^2*–критерий* [15], [20]) или метода АМБС, основанного на расчете весов терминов с помощью модели библиографического документа.

2. Проведение фильтрации шума для исключения из обучающей выборки шумовых примеров с использованием РМБС.

3. Обучение ММБС на редуцированной выборке, полученной после фильтрации шума, и проведение классификации новых наблюдений.

Таким образом, использование в ОМБС на первом шаге АМБС позволяет улучшить точность классификации за счет эффективного выбора весов информативных признаков, а проведение на втором шаге фильтрации шума с помощью РМБС приводит к удалению из обучающей выборки нерелевантных примеров, что также способствует улучшению точностных характеристик. Однако необходимо отметить, что в ОМБС значительно увеличивается время обучения. Более того, для ряда относительно "простых" выборок результаты классификации могут оказаться даже хуже, чем для отдельно взятых рассмотренных выше методов из семейства МБС (в частности, ММБС).

3.14. Заключительные замечания

Рассмотренные методы классификации позволяют в определенной степени воспроизводить фундаментальную способность человеческого мозга распознавать, обобщать, разделять, выявлять имеющиеся в данных закономерности.

Вместе с тем существующие численные процедуры группировки данных могут привести к сильно различающимся результатам, в ряде случаев они "навязывают" отличную от истинной систематизацию данных, поэтому вряд ли имеет смысл применять эти методы рутинно, без особых раздумий и дополнительных исследований. Как правило, для выбора из имеющегося разнообразия методов того, который может быть наиболее эффективен для решения конкретной задачи, исследователь должен проявить большую искушенность в теории, в частности, знать методы многомерного статистического анализа и систем искусственного интеллекта. Более того, в ряде случаев он должен принять по собственному усмотрению (обычно на основе проведенных экспериментов и опыта) некоторые интуитивные "внестатистические" решения.

Необходимо отметить, что большинство проблем при классификации реальных выборок возникает по двум причинам: выбранный *метод* не соответствует поставленной задаче (не выполняются предположения, в рамках которых предполагается его функционирование); *данные*, отобранные для исследования отражают свойства лишь конкретной выборки, а не всей генеральной совокупности (из-за этого получаются плохие настройки параметров для процедуры классификации).

Другими чрезвычайно важными факторами, влияющими на качество группировки являются структура и размер классов, их количество, степень перекрытия; процедура отбора наиболее информативных признаков и способ их взвешивания; наличие в выборке нехарактерных элементов – выбросов; выбор метрики близости (для тех методов, в которых она рассчитывается, хотя в то же время влияние вида метрики во многом нивелируется воздействием других вышеперечисленных факторов).

На практике представляется целесообразным проводить классификации не одним, а несколькими методами (желательно использующими различные теоретические принципы – решающие правила), сравнивая полученные результаты и выявляя области компетенции каждого из методов. Кроме того, процедуру группировки следует проводить несколько раз, подстраивая параметры методов под реальную выборку, изменяя метрики, способы взвешивания и выявления информативных признаков и т.п. В ряде случаев для улучшения точности классификации необходимо привлекать консультанта по предметной области для экспертной оценки и выработки рекомендаций или организовывать так называемую *обратную связь с пользователем* [3].

Так как не существует единственно правильной классификации (имеется лишь множество возможных решений этой задачи), то полученные результаты лучше всего рассматривать не столько как окончательную структуру, являющуюся объективным свойством многомерных данных, а как некоторую "подсказку", предназначенную для выявления самим исследователем существующих закономерностей и интересующих его характеристик, слишком громоздких для непосредственного анализа. Таким образом, окончательный выбор того или иного результата классификации остается за исследователем и определяется его квалификацией, целью исследования, имеющимися данными.

Контрольные вопросы

1. Какие требования предъявляются к методам классификации текстовой информации?
2. Какое решающее правило используется в центроидном методе классификации?
3. Какое решающее правило используется в методе потенциальных функций?
4. Какое решающее правило используется в наивном байесовском классификаторе?
5. Какое решающее правило используется в методе деревьев решений?
6. В чем заключаются преимущества и недостатки методов коллективной классификации?
7. Какое решающее правило используется в методе ближайшего соседа?
8. Какое решающее правило используется в методе *к*-ближайших соседей?
9. Проведите сравнительный анализ модификаций метода ближайшего соседа?
10. В чем заключается процедура фильтрации нерелевантных примеров?
11. Объясните, за счет чего удается добиться большей точности в обобщенном методе ближайшего соседа?

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. **Прикладная** статистика. Классификация и снижение размерности/ Под ред. С.А. Айвазяна. – М.: Финансы и статистика, 1989. – 607 с.
2. **Загоруйко Н.Г.** Прикладные методы анализа данных и знаний. – Новосибирск: Изд-во Института математики, 1999. – 270 с.
3. **Вапник В.Н., Червоненкис А.Я.** Теория распознавания образов. – М.: Наука, 1974. – 415 с.
4. **Дуда Р., Харт П.** Распознавание образов и анализ сцен. – М.: Мир, 1976. – 511 с.
5. **Солтон Дж.** Динамические библиотечно–информационные системы. – М.: Мир, 1979. – 557 с.
6. **Хонкела Т., Лагус К., Каски С.** Самоорганизующиеся карты для анализа обширных архивов документов. В кн.: Дебок Г., Кохонен Т. Анализ финансовых данных с помощью самоорганизующихся карт. – М.: Альпина. 2001. – с. 232–244.
7. **Корнеев В.В., Гареев А.Ф., Васютин С.В., Райх В.В.** Базы данных. Интеллектуальная обработка информации. – М.: Нолидж, 2000. – 352 с.
8. **Сэлтон Г.** Автоматическая обработка, хранение и поиск информации. – М: Советское радио, 1973. – 560 с.
9. **Айвазян С.А., Мхитарян В.С.** Теория вероятностей и прикладная статистика. Том 1. – М.: ЮНИТИ, 2001. – 656 с.
10. **Ким Дж., Мьюллер Ч., Клекка У. и др.** Факторный, дискриминантный и кластерный анализ. – М.: Финансы и статистика, 1989. – 215 с.
11. **Елисеева И.И., Руковишников В.О.** Логика прикладного статистического анализа. – М.: Финансы и статистика, 1982. – 192 с.
12. **Толчеев В.О., Ягодкина Т.В.** Методы идентификации одномерных линейных динамических систем. – М: Издательство МЭИ, 1997. – 107 с.
13. **Aas K., Eikvil L.** Text Categorization: A Survey. – Norwegian Computing Center, Oslo, 1999, pp.1–37.
14. **Браверман Э.М., Мучник И.Б.** Структурные методы обработки эмпирических данных. – М.: Наука, 1983. – 464 с.
15. **Статистические** методы для ЭВМ/ Под ред. Энслейна К., Рэлстона Э., Уилфа Г. – М.:Наука, 1986. – 464 с.
16. **Дюран Б., Оделл П.** Кластерный анализ. – М.: Статистика, 1977. – 128 с.
17. **Дубров А.М., Мхитарян В.С., Трошин Л.И.** Многомерные статистические методы. – М.: Финансы и статистика, 2000. – 351 с.
18. **Архиреева И.Н., Бородюк В.П., Голяс Ю.Е., Киреева В.Г.** Факторный анализ в задачах обработки экспериментальных данных. – М: Издательство МЭИ, 1994. – 38 с.
19. **Прикладная** статистика. Основы моделирования и первичная обработка данных/ Под ред. С.А. Айвазяна.–М.: Финансы и статистика, 1983. – 471 с.
20. **Вероятность** и математическая статистика: Энциклопедия/ Под ред. Ю.В. Прохорова. – М.: Большая российская энциклопедия, 1999. – 910 с.

21. **Yang Y.** An Evaluation of Statistical Approaches to Text Categorization// Information Retrieval Journal. 1999. 1. P. 67– 88.
22. **Классификация** и кластер/ Под ред. Райзина Дж.– М.: Мир, 1980. – 389 с.
23. **Дейвисон М.** Многомерное шкалирование: методы наглядного представления данных. – М.: Финансы и статистика, 1987. – 254 с.
24. **Терехина А.Ю.** Анализ данных методами многомерного шкалирования. – М.: Наука, 1986. – 168 с.
25. **Статистические** методы в инженерных исследованиях/ Под ред. Г.К. Круга – М.: Высшая школа, 1983. – 216 с.
26. **Айзерман М.А., Браверман Э.М., Розоноэр Л.И.** Метод потенциальных функций в теории обучения машин. – М.: Наука, 1970. – 384 с.
27. **Болотов А.А., Фролов А.Б.** Классификация и распознавание в дискретных системах. – М: Издательство МЭИ, 1997. – 99 с.
28. **Quinlan J.R.** C 4.5: Programs for Empirical Learning. – San Mateo, Morgan Kaufmann, 1992.
29. **Романов В.П.** Интеллектуальные информационные системы в экономике. – М.: Экзамен, 2003. – 494 с.
30. **Cohen W.W., Singer Y.** Context–Sensitive Learning Method for Text Categorization// Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 1996, pp.307–315.
31. **Растригин Л.А., Эренштейн Р.Х.** Метод коллективного распознавания. – М.: Энергоиздат, 1981. – 79 с.
32. **Реброва М.П.** Автоматическая классификация в системах обработки информации: поиск документов. – М.: Радио и связь, 1983. – 97 с.
33. **Айвазян С.А., Бежаева З.И., Староверов О.В.** Классификация многомерных наблюдений. – М.: Статистика, 1974. – 240 с.
34. **Елисеева И.И., Руковишников В.О.** Группировка, корреляция, распознавание образов. – М.: Статистика, 1977. – 143 с.
35. **Dudani S.A.** The Distance–Weighted k–Nearest–Neighbor Rule// IEEE Transactions on Systems, Man and Cybernetics. 1976. Vol. SMC–6. April. P. 325–327.
36. **Препарата Ф., Шеймос М.** Вычислительная геометрия. – М.:Мир,1989. – 478 с.
37. **Chang C–L.** Finding Prototypes for Nearest Neighbor Classifiers// IEEE Transactions on Computers. 1974. Vol. C–23. November. P. 1179–1184.

ОГЛАВЛЕНИЕ

Учебное издание

Толчеев Владимир Олегович

СОВРЕМЕННЫЕ МЕТОДЫ ОБРАБОТКИ И АНАЛИЗА ТЕКСТОВОЙ ИНФОРМАЦИИ

Учебное пособие
по курсу
«Интеллектуальные информационные системы»
для студентов, обучающихся по специальности
«Управление и информатика в технических системах»

Редактор издательства Г.Ф. Раджабова

Темплан издания МЭИ 2004 (II),учеб. Печать офсетная
Подписано в печать 16.01.06 Формат 60x84/16 Физ.печ.л. 4,75
Тираж 200 Изд.№ 186 Заказ16т Цена 14 руб.

Издательство МЭИ, 111250, Москва, ул. Красноказарменная, д.14.

Отпечатано в типографии ФГУП «НИИ «Геодезия», 141292
Московская обл., г. Красноармейск, пр-т Испытателей, д. 14