Методы добычи данных при построение локальной метрики в системах вывода по прецедентам

Л. Е. Карпов, В. Н. Юдин[1]

1. Введение

При современном уровне развития информационных технологий и, более конкретно, систем поддержки принятия решений различают два направления в развитии логического вывода знаний [Каменнова 95]:

- развитие систем логического вывода, основанного на правилах;
- развитие систем логического вывода, основанного на прецедентах.

Практически все ранние экспертные системы моделировали ход принятия решения экспертом как чисто дедуктивный процесс с использованием логического вывода, основанного на правилах. Это означало, что в систему закладывалась совокупность правил вида "*если...то...*", согласно которым на основании входных данных генерировалось то или иное заключение по интересующей проблеме. Выбранная модель являлась основой для создания экспертных систем первых поколений, которые были достаточно удобны как для разработчиков, так и для пользователей-экспертов. Однако с течением времени было осознано, что дедуктивная модель моделирует один из наиболее редких подходов, которому следует эксперт при решении проблемы.

Идея вывода по правилам является привлекательной, потому что она подразумевает наличие хорошо формализованной задачи, для которой существуют научные методы, доказавшие свою применимость и позволяющие получить решение, не требующее доказательств.

Но окружающий мир сложен. Существует много слабо формализованных задач, для которых, возможно, будут найдены решения. Кроме того, существует ряд задач, для которых никогда не будет найдено формальное решение (судопроизводство, медицина). Актуальность проблемы обусловлена и многочисленностью таких задач, и практической потребностью найти хотя бы одно сколько-нибудь подходящее решение там, где из-за отсутствия строго формализованного метода нельзя найти все или самое оптимальное из всех.

На самом деле, вместо того, чтобы решать каждую задачу, исходя из первичных принципов, эксперт часто анализирует ситуацию в целом и

[1] Работа поддержана грантами Российского фонда фундаментальных исследований № 06-07-89098 и № 06-01-00503

вспоминает, какие решения принимались ранее в подобных ситуациях. Затем он либо непосредственно использует эти решения, либо, при необходимости, адаптирует их к обстоятельствам, изменившимся для конкретной проблемы.

Моделирование такого подхода к решению проблем, основанного на опыте прошлых ситуаций, привело к появлению технологии логического вывода, основанного на прецедентах (по-английски – Case-Based Reasoning, или CBR), и в дальнейшем – к созданию программных продуктов, реализующих эту технологию.

В ряде ситуаций метод вывода по прецедентам имеет серьезные преимущества по сравнению с выводом, основанным на правилах, и особенно эффективен, когда:

- основным источником знаний о задаче является опыт, а не теория,
- решения не уникальны для конкретной ситуации и могут быть использованы в других случаях;
- целью является не гарантированное верное решение, а лучшее из возможных.

Таким образом, вывод, основанный на прецедентах, представляет собой метод построения экспертных систем, которые делают заключения относительно данной проблемы или ситуации по результатам поиска аналогий, хранящихся в базе прецедентов.

Системы вывода по прецедентам показывают очень хорошие результаты в самых разнообразных задачах, но обладают рядом существенных недостатков.

Во-первых, они вообще не создают каких-либо моделей или правил, обобщающих предыдущий опыт, – в выборе решения они основываются на всем массиве доступных исторических данных, поэтому невозможно сказать, на основе каких конкретно факторов системы вывода по прецедентам строят свои конкретные ответы.

Можно выделить две основные проблемы, с которыми сталкиваются подобные системы: поиск наиболее подходящих прецедентов и последующая адаптация найденного решения.

В основе всех подходов к отбору прецедентов лежит тот или иной способ измерения степени близости прецедента и текущего случая. При таких измерениях вычисляется численное значение некоторой меры, определяющей состав множества прецедентов, которые нужно обработать для достижения удовлетворительной классификации или прогноза. Основным недостатком таких систем является произвол, который допускают системы при выборе меры близости. Кроме того, безосновательным выглядит распространение общей меры близости на выборку данных в целом.

Еще один недостаток метода связан с конструкцией прецедентов и назначения уместных весов их атрибутам, что уменьшает применимость таких систем в разных предметных областях.

В большинстве случаев методы поиска прецедентов сводятся к индукции деревьев решений или к алгоритму "ближайшего соседа", дополненному, может быть, использованием знаний о предметной области. Что касается

адаптации и использования найденного решения, эта задача до сих пор остается недостаточно формализованной и сильно зависящей от предметной области.

Обе проблемы – поиск прецедентов и адаптация выбранного решения – решаются (полностью или частично) с привлечением фонового знания, иными словами, знания о предметной области (domain knowledge). Существуют разные способы получения информации о предметной области:

1. Привлечение экспертного знания. Оно может выражаться, например, в ограничениях, накладываемых на диапазоны изменений признаков объектов, или же в формулировании набора правил для разбиения базы прецедентов на классы (построение классификатора).
2. Получение необходимых знаний из набора имеющихся данных методами добычи данных (по-английски – Data Mining). Сюда относятся все методы выявления отношений в данных, в частности, кластеризация, регрессия, поиск ассоциаций. Использование методов добычи данных может выделить узкую группу показателей, от которых зависит интересующая исследователя характеристика, и представить обнаруженную закономерность в аналитической форме.
3. Формирование знаний на основе обучающей выборки, представленной экспертом (обучение с учителем). Этот способ включает в себя оба первых.

Изначально в системах вывода по прецедентам в качестве источников фонового знания выступали эксперты – высококвалифицированные специалисты предметных областей, а также текстовые материалы – от учебников до протоколов, и, разумеется, базы данных (имплицитные источники знаний). Роль эксперта (затратная по ресурсам и времени) заключалась в вербализации, то есть переводе таких источников в эксплицитную форму. Учитывая, что важнейшей задачей в процессе формализации извлечения знаний является минимизация роли эксперта, его роль должны взять на себя средства добычи данных.

Среди извлекаемых закономерностей на практике чаще всего встречаются отношения эквивалентности и порядка. Первые присущи, в частности, задачам классификации, диагностики и распознавания образов. С другой стороны, отношения порядка свойственны задачам шкалирования, прогнозирования и т.п.

Авторы поставили перед собой задачу предложить подход к построению интегрированных систем, при котором минимизируются указанные ранее недостатки. Этот подход основан на привлечении дополнительных знаний о предметной области с помощью методов добычи данных – классификации и кластеризации.

Предлагается ввести в базе прецедентов отношения эквивалентности, которые выражают принадлежность оцениваемых объектов к каким-либо классам,

рассматриваемым как самостоятельные семантические единицы. Классы представляют номинальную шкалу (шкала наименований – не количественная, а строго качественная, она не приписывает классам никаких численно выражаемых атрибутов). Можно считать, что объекты, отнесенные к одному и тому же классу, эквивалентны с точки зрения данной номинальной шкалы. Такие классы (или основные понятия) в базе прецедентов могут быть построены различными способами: с помощью привлечения экспертного знания или путем предварительной кластеризации базы прецедентов. Эти классы, в свою очередь, предлагается использовать как основу для предлагаемой меры близости прецедентов.

К сожалению, реальные приложения редко укладываются в рамки фиксированного признакового пространства. Одной из причин этого является недостаток информации в описании объектов (прецедентов или текущего случая). Это приводит к тому, что текущий случай может попасть в смешение понятий, иными словами – в пересечение классов.

Авторы предлагают уйти от распространения общей меры близости на выборку данных в целом, введя понятие локальной контекстно-зависимой метрики для текущего случая. Эта метрика называется локальной, так как она привязывается к текущему случаю, а контекстно-зависимой – из-за того, что она определяется отношениями между объектами. В частности, от степени описания текущего случая зависят проекции классов на пространство его признаков и степень их пересечения. Само понятие пересечения используется при построении этой метрики.

2. Вывод на основе прецедентов в системах поддержки принятия решений

2.1. Концепция вывода

Вывод на основе прецедентов – это метод принятия решений, в котором используются знания о предыдущих ситуациях или случаях (прецедентах). При рассмотрении новой проблемы (текущего случая) отыскивается похожий прецедент в качестве аналога. Вместо того, чтобы искать решение каждый раз сначала, можно пытаться использовать решение, принятое в сходной ситуации, возможно, адаптировав его к изменившейся ситуации текущего случая. После того, как текущий случай будет обработан, он вносится в базу прецедентов вместе со своим решением для его возможного последующего использования в будущем. Более формальное определение дано в [Bundy 97].

Прецедент – это описание проблемы или ситуации в совокупности с подробным указанием действий, предпринимаемых в данной ситуации или для решения данной проблемы.

Согласно [Althof 95/1] прецедент включает:

1. Описание проблемы,
2. Решение этой проблемы,
3. Результат (обоснованность) применения решения.

Описание проблемы должно содержать всю информацию, необходимую для достижения цели вывода (выбора наиболее подходящего решения). Например, если цель состоит в диагностике заболеваний, то описательная информация должна содержать симптомы больного, результаты лабораторных исследований. Если цель – выбор лечения, то понадобятся еще хронология состояния больного, сведения о возможной аллергической реакции на те или иные лекарственные средства и т. д. Все этапы примененного к больному лечения сохраняются в описании решения.

Исход как результат применения решения – это обратная связь, полученная от применения решения. Описание результата может содержать перечень выполненных операций, результат их выполнения, способ восстановления (в случае отказа), указания на то, что можно сделать, чтобы избежать отказа, результаты восстановления. Описание результата может также включать ссылки на другие прецеденты, дополнительную текстовую информацию.

Прецедент может содержать не только положительный исход. Информацию о том, что у больного не наступило улучшение самочувствия в результате примененного лечения, надо сохранять, чтобы избежать бесполезных назначений в будущем. Объяснение того, какой отказ произошел и почему, может быть использовано в будущем. Некоторые системы могут сохранять обоснование решения и даже его альтернативы.

Имеется множество способов представления прецедента: от записей в базах данных, древовидных структур – до предикатов и фреймов. Конкретное выбранное представление прецедентов должно соответствовать общим целям системы. Проблема представления прецедента – прежде всего проблема выбора информации, которую надо включать в описание прецедентов, нахождение соответствующей структуры для описания содержания прецедента, а также определения, каким образом должна быть организована и индексирована база знаний прецедентов для эффективного поиска и многократного использования.

2.2. Декомпозиция метода (основные фазы)

Хотя не все системы вывода, основанного на прецедентах, полностью включают этапы, приведенные ниже (Рис. 1), подход, основанный на прецедентах, в целом состоит из следующих компонентов [Aamodt 94]:

1. Извлечение наиболее релевантных прецедентов для текущего случая из библиотеки прецедентов.
2. Адаптация выбранного решения для текущего случая, если это необходимо.
3. Применение решения.

4. Оценка применения (проверка корректности).
5. Сохранение. Добавление текущего случая в базу прецедентов.

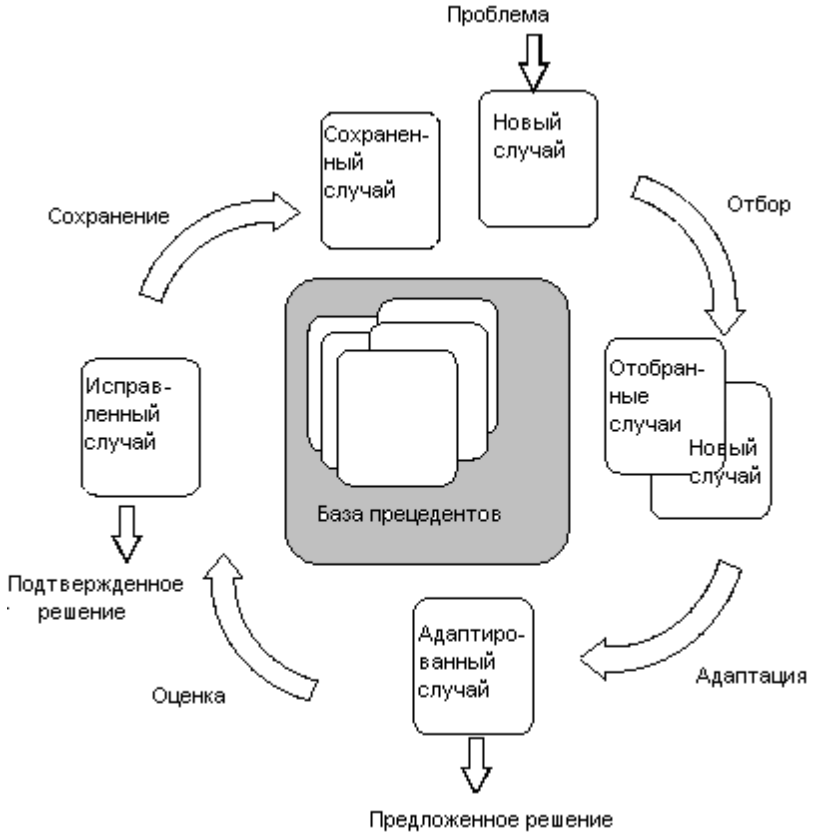

Рис 1. Цикл вывода на основе прецедентов

Проблема выбора подходящего прецедента является одной из самых важных в таких системах. Естественно искать подходящий прецедент в той области пространства поиска, где находятся решения сходных проблем, иначе говоря, поиск должен быть организован сообразно цели. Но как определить, какие именно решения считать сходными?

Эффективность поиска прецедентов для текущего случая во многом зависит от того, по каким признакам организован индекс в базе прецедентов. Это, в свою очередь, требует хороших знаний о предметной области и конечной цели решения проблемы. Однако выбор наилучшего индекса не может быть столь же прост, как это звучит, так как не имеется никаких общих рекомендаций для этого. Для ориентира, однако, можно привести четыре свойства хороших индексов [Kolodner 83]:

1. Направленность: Индексы должны быть направлены на решение цели.
2. Абстрактность: Индексы должны быть достаточно абстрактны, чтобы прецедент мог быть использован в разных запросах.
3. Конкретность: Индексы должны быть распознаваемы в других ситуациях без дальнейшей обработки.
4. Полноценность: Индексы должны быть способны дифференцировать прецеденты.

После того, как прецеденты извлечены, нужно выбрать "наиболее подходящий" из них. Это определяется сравнением признаков текущего случая и выбранных прецедентов. Определение метода, на котором будет основываться нахождение меры сходства прецедентов, решается во время создания системы ее разработчиками. Наиболее популярным и часто используемым является метод "ближайшего соседа" (nearest neighbour) [Anand 99]. В его основе лежит тот или иной способ измерения степени близости прецедента и текущего случая по каждому признаку (будь это текстовый, числовой или булевский), который пользователь сочтет полезным для достижения цели.

Говоря более строгим языком, вводится метрика на пространстве всех признаков, в этом пространстве определяется точка, соответствующая текущему случаю, и в рамках этой метрики находится ближайшая к ней точка из точек, представляющих прецеденты. Описанный здесь алгоритм очень прост – реально применяются некоторые его модификации. Обычно прогноз делается на основе нескольких ближайших точек, а не одной (K-nearest neighbours). Такой метод более устойчив, поскольку позволяет сгладить отдельные выбросы, случайный шум, всегда присутствующий в данных.

Каждому признаку назначают вес, учитывающий его относительную ценность. Полностью степень близости прецедента по всем признакам можно вычислить, используя обобщенную формулу вида:

$$\frac{\sum_j w_j * \mathrm{sim}(x_{ij}, x_{kj})}{\sum_j w_j}$$

где w_j – вес j-го признака, sim – функция подобия (метрика), x_{ij} и x_{ik} – значения признака x_j для текущего случая и прецедента, соответственно. После вычисления степеней близости все прецеденты выстраиваются в единый ранжированный список.

Метод прост, он может быть реализован очень эффективно, правда требует для работы большой памяти, так как в процессе нахождения значения зависимой переменной для новой записи используется вся существующая база данных.

Выбор метрики (или меры близости) считается узловым моментом, от которого решающим образом зависит поиск подходящих прецедентов. В каждой конкретной задаче этот выбор производится по-своему, с учетом главных целей исследования, физической и статистической природы используемой информации и т. п. В некоторых методах выбор метрики достигается с помощью специальных алгоритмов преобразования исходного пространства признаков.

Пусть имеются образцы X_i и X_k в N-мерном пространстве признаков. Основные метрики, традиционно используемые при выборе прецедентов, приводятся в таблице 1.

После того, как выбран подходящий прецедент, при поиске решения для целевой проблемы выполняется адаптация – модификация имеющегося в нем решения с целью его оптимизации. Невозможно выработать единый вариант для такой адаптации, так как это в большой степени зависит от предметной области. Если существуют алгоритмы адаптации, они обычно предполагают наличие зависимости между признаками прецедентов и признаками содержащихся в них решений. Такие зависимости могут задаваться человеком при построении базы прецедентов или обнаруживаться в базе автоматически.

Процесс модификации решения при его адаптации к текущему случаю может включать ряд шагов, от простой замены некоторых компонентов в имеющемся решении, корректировки или интерполяции (числовых) признаков или изменения порядка операций, до более существенных. Имеются и другие подходы:

- Повторная конкретизация переменных в существующем прецеденте и присвоение им новых значений.
- Уточнение параметров. Некоторые прецеденты могут содержать числовые значения, например, время выполнения какого-либо этапа плана. Это значение должно быть уточнено в соответствии с новым значением другого свойства.
- Поиск в памяти. Иногда требуется найти способ преодоления затруднения, возникшего как побочный эффект замены одних компонентов решения другими.

Обратная связь, возникающая при сохранении решений для новых проблем, означает, что вывод по прецедентам по своей сути является "самообучающейся" технологией, благодаря чему рабочие характеристики каждой базы прецедентов с течением времени и накоплением опыта непрерывно улучшаются. Разработка баз прецедентов по конкретной проблеме или области деятельности происходит на естественном (русском, английском) языке, то есть не требует никакого программирования, и может быть выполнена наиболее опытными сотрудниками – экспертами, работающими в данной конкретной области.

Не стоит, однако, рассчитывать, что экспертная система будет действительно принимать решения. Принятие решения всегда остается за человеком, а

система лишь предлагает несколько возможных вариантов и указывает на самый "разумный" из них с ее точки зрения.

Наименование метрики	Тип признаков	Формула для оценки меры близости (метрики)
Эвклидово расстояние	Количественные	$d_{ik} = \left(\sum_{j=1}^{N} (x_{ij} - x_{kj})^2 \right)^{\frac{1}{2}}$
Манхэттенская метрика	Количественные	$d_{ik}^{(1)} = \sum_{j=1}^{N} \lvert x_{ij} - x_{kj} \rvert$
Мера сходства Хэмминга	Номинальные (качественные)	$\mu_{ij}^{H} = \frac{n_{ik}}{N}$, где n_{ik} – число совпадающих признаков у образцов X_i и X_k.
Мера сходства Роджерса-Танимото	Номинальные шкалы	$\mu_{ii}^{R-T} = n_{ik}''(n_i' + n_k' - n_{ik}'')$ где n_{ik}'' – число совпадающих единичных признаков у образцов X_i и X_k; n_i', n_k' – общее число единичных признаков у образцов X_i и X_k соответственно.
Расстояние Махалонобиса	Количественные	$d_{ik}^{M} = (x_{ij} - x_{kj})^T W^{-1} (x_{ij} - x_{kj})$ W – ковариационная матрица выборки $X = (X_1, X_2, \ldots, X_n)$
Расстояние Журавлева	Смешанные	$d_{ik} = \sum_{j=1}^{N} I_{ik}^{j}$, где $I_{ik}^{j} = \begin{cases} 1, \text{если } \lvert x_{ij} - x_{kj} \rvert < \varepsilon \\ 0, \text{иначе} \end{cases}$

Таблица 1. Основные типы метрик.

2.3. Примеры систем вывода на основе прецедентов

Вывод по прецедентам – не новая технология, ее возникновение прослеживается в работе Роджера Шанка по динамической памяти [Schank 82]. Изложенные этим исследователем идеи были далее расширены Джанет Колоднер, которая разработала систему CYRUS [Kolodner 83]. С 1988 года проводился ряд ежегодных семинаров под эгидой DARPA (Управление

Перспективных Исследований Министерства Обороны США) [Kolodner 88, Hammond 89, Bareiss 91]. Интерес к методу вырос в последние годы, регулярно проводятся конференции и семинары типа широко известного Европейского семинара (EWCBR, в последующем – ECCBR) [Wess 93, Haton 94, Smith 96, Smyth 98, Blanzieri 00, Craw 02, Funk 04], семинара Соединенного Королевства (UKCBR) [UKCBR 04, UKCBR 05].

К ранним разработкам относят CHEF [Hammond 86] – систему, которая предназначалась для формирования кулинарных рецептов. Эта программа принимает информацию о целевых характеристиках блюда (тип, вкусовые качества, своеобразие) и формирует подходящий рецепт. Результатом работы программы должен быть рецепт – последовательность операций, позволяющая приготовить такое блюдо. Получив заказ, программа просматривает свою базу прецедентов, отыскивает в ней рецепт приготовления аналогичного блюда и адаптирует его в соответствии с особенностями текущего заказа.

Из других систем можно отметить PROTOS [Bareiss 88] для классификации и диагностирования нарушений слуха, MEDIATOR [Simpson 85] для области посредничества в спорных судебных вопросах.

При работе над экспертными системами исследователи пришли к выводу, что представление знаний внутри экспертной системы должно со временем привести к созданию систем обучения с помощью компьютера.

В основу программы HYPO [Ashley 88, Ashley 90], которая была создана для обучения студентов-юристов методике ведения судебных дел, положена абстрактная модель процесса прения сторон. И расследования, и рассуждения в юриспруденции направляются аргументацией, а более точно – аргументами, выражающими противоположные интересы, с помощью которых стороны процесса пытаются склонить на свою сторону судью или присяжных, убедить их в том, что именно предлагаемая интерпретация закона и фактов является корректной в данном случае.

Базовая предпосылка, сделанная авторами, состоит в том, что эти ходы могут быть описаны в рамках какой-то системы и затем использованы для обучения. Для выполнения очередного хода в игре нужно выбрать прецеденты в базе знаний о прецедентах, причем выбор должен учитывать как информацию о текущем случае, так и возможные ходы оппонентов. Таким образом, поведение сторон можно рассматривать как планирование трехходовой комбинации в игре:

1. одна сторона с помощью своего набора прецедентов "продвигает" свою позицию в игре;
2. противоположная сторона выдвигает другой набор прецедентов для представления своих аргументов;
3. первая сторона наносит новый удар, выдвигая соображения, парирующие в определенной степени аргумент противной стороны.

Ходы и контрходы можно анализировать в терминах суждений, основанных на прецедентах. Авторы системы считают, что в основе такой модели аргументации лежит следующий процесс [Ashley 90]:

1. Сравнение текущего случая с прецедентом с прицелом на обоснование аналогичного результата.
2. Определение отличия (противопоставления) между текущим случаем и прецедентом, чтобы найти аргумент против того же результата.
3. Поиск контрпримера к (1), в котором аналогичный прецедент привел к другому результату.
4. Формулировка гипотетических прецедентов, которые дали бы аргументы за и против определенной позиции.
5. Комбинирование сравнений и противопоставлений в аргумент, который включает оценку конкурирующих аргументов.

Эта модель аргументации реализована в системе HYPO, в которой процесс формирования аргумента выполняется за шесть шагов.

1. Анализ факторов, присущих текущему случаю.
2. Извлечение прецедентов на основе этих факторов.
3. Упорядочение извлеченных прецедентов по степени близости к текущему случаю.
4. Выбор наиболее подходящих прецедентов как с точки зрения одной стороны, так и с точки зрения другой.
5. Формирование аргументов для трехходовой комбинации по каждому из пунктов текущего дела.
6. Проверка результатов на гипотетических случаях.

Выполнение шага (5) усложняется тем обстоятельством, что судебное дело может содержать более одного пункта. Например, дело о разводе может содержать множество пунктов, касающихся раздела имущества, обеспечения детей и т.д., по каждому из которых стороны должны представить свои аргументы.

Что касается доступности коммерческих систем и успеха в информационных приложениях – это система SMART [Acorn 92], которая дала импульс этой технологии. Система SMART предназначена для технической поддержки заказчиков корпорации COMPAQ. Когда заказчик сталкивается с проблемой (например, печать принтера блекнет), подробности передаются в систему. Выполняется начальный поиск в библиотеке прецедентов, чтобы найти случаи с подобными признаками. При недостатке информации система задает дополнительные вопросы. Как только определенный порог достигнут (скажем, прецедент совпадает не менее, чем на 80%), предлагается решение от прецедента. В дополнение к этому, система может быть использована как инструмент обучения.

В дальнейшем COMPAQ расширила эту систему, продвинув ее непосредственно к покупателям. Система QUICKSOURCE [Nguyen 93] позволяет пользователю самому справляться с проблемами и обращаться в центр поддержки в качестве последнего прибежища.

В системе KATE TOOLS компании Acknosoft (Франция) [Althof 95/1] поддерживается упрощенный взгляд на процесс вывода. Входная информация для KATE – это файл, который содержит описания признаков и их значения на специальном языке CASUAL [Althof 95/2]. KATE может работать со сложными данными, представленными в виде структурированных объектов, отношениями или даже общими знаниями о проблемной области. Но для выявления сходства между прецедентами используется одна простая метрика.

Основной акцент делается на отбор прецедентов с помощью алгоритма "ближайшего соседа". KATE использует версию алгоритма ближайшего соседа для вычисления метрики подобия. Близость между двумя случаями x и y, имеющими p признаков вычисляется по формуле:

$$\text{Similarity}(x,y) = -\sqrt{\sum_{i=1}^{p} f(x_i, y_i)}$$, где f определяется как

$$f(x_i, y_i) = \begin{cases} (x_i - y_i)^2 & \text{if } x_i, y_i \text{ are numeric} \\ (x_i \neq y_i) & \text{if } x_i, y_i \text{ are symbolic} \end{cases}$$

Алгоритм работы системы может быть описан следующим образом:

```
Classified Data = 0
for each Case x in Casebase do
 1. for each y in Classified Data do
     Sim(y) = Similarity(y,x)
 2. y_max = (y_1,...,y_k) such that
     Sim(y_k)= max(K-nearest neighbors)
 3. if class(y_max) = class(x)
          then      classification is correct
              Classified Data = Classified Data + {x}
          else classification is incorrect
```

Система KATE не предлагает возможностей для автоматической адаптации решения. Проверка корректности решения невозможна, но есть проверка базы прецедентов на наличие контрпримеров. Все же, KATE – это эффективная индустриальная система, которая позволяет использовать взвешенные признаки при вычислении метрики подобия, а также использовать определяемую пользователем метрику. Ее легко расширять, потому что все функции KATE доступны при подключении сопутствующих динамических библиотек (.dll).

В настоящее время на рынке программных продуктов реально предлагается лишь несколько коммерческих продуктов, реализующих технологию вывода, основанного на прецедентах. Это объясняется, в первую очередь, сложностью алгоритмов и их эффективной программной реализации. Наиболее успешные и известные из присутствующих на рынке продуктов – CBR Express и Case Point (Inference Corp.), Apriori (Answer Systems), DP Umbrella (VYCOR Corp.) [Althof 96].

CBR Express и CasePoint – продукты, предназначенные для разработки экспертных систем, основанных на прецедентах. CBR Express тоже накапливает "опыт", обеспечивая ввод, сопровождение и динамическое добавление прецедентов, а также простой доступ к ним с помощью вопросов и ответов. Обе системы используются при автоматизации информационно-справочных служб и "горячих линий", а также при создании интеллектуальных программных продуктов, систем доступа к информации, систем публикации знаний и т. д.

При общении с системой сначала вводится простой запрос, например: "Мой компьютер не работает". Далее происходит выделение ключевых слов, поиск в базе прецедентов, и генерируется перечень потенциальных решений. Пользователю могут быть также заданы уточняющие вопросы. Предлагаемые варианты решения проблемы могут включать в себя видео- или фотоматериалы. Технология вывода по прецедентам представляет собой основу для практически безграничных приложений, которые наращиваются за счет постоянного сбора информации (причем обеспечивается совмещение структурированных и неструктурированных данных, включая мультимедиа). По мнению компаний, активно использующих эту технологию, таких как Nippon Steel, Lockheed и некоторых других, создается самообучающаяся коллективная память, исключительно удобная для накопления и передачи профессионального опыта.

3. Добыча данных в системах поддержки принятия решений и прогнозирования

Русскоязычному термину "добыча данных" или "раскопка данных" в английском языке соответствует термин Data Mining. Нередко встречаются слова "обнаружение знаний в базах данных" (Knowledge Discovery in Databases) и "интеллектуальный анализ данных" (ИАД). Возникновение всех указанных терминов связано с новым витком в развитии средств и методов обработки данных. Цель добычи данных состоит в выявлении скрытых правил и закономерностей в наборах данных. Дело в том, что человеческий разум сам по себе не приспособлен для восприятия больших массивов разнородной информации. Человек обычно не способен улавливать более двух-трех взаимосвязей даже в небольших выборках. Но и традиционная математическая статистика, долгое время претендовавшая на роль основного инструмента анализа данных, также нередко пасует при решении задач из реальной сложной жизни. Она оперирует усредненными характеристиками выборки, которые

часто являются фиктивными величинами (типа средней температуры пациентов по больнице, средней высоты дома на улице, состоящей из дворцов и лачуг и т. п.). Поэтому методы математической статистики оказываются полезными главным образом для проверки заранее сформулированных гипотез и для "грубого" разведочного анализа, составляющего основу оперативной аналитической обработки данных (online analytical processing – OLAP).

В основу современной технологии добычи данных положена концепция шаблонов (паттернов), отражающих фрагменты многоаспектных взаимоотношений в данных. Эти шаблоны представляют собой закономерности, свойственные подвыборкам данных, которые могут быть компактно выражены в понятной человеку форме. Поиск шаблонов производится методами, не ограниченными рамками априорных предположений о структуре выборки и виде распределений значений анализируемых показателей.

К задачам, использующим методы добычи данных, обычно относятся задачи, при решении которых требуется получить ответы, например, на следующие вопросы:

1. Какие факторы лучше всего предсказывают несчастные случаи (встречаются ли точные шаблоны в описаниях людей, подверженных повышенному травматизму)?
2. Какие характеристики отличают клиентов, которые, по всей вероятности, собираются отказаться от услуг телефонной компании?
3. Какие схемы покупок характерны для мошенничества с кредитными карточками?

Важное положение – нетривиальность разыскиваемых шаблонов. Это означает, что найденные шаблоны должны отражать неочевидные, неожиданные регулярности в данных, составляющие так называемые скрытые знания. К обществу пришло понимание того, что сырые (первичные) данные содержат глубинный пласт знаний, при грамотной раскопке которого могут быть обнаружены настоящие самородки.

В целом технологию добычи данных достаточно точно определяет Григорий Пиатецкий-Шапиро [Fayyad 96] – один из основателей этого направления. Добыча данных – это процесс обнаружения в сырых данных:

- ранее неизвестных;
- нетривиальных;
- практически полезных;
- доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности.

3.1. Различные подходы к классификации области добычи данных

Существуют различные подходы к классификации самой области добычи данных. В частности, выделяют "*дескриптивный*" и "*предиктивный*" подход. Отличие этих двух подходов друг от друга заключается в том, что в первом случае мы обнаруживаем знания описательного характера, а во втором – знания, которые можно использовать для прогноза. В литературе встречаются также термины "*структурный*" и "*неструктурный*" подход. Такая терминология исходит из того, что предметом анализа могут быть как хорошо структурированные данные, например, таблицы в реляционной базе, так и неструктурированные данные, например, текст или изображения. "*Исторический*" подход основывается на том, что существует целый ряд исторически сложившихся дисциплин, например, теория нейронных сетей или нечеткая логика, перечисляя которые, определяют предмет и границы области добычи данных. Мы, следуя уже сложившейся в литературе традиции, будем придерживаться более формального подхода, а именно, называть все технологии добычи данных машинным обучением и делить их на две большие группы: "*управляемое*" и "*неуправляемое*" обучение.

В первом случае – "обучении с учителем" – задача анализа данных, например, классификация, осуществляется в несколько этапов. Сначала с помощью какого-либо алгоритма строится модель анализируемых данных – классификатор. Затем, этот классификатор подвергается "обучению". Другими словами, проверяется качество его работы и, если оно неудовлетворительно, происходит "дополнительное обучение" классификатора. Этот процесс повторяется до тех пор, пока не будет достигнут требуемый уровень качества и не возникнет убеждение, что выбранный алгоритм работает корректно с данными, либо же сами данные не имеют структуры, которую можно выявить.

Неуправляемое обучение объединяет технологии, выявляющие дескриптивные модели, например, закономерности, проявляющиеся при покупках, совершаемых клиентами большого магазина. Очевидно, что если эти закономерности есть, то модель должна их представить, и неуместно говорить об ее обучении. Отсюда и название – неуправляемое обучение. Дальнейшая классификация технологий добычи данных опирается на то, какие задачи этими технологиями решаются. Управляемое обучение, таким образом, подразделяется на классификацию и регрессию, а неуправляемое обучение – на анализ рыночных корзин, анализ временных последовательностей (секвенциальный анализ) и кластеризацию.

3.2. Классификация задач добычи данных

Целью технологии добычи данных является производство нового знания, которое пользователь может в дальнейшем применить для улучшения результатов своей деятельности. Рассмотрим основные виды моделей, которые используются для нахождения нового знания. Результат моделирования – это

выявленные отношения в данных. Можно выделить, по крайней мере, семь методов выявления и анализа знаний:

1. классификация,
2. регрессия,
3. кластеризация,
4. анализ ассоциаций,
5. прогнозирование временных последовательностей (рядов),
6. агрегирование (обобщение),
7. обнаружение отклонений.

Методы 1, 2 и 4 используются, главным образом, для предсказания, в то время как остальные удобны для описания существующих закономерностей в данных.

Вероятно, наиболее распространенной сегодня операцией интеллектуального анализа данных является *классификация*. С ее помощью выявляются признаки, характеризующие группу, к которой принадлежит тот или иной объект. Это делается посредством анализа уже классифицированных объектов и формулирования некоторого набора правил. Во многих видах бизнеса болезненной проблемой считается потеря постоянных клиентов. В разных сферах (таких, как сотовая телефонная связь, фармацевтический бизнес или деятельность, связанная с кредитными карточками) ее обозначают различными терминами – "переменой моды", "истощением спроса" или "покупательской изменой", – но суть при этом одна. Классификация помогает выявить характеристики "неустойчивых" покупателей и создает модель, способную предсказать, кто именно склонен уйти к другому поставщику. Используя ее, можно определить самые эффективные виды скидок и других выгодных предложений, которые будут наиболее действенны для тех или иных типов покупателей. Благодаря этому удается удержать клиентов, потратив ровно столько денег, сколько необходимо, и не более.

Однажды определенный эффективный классификатор используется для классификации новых записей в базе данных в уже существующие классы, и в этом случае он приобретает характер прогноза. Например, классификатор, который умеет идентифицировать риск выдачи займа, может быть использован для целей принятия решения, велик ли риск предоставления займа определенному клиенту. То есть классификатор используется для прогнозирования возможности возврата займа.

Классическим примером применения классификации на практике является решение проблемы о возможной некредитоспособности клиентов банка. Этот вопрос, тревожащий любого сотрудника кредитного отдела банка, можно, конечно, разрешить интуитивно. Если образ клиента в сознании банковского служащего соответствует его представлению о кредитоспособном клиенте, то кредит выдавать можно, иначе – отказать. По схожей схеме работают установленные в тысячах американских банках системы добычи данных. Лишенные субъективной предвзятости, они опираются в своей работе только

на историческую базу данных банка, где записывается детальная информация о каждом клиенте и, в конечном итоге, факт его кредитоспособности (вернул клиент ранее выданный кредит или нет). Клиенты банка в этих системах интерпретируются как векторы в пространстве $\Re^d$, атрибутам которых соответствуют различные данные о клиентах (возраст, годовой доход, профессия и т. д.). Факт возврата (невозврата) кредита играет роль функции $y_i = \{+1, -1\}$. Часть упомянутой исторической базы можно рассматривать как тренировочный набор данных, а другую часть – как обучающий набор. При таком подходе задача определения риска, связанного с клиентами, сводится к задаче построения классификатора. Решить ее можно с помощью нескольких методик. Также очевидно, что она имеет решение, поскольку интуитивно понятно, какого рода закономерности в данных обуславливают риск, связанный с клиентом. Высокий доход и хорошая профессия, безусловно, хороший аргумент в пользу благонадежности клиента.

В качестве методов решения задачи классификации могут использоваться алгоритмы типа Lazy-Learning [Wettschereck; Wang 99], в том числе известные алгоритмы ближайшего соседа и k-ближайших соседей [Aha 93, Aha 95], байесовские сети [Brand 98/2, Heckerman 95, Heckerman 97], деревья решений[Brand 98/1, Breiman 84, Quinlan 87, Quinlan 93, Гупал 93, Цветков 93], индукция символьных правил [Fuernkranz 96, Parsaye 97], нейронные сети [Уоссермен 92].

Регрессионный анализ используется в том случае, если отношения между переменными могут быть выражены количественно в виде некоторой комбинации этих переменных. Полученная комбинация далее используется для предсказания значения, которое может принимать целевая (зависимая) переменная, вычисляемая на заданном наборе значений входных (независимых) переменных. В простейшем случае для этого используются стандартные статистические методы, такие как линейная регрессия. К сожалению, большинство реальных моделей не укладываются в рамки линейной регрессии. Например, размеры продаж или фондовые цены очень сложны для предсказания, потому что могут зависеть от комплекса взаимоотношений множества переменных. Таким образом, необходимы комплексные методы для предсказания будущих значений.

Основные проблемы, с которыми сталкиваются при решении задач классификации и регрессии – это неудовлетворительное качество исходных данных, в которых встречается как шум, так и пропущенные значения, различные типы атрибутов – числовые и категорические, разная значимость атрибутов, а также, так называемые, проблемы "*overfitting*" и "*underfitting*". Суть первой из них, заключается в том, что классификационная функция при построении "чересчур хорошо" адаптируется к данным. И встречающийся в данных шум, и аномальные значения эта функция пытается интерпретировать как часть внутренней структуры данных. Очевидно, что такой классификатор будет некорректно работать в дальнейшем с другими данными, где характер шума будет несколько иной. Термином "*underfitting*" обозначают ситуацию,

когда слишком велико количество ошибок при проверке классификатора на обучающем множестве. Это означает, что особых закономерности в данных не было обнаружено и, либо их нет вообще, либо необходимо выбрать иной метод их обнаружения.

Кластеризация логически продолжает идею классификации на более сложный случай, когда сами классы не предопределены. Результатом использования метода, выполняющего кластеризацию, как раз является определение (посредством свободного поиска) присущего исследуемым данным разбиения на группы.

Так можно выделить родственные группы клиентов или покупателей с тем, чтобы вести в их отношении дифференцированную политику. Например, "группы риска" – категории клиентов, готовых уйти к другому поставщику – средствами кластеризации могут быть определены до начала процесса ухода, что позволит производить профилактику проблемы, а не экстренное исправление положения. В большинстве случаев кластеризация очень субъективна: любой вариант разбиения на кластеры напрямую зависит от выбранной меры расстояния между объектами.

Для научных исследований изучение результатов кластеризации, а именно, выяснение причин, по которым объекты объединяются в группы, способно открыть новые перспективные направления. Традиционным примером, который обычно приводят для этого случая, является периодическая таблица элементов. В 1869 году Дмитрий Менделеев разделил 60 известных в то время элементов, на кластеры или периоды. Элементы, попадавшие в одну группу, обладали схожими характеристиками. Изучение причин, по которым элементы разбивались на явно выраженные кластеры, в значительной степени, определило приоритеты научных изысканий на годы вперед. Но лишь спустя пятьдесят лет квантовая физика дала убедительные объяснения периодической системы.

Кластеризация в чем-то аналогична классификации, но отличается от нее тем, что для проведения анализа не требуется иметь выделенную целевую переменную. Ее удобно использовать на начальных этапах исследования, когда о данных мало что известно. В большинстве других методов добычи данных исследование начинается, когда данные уже предварительно как-то расклассифицированы, хотя бы на обучающее множество данных и данные, по которым проверяется найденная модель или для которых надо предсказать целевую переменную. Для этапа кластеризации характерно отсутствие каких-либо различий, как между переменными, так и между записями. Напротив, ищутся группы наиболее близких, похожих записей.

Техника кластеризации применяется в самых разнообразных областях. Хартиган [Hartigan 1975] дал обзор многих опубликованных исследований, содержащих результаты, полученные методами кластерного анализа. Например, в области медицины кластеризация заболеваний, лечения заболеваний или симптомов заболеваний приводит к широко используемым таксономиям. В области психиатрии правильная диагностика кластеров

симптомов, таких как паранойя, шизофрения и т. д., является решающей для успешной терапии. В археологии с помощью кластерного анализа исследователи пытаются установить таксономии каменных орудий, похоронных объектов и т.д. Известны широкие применения кластерного анализа в маркетинговых исследованиях. В общем, всякий раз, когда необходимо классифицировать "горы" информации к пригодным для дальнейшей обработки группам, кластерный анализ оказывается весьма полезным и эффективным.

Существует целый ряд алгоритмов кластеризации, позволяющих обнаруживать кластеры данных с любой степенью точности. Наиболее распространенные алгоритмы – это иерархическая кластеризация [Johnson 67, Gruvaeus 72] и метод *k-средних* [Hartigan 75, Hartigan 78]. В качестве примера других используемых методов можно привести обучение "без учителя" особого вида нейронных сетей – сетей Кохонена [Уоссермен 92], а также индукцию правил [Fuernkranz 96].

Выявление ассоциаций (другие названия: *поиск ассоциативных правил*, *анализ рыночных корзин*). Ассоциация имеет место в том случае, если несколько событий связаны друг с другом. Например, исследование "покупательской корзины", проведенное в супермаркете, может показать, что 65 % купивших кукурузные чипсы берут также и "кока-колу", а при наличии скидки за такой комплект "колу" приобретают в 85 % случаев. Располагая сведениями о подобной ассоциации, менеджерам легко оценить, насколько действенна предоставляемая скидка.

Кроме обширных практических приложений в области маркетинга, эта задача представляется важной и в ряде других приложений, связанных с объединением данных из различных источников. В частности, результаты анализа ассоциаций позволяют получать паттерны типа ассоциативных правил, которые далее могут использоваться для формирования продукционных баз знаний в системах поддержки принятия решений, обнаружения причин отказов аппаратуры, причин негативного или, наоборот, позитивного развития событий, ситуаций и т.п. Например, анализ ассоциаций, зависящих от времени, в последовательности событий входящего трафика компьютерной сети является основным источником информации для различения нормальной и аномальной деятельности пользователей.

Прогнозирование временных последовательностей (*секвенциальный анализ*) есть установление закономерностей между связанными во времени событиями. Метод позволяет на основе анализа поведения временных рядов оценить будущие значения прогнозируемых переменных. Конечно, эти модели должны включать в себя особые свойства времени: иерархию периодов (декада-месяц-год или месяц-квартал-год), особые отрезки времени (пяти-, шести- или семидневная рабочая неделя, тринадцатый месяц), сезонность, праздники и др.

Анализ рыночных корзин (*Basket Analysis*) и секвенциальный анализ являются в настоящий момент одними из самых популярных приложений добычи данных.

Агрегированием (обобщением) называют задачу поиска компактного описания подмножества данных. Примерами могут служить задача отыскания вектора средних значений и матрицы отклонений для набора данных, поиск функциональных зависимостей между переменными или ассоциативных правил и другие задачи. Поиск агрегированных описаний интерпретируется часто как поиск другого, в каком-то смысле лучшего, пространства представления данных. Типичным примером такого преобразования пространства представления данных является замена описания данных в терминах первичных атрибутов описанием их в терминах так называемых "аргументов" в пользу того или иного решения [Aha 95/1, Bundy 97], истинностные значения которых на конкретных входных данных затем используются для их классификации [Bull 97].

Обнаружение отклонений. Целью задачи является поиск наиболее значимых в заданном смысле изменений в данных по сравнению со средними, нормативными показателями.

3.3. Классификация систем добычи данных

Добыча данных является мультидисциплинарной областью, возникшей и развивающейся на базе достижений прикладной статистики, распознавания образов, методов искусственного интеллекта, теории баз данных и др. Отсюда обилие методов и алгоритмов, реализованных в различных действующих системах добычи данных. Многие из таких систем интегрируют в себе сразу несколько подходов. Тем не менее, как правило, в каждой системе имеется какой-то ключевой компонент, на который делается главная ставка. Ниже приводится классификация указанных ключевых компонентов на основе работ [Киселев 97, Дюк 01]:

- статистические методы;
- нейронные сети;
- деревья решений;
- системы рассуждения на основе аналогичных случаев;
- нечеткая логика;
- генетические алгоритмы;
- эволюционное программирование;
- алгоритмы ограниченного перебора;
- комбинированные методы.

Несмотря на то, что последние версии почти всех известных *статистических пакетов* включают наряду с традиционными статистическими методами также элементы добычи данных, основное внимание в них уделяется все же классическим методикам: корреляционному, регрессионному, факторному анализу и другим. Детальный обзор пакетов для статистического анализа приведен в [HTTP/1] Недостатком систем этого класса считается требование к специальной подготовке пользователя. Еще более серьезным принципиальным недостатком статистических пакетов, ограничивающим их применение в

добыче данных, является то, что большинство методов, входящих в состав пакетов, опираются на статистическую парадигму, в которой главными фигурантами служат усредненные характеристики выборки. А они при исследовании реальных сложных жизненных феноменов часто являются фиктивными величинами.

Нейронные сети относятся к классу нелинейных адаптивных систем с архитектурой, условно имитирующей нервную ткань из нейронов. Математическая модель нейрона представляет собой некоторый универсальный нелинейный элемент с возможностью широкого изменения и настройки его характеристик. В одной из наиболее распространенных архитектур, многослойном перцептроне с обратным распространением ошибки, эмулируется работа нейронов в составе иерархической сети, где каждый нейрон более высокого уровня соединен своими входами с выходами нейронов нижележащего слоя. На нейроны самого нижнего слоя подаются значения входных параметров, на основе которых нужно принимать какие-то решения, прогнозировать развитие ситуации и т. д. Эти значения рассматриваются как сигналы, передающиеся в вышележащий слой, ослабляясь или усиливаясь в зависимости от числовых значений (весов), приписываемых межнейронным связям. В результате на выходе нейрона самого верхнего слоя вырабатывается некоторое значение, которое рассматривается как ответ, реакция всей сети на введенные значения входных параметров. Для того чтобы сеть можно было применять в дальнейшем, ее прежде надо "натренировать" на полученных ранее данных, для которых известны и значения входных параметров, и правильные ответы на них. Эта тренировка состоит в подборе весов межнейронных связей, обеспечивающих наибольшую близость ответов сети к известным правильным ответам.

Основным недостатком нейросетевой парадигмы является необходимость иметь очень большой объем обучающей выборки. Другой существенный недостаток заключается в том, что даже натренированная нейронная сеть представляет собой черный ящик. Знания, зафиксированные как веса нескольких сотен межнейронных связей, совершенно не поддаются анализу и интерпретации человеком, а известные попытки дать интерпретацию структуре настроенной нейросети выглядят неубедительными.

Деревья решений являются одним из наиболее популярных подходов к решению задач добычи данных. Они создают иерархическую структуру классифицирующих правил типа "*если...то...*", имеющую вид дерева. Для того чтобы решить, к какому классу отнести некоторый объект или ситуацию, требуется ответить на вопросы, стоящие в узлах этого дерева, начиная с его корня. Вопросы имеют вид "значение параметра *A* больше *x*?". Если ответ положительный, осуществляется переход к правому узлу следующего уровня, если отрицательный – то к левому узлу; затем снова следует вопрос, связанный с соответствующим узлом. В результате мы добираемся до одного из возможных вариантов решения.

Популярность деревьев решений связана с наглядностью и понятностью. Но для них очень остро стоит проблема значимости. Дело в том, что отдельным узлам на каждом новом построенном уровне дерева соответствует все меньшее и меньшее число записей данных – дерево дробит данные на большое количество частных случаев. Чем их больше, чем меньше обучающих примеров попадает в каждый такой частный случай, тем менее уверенной будет их классификация. Если построенное дерево слишком "кустистое" – состоит из неоправданно большого числа мелких веточек – оно не будет давать статистически обоснованных ответов. Как показывает практика, в большинстве систем, использующих деревья решений, эта проблема не находит удовлетворительного решения. Кроме того, общеизвестно, и это легко показать, что деревья решений дают полезные результаты только в случае независимых признаков. В противном случае они лишь создают иллюзию логического вывода.

Деревья решений принципиально не способны находить "лучшие" правила в данных. Они реализуют наивный принцип последовательного просмотра признаков и "цепляют" фактически осколки настоящих закономерностей, создавая лишь иллюзию логического вывода. Вместе с тем, большинство систем используют именно этот метод.

Нечеткая логика применяется для таких наборов данных, где причисление данных к какой-либо группе является вероятностью, находящейся в интервале от 0 до 1, но не принимающей крайние значения. Четкая логика манипулирует результатами, которые могут быть либо истиной, либо ложью. Нечеткая логика применяется в тех случаях, когда необходимо манипулировать степенью "может быть" в дополнении к "да" и "нет".

Добыча данных – далеко не основная область применения *генетических алгоритмов*, которые, скорее, нужно рассматривать в качестве мощного средства решения разнообразных комбинаторных задач и задач оптимизации. Тем не менее, генетические алгоритмы вошли сейчас в стандартный инструментарий методов добычи данных. Этот метод назван так потому, что в какой-то степени имитирует процесс естественного отбора в природе.

Предположим, требуется найти решение задачи, наиболее оптимальное с точки зрения некоторого критерия. Пусть каждое решение полностью описывается некоторым набором чисел или величин нечисловой природы. Если необходимо выбрать совокупность фиксированного числа параметров рынка, наиболее выразительным образом влияющих на его динамику, это будет набор имен этих параметров. О нем можно говорить как о совокупности хромосом, определяющих качества индивида – конкретного решения поставленной задачи. Значения параметров, определяющих решение, будут в этом случае называться генами. Поиск оптимального решения при этом похож на эволюцию популяции индивидов, представленных их наборами хромосом. В этой эволюции действуют три механизма: во-первых, отбор сильнейших, то есть тех наборов хромосом, которым соответствуют наиболее оптимальные решения; во-вторых, скрещивание – производство новых индивидов при

помощи смешивания хромосомных наборов отобранных индивидов; и, в-третьих, мутации – случайные изменения генов у некоторых индивидов популяции. В результате смены поколений вырабатывается такое решение поставленной задачи, которое не может быть далее улучшено.

Генетические алгоритмы удобны тем, что их легко распараллеливать. Например, можно разбить поколение на несколько групп и работать с каждой из них независимо, обмениваясь, время от времени несколькими хромосомами. Существуют также и другие методы распараллеливания генетических алгоритмов.

Генетические алгоритмы имеют ряд недостатков. Критерий отбора хромосом и используемые процедуры являются эвристическими и далеко не гарантируют нахождения "лучшего" решения. Как и в реальной жизни, эволюцию может "заклинить" на какой-либо непродуктивной ветви. И, наоборот, можно привести примеры, как два неперспективных родителя, которые будут исключены из эволюции генетическим алгоритмом, оказываются способными произвести высокоэффективного потомка. Это особенно становится заметно при решении высокоразмерных задач со сложными внутренними связями.

Сама постановка задачи в терминах генетических алгоритмов не дает возможности проанализировать статистическую значимость получаемого с их помощью решения. Кроме того, эффективно сформулировать задачу, определить критерий отбора хромосом под силу только специалисту. В силу этих факторов сегодня генетические алгоритмы надо рассматривать скорее как инструмент научного исследования, чем как средство анализа данных для практического применения в бизнесе и финансах.

Эволюционное программирование – сегодня самая молодая и наиболее перспективная ветвь добычи данных. Суть метода заключается в том, что гипотезы о виде зависимости целевой переменной от других переменных формулируются системой в виде программ на некотором внутреннем языке программирования. Если это универсальный язык, то теоретически на нем можно выразить зависимость любого вида. Процесс построения этих программ строится подобно эволюции в мире программ (этим метод похож на генетические алгоритмы). Когда система находит программу, достаточно точно выражающую искомую зависимость, она начинает вносить в нее небольшие модификации и отбирает среди построенных таким образом дочерних программ те, которые повышают точность. Таким образом, система "выращивает" несколько генетических линий программ, которые конкурируют между собой в точности выражения искомой зависимости. Специальный транслирующий модуль переводит найденные зависимости с внутреннего языка системы на понятный пользователю язык (математические формулы, таблицы и пр.), делая их легкодоступными. Для того чтобы сделать полученные результаты еще понятнее для пользователя-нематематика, имеется богатый арсенал разнообразных средств визуализации обнаруживаемых зависимостей.

Поиск зависимости целевых переменных от остальных ведется в форме функций какого-то определенного вида. Например, в одном из наиболее удачных алгоритмов этого типа – методе группового учета аргументов (МГУА) зависимость ищут в форме полиномов. Причем сложные полиномы заменяются несколькими более простыми, учитывающими только некоторые признаки (групп аргументов). Обычно для этого используются попарные объединения признаков.

Алгоритмы ограниченного перебора были предложены в середине 60-х годов М. М. Бонгардом [Айвазян 89] для поиска логических закономерностей в данных. С тех пор они продемонстрировали свою эффективность при решении множества задач из самых различных областей.

Эти алгоритмы вычисляют частоты комбинаций простых логических событий в подгруппах данных. Примеры простых логических событий: $X = a$; $X < a$; $X > a$; $a < X < b$ и др., где X — какой либо параметр, “a” и “b” — константы. Ограничением служит длина комбинации простых логических событий (у М. Бонгарда она была равна 3). На основании анализа вычисленных частот делается заключение о полезности той или иной комбинации для установления ассоциации в данных, для классификации, прогнозирования и пр.

Наиболее ярким современным представителем этого подхода является система WizWhy предприятия WizSoft [HTTP/2]. Хотя автор системы Абрахам Мейдан не раскрывает специфику алгоритма, положенного в основу работы WizWhy, по результатам тщательного тестирования системы были сделаны выводы о наличии здесь ограниченного перебора (изучались результаты, зависимости времени их получения от числа анализируемых параметров и др.).

Автор WizWhy утверждает, что его система обнаруживает *все* логические правила вида "*если…то…*" для поступающих данных. На самом деле это, конечно, не так. Во-первых, максимальная длина комбинации в правиле "*если…то…*" в системе WizWhy равна 6, и, во-вторых, с самого начала работы алгоритма производится эвристический поиск простых логических событий, на которых потом строится весь дальнейший анализ. Поняв эти особенности WizWhy, нетрудно было предложить простейшую тестовую задачу, которую система не смогла вообще решить. Другой момент — система выдает решение за приемлемое время только для сравнительно небольшой размерности данных (не более 20).

Тем не менее, система WizWhy является на сегодняшний день одним из лидеров на рынке продуктов добычи данных, что совсем не лишено оснований. Система постоянно демонстрирует более высокие показатели при решении практических задач, чем все остальные алгоритмы.

Комбинированные методы. Часто производители сочетают указанные подходы. Объединение в себе средств нейронных сетей и технологии деревьев решений должно способствовать построению более точной модели и повышению ее быстродействия. Программы визуализации данных в каком-то смысле не являются средством анализа информации, поскольку они только представляют ее пользователю. Тем не менее, визуальное представление,

скажем, сразу четырех переменных достаточно выразительно обобщает очень большие объемы данных.

Для того чтобы найти новое знание на основе данных большого хранилища, недостаточно просто взять алгоритмы добычи данных, запустить их и ждать появления интересных результатов. Нахождение нового знания – это процесс, который включает в себя несколько шагов, каждый из которых необходим для уверенности в эффективном применении средств добычи данных.

4. Интегрированный подход к построению систем поддержки принятия решений

4.1. Два подхода к интеграции вывода на основе прецедентов и добычи данных

Тому, что вывод по прецедентам – не только парадигма, но и равноправный партнер добычи данных, когда оба метода могут использовать результаты работы друг друга, до сих пор уделялось небольшое внимание, хотя и было признано важным [Fayyad 96].

Какова мотивация для интеграции двух методов? Оба используются для обработки информации в целях улучшения качества решений, однако, используя интегрированный подход, можно, по-видимому, получить большую отдачу от информации, чем, используя любой из методов в отдельности. Сочетание двух методов позволяет сформулировать и реализовать на практике принципиально новый подход к построению интеллектуальных систем. Можно привести слова математика Сеймура Паперта: "Некоторые из наиболее серьезных шагов в умственном развитии человечества основаны не просто на приобретении новых знаний, а на приобретении новых административных способов использовать то, что каждый уже знает".

Вывод по прецедентам сильно зависит от качества и количества собранных данных, от знаний о проблемной области и способов отбора наиболее релевантных прецедентов. Метод больше подходит для областей, о которых мы имеем недостаточно знаний.

В свою очередь, некоторые алгоритмы добычи данных сами требуют фонового знания, которое может быть получено с помощью прецедентов.

Вывод по прецедентам и добыча данных могут быть интегрированы несколькими способами. В зависимости от этого один из методов можно рассматривать как главный (*master*), а другой – в качестве вспомогательного (*slave*).

4.2. Использование методов добычи данных в системах вывода по прецедентам

Добыча данных позволяет находить дополнительные знания о проблемной области в виде паттернов, которые могут использоваться как фоновое знание в выводе по прецедентам:

- для вычисления степени близости между прецедентами (одним из таких способов является разбиение прецедентов на классы эквивалентности, когда близкими текущему случаю считаются прецеденты того же класса),
- для получения дополнительных знаний из базы прецедентов, что позволяет, например, выявлять значимость признаков и заполнять отсутствующие признаки,
- при адаптации решения,
- и даже при добавлении прецедентов (добыча данных может помочь найти дополнительные знания в базе данных и представить это как сконструированный прецедент).

4.3. Использование прецедентов в системах добычи данных

Учитывая, что процесс добычи данных может быть затратным, информация о достигнутых результатах и о процессе в целом может быть сохранена в виде прецедента, чтобы не тратить время на выработку одних и те же паттернов. Потребность в таком подходе впервые была озвучена в рамках обсуждения проекта CRISP-DM [Anand 97/1] при попытке выработать стандартную модель процесса добычи данных. В ходе него было заявлено: "Стандартная методология добычи данных должна обеспечить возможность фиксации и многократного использования опытов, а также управления проектами".

Прецеденты могут использоваться для нахождения некоторого фонового знания в базе данных, например, весов признаков для классификатора. В байесовской сети структура сети может быть изначально установлена с помощью "экспертного знания" (на основе прецедентов), а параметры уточнены с помощью алгоритмов добычи данных.

Прецеденты могут также использоваться, чтобы обеспечить утилитарность, критический анализ (обоснованность) и проверку новизны для алгоритмов добычи данных.

5. Использование методов добычи данных для отбора прецедентов

На примере различных систем можно увидеть, что интеграция является не только возможной, но и заслуживающей внимания. В настоящее время имеются несколько интегрированных систем, в том числе и на стадии разработки. В основном – это системы вывода по прецедентам, использующие методы добычи данных для работы с прецедентами.

Большая часть существующих подходов к использованию методов добычи данных в системах вывода по прецедентам сосредоточена на одном аспекте такого использования: выборе наиболее релевантных прецедентов. Здесь применяются различные методы добычи данных, среди них – деревья решений, байесовские сети, нейронные сети, методы k-ближайших соседей, и

т.д. Все они предлагают тот или иной способ измерения степени близости прецедента и текущего случая по их признакам.

Приведем два способа оценить близость прецедентов. Первый – статистический, где для отбора прецедентов используется байесовская сеть. Второй способ – введение классов эквивалентности на множестве прецедентов.

5.1. Байесовские сети

Байесовские сети (Bayesian Networks) – это статистический метод описания закономерностей в данных. На основе первичной информации, содержащейся в базах данных, строится модель в виде сети, где множество вершин описывает события, а ребра интерпретируются как причинные связи между событиями.

В основе байесовских сетей лежит теорема Байеса теории вероятностей для определения апостериорных вероятностей попарно несовместных событий Y_i по их априорным вероятностям:

$$P(Y_i|X) = \frac{P(Y_i) \cdot P(X|Y_i)}{P(X)}$$

Всякое множество ребер, представляющее собой все пути между некоторыми двумя вершинами, соответствует условной зависимости между этими вершинами. Если задать некоторое распределение вероятностей на множестве переменных, соответствующих вершинам этого графа, то полученная сеть будет называться *байесовской сетью*. На такой сети можно использовать, так называемый байесовский вывод для вычисления вероятностей следствий событий.

Критерий отбора прецедентов заключается в следующем. Если нет полностью совпадающего прецедента, вычисляется распределение вероятностей по тем признакам, которые не совпадают с признаками текущего случая. Выбирается тот прецедент, для которого эта вероятность наибольшая.

Существуют два способа обучения байесовских сетей с помощью прецедентов: уточнение параметров сети, если структура сети известна, и выбор из множества моделей, применяя введенную метрику ко всей базе прецедентов.

Экерман [Heckerman 97] отмечает четыре достоинства байесовских сетей как средства извлечения данных:

- поскольку в модели определяются зависимости между всеми переменными, легко обрабатываются ситуации, когда значения некоторых переменных неизвестны;
- построенные байесовские сети просто интерпретируются и позволяют на этапе прогностического моделирования легко производить анализ по сценарию "*что если*…";
- подход позволяет естественным образом совмещать закономерности, выведенные из данных, и фоновые знания, полученные в явном виде, например, от экспертов;

- использование байесовских сетей позволяет избежать проблемы переподгонки (overfitting), то есть избыточного усложнения модели, чем страдают многие методы (например, деревья решений и индукция правил) при слишком буквальном следовании распределению зашумленных данных.

Несмотря на свою простоту, скорость и интерпретируемость результатов, наивно-байесовский алгоритм имеет недостатки:

- перемножать условные вероятности корректно только тогда, когда все входные переменные действительно статистически независимы; допущение этой независимости и обуславливает уточнение "*наивно*-" в названии алгоритма, хотя, по приведенным в [Brand 98/2] примерам он показывает неплохие практические результаты даже при несоблюдении условия статистической независимости; корректно данная ситуация обрабатывается только более сложными методами, основанными на обучении байесовских сетей [Heckerman 95, Heckerman 97];
- невозможна непосредственная обработка непрерывных переменных – их требуется разбивать на множество интервалов, чтобы атрибуты были дискретными; такое разбиение в ряде случаев приводит к потере значимых закономерностей [Brand 98/2];
- наивно-байесовский подход учитывает только индивидуальное влияние входных переменных на результат классификации, не принимая во внимание комбинированного влияния пар или троек значений разных атрибутов [Brand 98/2], что было бы полезно с точки зрения прогностической точности, но значительно увеличило бы количество проверяемых комбинаций.

Байесовские сети активно использовались для формализации знаний экспертов в экспертных системах [Heckerman 95], но с недавних пор стали применяться для извлечения знаний из наборов данных. Приведем несколько примеров систем, в которых используется интеграция байесовских сетей и вывод по прецедентам.

Компания Microsoft разработала прототип системы для диагностики неисправностей с кодовым именем ALADDIN [Breese 95]. В системе используется трехуровневая байесовская сеть. Первый уровень описывает одну или несколько *причин* – факторов, приведших к сбою, второй – *результат,* который будет получен при наличии всех причин, и третий – *симптомы*, вызываемые результатом. Байесовская сеть конструируется экспертом и корректируется при каждом использовании. Microsoft прекратила использование системы в связи с малым объемом базы знаний.

В Университете Salford'a [Rodriguez 97], была разработана система, в которой одна байесовская сеть используется для индексации *категорий* – групп

прецедентов, объединенных по принципу общности свойств, а другая – для индексации *экземпляров*, то есть единичных прецедентов внутри категории.

INBANCA [Aha 96] – система, разработанная Центром Прикладных Исследований BMC США для принятия плана действий, адекватных текущей ситуации. Байесовская модель используется при описании окружающей среды.

D-SIDE [Tirri 96] – это программный пакет, разработанный в Университете Хельсинки. Здесь прецеденты рассматриваются как вектора. Допускается, что база может иметь некорректные прецеденты. Байесовская модель используется при адаптации решения прецедента к текущему случаю, в частности, для предсказания наиболее вероятных значений отсутствующих признаков.

Существующие системы используют разные подходы. Первые две используют добычу данных для выявления знаний о предметной области. INBANCA использует прецеденты для выбора плана действий, ALADDIN использует прецеденты для устранения ошибки, найденной с помощью байесовских сетей, система в Salford'e использует байесовские сети для манипулирования прецедентами, и, наконец, D-SIDE использует прецеденты для классификации.

5.2. Разбиение базы прецедентов на классы

Одним из способов введения меры близости между объектами является разбиение их на классы эквивалентности. Задать классы эквивалентности – значит разбить множество объектов на группы, внутри которых объекты считаются (в некотором смысле) равными. Считается, что классы соответствуют различным внутренним понятиям базы и, соответственно, предполагают различные решения проблем. Разбиение на кластеры можно считать частным случаем разбиения на классы, за одним исключением: в этом случае не требуется этап предварительного обучения.

Так, например, применение методов классификации (в частности, кластерного анализа) позволяет в области торговли недвижимостью предварительно разбить все объекты на классы (например, дворцы и бунгало) не только по стоимости, но и по характеру жилья. Внутри класса объекты могут отличаться в меньшей степени, например, по количеству спальных или ванных комнат, и могут ранжироваться по некоторым другим признакам.

В решении, предложенном авторами системы M^2 [Anand 97/2, Anand 98], используется предварительная кластеризация базы прецедентов. Кластеризация применяется в двух аспектах: сбор прецедентов и отыскание недостающих знаний при адаптации решения. В [Anand 98] подробно обсуждается подход к обнаружению прецедентов и в кратких чертах – методология адаптации решения.

В этой системе задачу кластеризации входных данных выполняет нейронная сеть Кохонена. При решении этой задачи образуются начальные кластеры, которые затем анализируются с использованием алгоритма построения дерева решений C4.5 [Quinlan 93]. Неуникальные кластеры группируются.

На последней стадии используется алгоритм индукции регрессионного дерева, чтобы гарантировать, что эти понятия информационно полны.

Основная идея заключается в том, что если текущий случай попадает в кластер, наиболее удачным аналогом для него считается центр этого кластера. Авторы показали, что предложенный подход достигает высокой редукции размера базы прецедентов.

Однако на практике не всегда удается четко разграничить кластеры, куда попадает текущий случай. Одной из причин этого является недостаток информации в описании текущего случая. Но главная причина заключается в том, что реальные приложения редко укладываются в рамки фиксированного признакового пространства. Попадание текущего случая в область пересечения кластеров в этом случае становится непреодолимой проблемой.

Так, в медицине разные наборы признаков (иными словами, показателей, симптомов) могут быть не только у разных заболеваний, но и у разных пациентов с одним и тем же заболеванием. И, наконец, пациент может иметь признаки, не совпадающие ни с одним из признаков заболеваний, ранее введенных в систему.

Для наглядности приведем пример из медицины. Текущий случай – это пациент, описываемый тремя признаками (симптомы острого живота):

1. боли в животе,
2. напряжение передних мышц брюшной стенки,
3. болезненная перкуссия по брюшной стенке.

В пространстве этих признаков точка, соответствующая текущему случаю, попадает в пересечение кластеров (заболеваний):

1. прободная язва желудка,
2. спонтанный разрыв пищевода,
3. перитонит,
4. базальная плевропневмония.

Разрешить эту проблему, иными словами, дифференцировать эти кластеры, можно только увеличив размерность пространства, добавив новые признаки для текущего случая, если такие найдутся. Последнее не всегда возможно.

5.3. Другие примеры систем, использующих интегрированный подход

Корпорацией NEC для управления общей корпоративной базой данных разработан Case-метод [Kitano 96, Leake 96]. Утверждается, что накопление данных в виде прецедентов, выявление закономерностей и последующее внедрение полученных знаний повышают общий профессиональный уровень сотрудников корпорации. В качестве алгоритмов добычи данных используются извлечение правил, статистические методы, и другие.

В Университете города Росток (ФРГ) разработана система для прогноза эпидемий [Bull 97]. В качестве прецедентов используются последовательности т.н. *сценариев*, в которые входят новые случаи заболеваний, нагрузка на

органы здравоохранения, контекстная информация о сезоне и погоде. Чтобы обнаружить "заметные" изменения в последовательностях сценариев, используется статистический метод добычи данных, называемый G-тест. Для сверки результата используются методы вывода по прецедентам.

6. Понятие контекстно-зависимой локальной метрики

Обычно в методе "ближайшего соседа" применяется простая евклидова метрика – сумма квадратов отклонений по разным параметрам. Это быстрый и часто неплохо работающий метод. Первый его минус заключается в том, что когда число анализируемых показателей, или количество полей записей, сравнимо с числом самих записей, получается пространство очень большого числа переменных с редким облачком точек. В этом случае соседство точек в терминах евклидовой метрики часто не означает естественной близости значений соответствующих записей, а в значительной степени обусловлено выбранным для анализа набором показателей. Когда же, как это довольно часто бывает, число параметров превышает число записей, облако точек становится настолько редким, что никаких разумных оценок этот метод, как правило, не дает. Другим слабым местом рассматриваемого метода, также как и у нейросетей, является удовлетворительный прогноз лишь достаточно непрерывных и гладких зависимостей.

Применение метода ближайших соседей приводит и к более глубоким проблемам. Например, если все независимые переменные имеют одну и ту же размерность, то есть, допустим, все измеряются в молях на литр (как, например, концентрации различных химических соединений в крови человека), то евклидова метрика имеет естественный смысл, понятна и адекватна. Но если одна из независимых переменных – это вес пациента, а вторая, скажем, его рост, непонятно, как соотнести разницу по одной оси в 1 кг с разницей в 1 см по другой оси. По существу, в этом случае пространство независимых переменных – это аффинное пространство, а не метрическое. Один из возможных способов преодоления этой трудности – нормирование всех независимых переменных на некоторое естественное значение этой переменной или характерный масштаб. Если естественные характерные значения переменных неизвестны (а это наиболее распространенный случай), каждую независимую переменную можно разделить на величину ее дисперсии. При этом дисперсии всех независимых переменных становятся равными единице, и это дает основания надеяться, что их изменения на одну и ту же величину сопоставимы между собой. Однако это предположение оправдано далеко не всегда.

Другая проблема возникает при работе с большими базами прецедентов, когда вероятность выбора близкого соответствия высока, а потребность в адаптации решения – низка, что ведет к дополнительным проблемам. Исследование в [Smyth 95] показало, что такие системы страдают от так называемого "заболачивания", которое происходит, когда стоимость поиска знания перевешивает выгоду от применения этого знания.

Традиционные методы анализа многомерных данных используют представление об общем пространстве признаков для всех объектов и об одинаковой мере, применяемой для оценки их сходства или различия. Такое представление уместно, например, при изучении однородных физических феноменов на статистическом уровне системной организации, в которых объект можно рассматривать как реализацию многомерной случайной величины с ясным физическим смыслом, когда есть все основания интерпретировать зафиксированные особенности объектов как случайные отклонения, обусловленные воздействием шумов, погрешностями измерительных приборов и т.п.

В задачах, которые можно объединить под общим названием "формирование знаний" (к ним относятся добыча данных и рассматриваемый нами метод вывода по прецедентам), каждый объект следует рассматривать как самостоятельный информационный факт (совокупность зафиксированных значений признаков), имеющий ценные уникальные особенности.

Эти особенности раскрываются путем конструирования собственного пространства признаков для любого объекта и нахождения индивидуальной меры его сходства с другими объектами. Без такого раскрытия описания объектов нивелированы, могут содержать много ненужных, шумящих, отвлекающих и даже вредных деталей.

Это, в свою очередь, требует знаний о предметной области, то есть сведений, выражающих закономерности, определяющие отношения между объектами из баз данных, в которых хранятся прецеденты.

Задачей методов добычи данных, которые включают в себя решение задач классификации, является не только поиск закономерностей, но и интерпретация этих закономерностей. Это позволяет сконструировать для каждого объекта индивидуальную локальную метрику, которая обеспечивает ему максимально возможную "сферу действия", которой нельзя достигнуть при построении общего пространства признаков и использовании одинаковой метрики для всех объектов.

Описание каждого эмпирического факта в этом случае оказывается полностью избавленным от неинформативных элементов, что позволяет в дальнейшем иметь дело с чистыми, "незашумленными" структурами данных. В этом описании остается только то, что действительно важно для отражения сходства и различия эмпирического факта с другими фактами в контексте решаемой задачи.

В свете представлений о локальных метриках, очевидно, что один и тот же объект может поворачиваться разными гранями своего многомерного описания сообразно заданному контексту. К любому объекту, запечатленному в памяти как целостная многомерная структура, может быть привязан набор различных локальных метрик, каждая из которых оптимизирует его сходства и различия с другими объектами соответственно целям определенной задачи отражения отношений между объектами.

В результате построения локальных метрик отношения между объектами выражаются матрицей удаленностей. Так как локальная метрика привязана к объекту, метрики разных объектов могут не совпадать, и для элементов матрицы могут не выполняться требования симметричности и неравенства треугольника. Поэтому данная матрица, хотя и отражает отношения различия между объектами, не может истолковываться как матрица расстояний.

Образно говоря, если взглянуть на множество объектов с точки, занимаемой объектом в пространстве, специально сконструированном для этого объекта, то для такого взора объекты выстроятся в специфический ряд по степени удаленности от данной точки. С другой точки и в другом пространстве ряд удаленностей тех же самых объектов будет иметь свой специфический вид.

Выбор конкретного преобразования зависит от того, на каком аспекте структуры данных исследователь решает сделать акцент. Например, может использоваться преобразование в ранговую величину. Выбор меры зависит, с одной стороны, от вида преобразования, с другой стороны – от того, какие особенности рядов и объектов имеется намерение оттенить при определении их сходства (различия).

В работах Дюка [Дюк 94, Дюк 96] предложен подход к конструированию собственного пространства признаков и нахождению индивидуальной меры, который назван *локальным преобразованием пространства признаков*. Это пространство образуется путем перехода к новой векторной переменной, например,

$$\Delta = X - X_i$$

где X_i – выбранный объект.

Подход Дюка заключается в комбинированном применении методов линейной алгебры и интерактивной графики. С одной стороны, алгебраическими методами ищется новая ось в локальном пространстве (весовой вектор), на которой распределение проекций объектов удовлетворяет заданному критерию (например, выражающему стремление сгруппировать около нулевой отметки объекты того же класса, что и у центрального объекта).

С другой стороны, так как интерес представляет только сравнительно небольшая область около нулевой отметки новой оси, удаленные от данной отметки объекты подвергаются исключению с использованием средств интерактивной графики. После каждого такого исключения параметры новой оси рассчитываются заново, и визуальный анализ полученного распределения дает основание для произведения еще одного акта исключения объектов, либо для останова процедуры поиска логической закономерности.

Здесь задача определения локальной метрики заключается в нахождении линейного преобразования новой векторной переменной. Для ее решения пригоден хорошо разработанный аппарат методов многомерного линейного анализа данных.

Другой вариант – преобразование в классификационный показатель. В этом случае ранг объектов по степени удаленности заменяется идентификатором своего класса, образно говоря, объекты "окрашиваются" в цвета своего класса. Но так как принадлежность классу – категориальная величина, все объекты, находящиеся в одном классе с рассматриваемым, будут считаться равными ему, а объекты других классов – нет. Локальная метрика для текущего объекта превращается в бинарную величину.

Как уже указывалось, особенности объекта раскрываются в собственном пространстве признаков. На практике это означает, что локальная метрика зависит от степени "описанности" объекта, от наличия тех или иных признаков. Так, у пациента некоторые показатели могут отсутствовать по причине нехватки средств или оборудования для проведения подробного анализа.

Как сами окружающие объекты, так и сформированные о них знания (например, описания классов) могут иметь свое признаковое пространство. В нозологии каждое заболевание характеризуется своим набором симптомов. По отношению к этому набору часть соответствующих признаков у пациента могут отсутствовать.

Если ввести понятие *контекста*, который определяет отношения между объектами и, в частности, степень описания самого объекта, то этот контекст проявляется в проекции классов на пространство признаков объекта. Недостаточно описанный объект может попасть в класс, к которому он не принадлежит, только потому, что у него не хватает признака, который дифференцировал бы его от этого класса. Очевидно, что чем меньше степень описания объекта, тем больше пересекаются проекции классов в этом пространстве, и тем худшего качества будет привязанная к объекту локальная метрика, которая определяет его сходство (различие) с другими объектами. Поэтому к такой метрике кроме понятия "локальная" мы добавляем понятие "контекстно-зависимая".

7. Описание контекстно-зависимой локальной метрики

Локальная метрика, основанная на классах эквивалентности, делит все объекты на две группы: входящие в один класс с текущим, и не входящие. Она может принимать только два значения. Если текущий случай попал в класс, то близкими (равными по метрике) ему могут считаться прецеденты этого же класса. Остальные – не равны. Такая метрика не полностью учитывает взаимоотношения между текущим объектом и окружающими (контекст), особенно когда они выражаются через пересечение классов и попадание объекта в область пересечения.

В каких случаях объект попадает в пересечение классов? Формирование классов происходит до рассмотрения исследуемого объекта, и естественно, не в его признаковом пространстве. На этапе предварительной обработки, когда объекты собирают в классы, признаковым пространством для класса будет общее для всех признаков этого класса пространство. Далее, после того, как классы сформированы, естественно рассматривать их в общем для них

признаковом пространстве (в транзитивном замыкании пространств всех объектов).

При рассмотрении исследуемого объекта отнесение его к нескольким классам может возникать, когда у этого объекта часть признаков по отношению к этим классам отсутствует. Другая причина возникает из-за недостаточной или некачественной информации при обучении или при разделении на классы.

Проиллюстрируем такой случай на простом примере (Рис. 2).

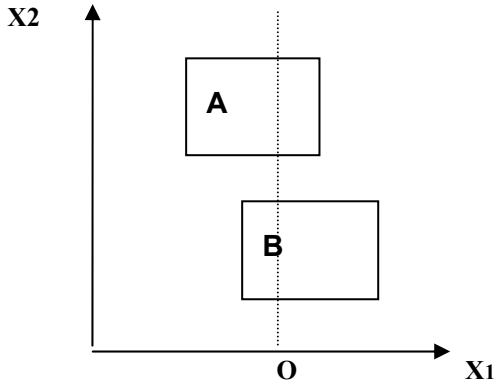

Рис. 2. Отнесение недостаточно описанного объекта к двум классам.

Два непересекающихся класса, A и B, описаны в пространстве признаков $\{X_1, X_2\}$. Объект исследования O представлен одним признаком X_1, признак X_2 у объекта отсутствует. На основе имеющейся информации объект может быть отнесен к обоим классам.

Для более точной оценки нужно было бы добавить контрольному объекту значение признака X_2 (так же поступают и в медицине: если имеющихся показателей не хватает для дифференцирования заболеваний, только дополнительное исследование позволит сделать окончательный вывод), но на практике это не всегда возможно.

До сих пор считалось, что попадание объекта в область пересечения классов является препятствием для оценки объекта. Когда от этой ситуации не уйти, ее надо постараться использовать. Для этого будем использовать *аналоги* – объекты соответствующих классов, попадающие в ту же область пересечения.

Предположим, база прецедентов подверглась предварительной обработке – разбиению на классы эквивалентности. Три способа такого разбиения были указаны в предыдущем разделе. Рассмотрим один из них – кластеризацию – как частный случай, не требующий предварительного обучения.

При рассмотрении текущего случая точка, соответствующая ему, сравнивается с пространственным расположением кластеров в проекции на пространство его

признаков. Близкими считаются прецеденты, принадлежащие кластеру, в который попадает случай. Если он попал в область пересечения кластеров, то ближайшими к нему будут прецеденты, также находящиеся в области пересечения (очевидно, что они могут быть наиболее полезны при выборе решения).

Допустим, он попал в область пересечения кластеров. В зависимости от сложности пересечения, мы можем разделить все объекты на группы (Рис. 3). Прецеденты, находящиеся в одной с текущим случаем области пересечения, естественно считать более близкими к нему, чем те, что находятся только в одном из кластеров, потому что с тем же набором признаков, что и текущий случай, они подобны ему по принадлежности к понятиям, обозначаемым кластерами.

Сравнив введенное понятие близости с тем, что говорилось в предыдущем разделе, нетрудно заметить, что предложенная метрика является локальной и контекстно-зависимой. Локальной, потому что привязана к текущему случаю, контекстной – потому что зависит от его набора признаков. Приведем более строгое определение предлагаемой меры:

> *Расстояние между текущим случаем и прецедентом равно разности количества кластеров, куда попал текущий случай, и количества кластеров из этого числа, в котором находится прецедент.*

Это значит, что расстояние между текущим случаем и прецедентом, находящимся в той же области пересечении кластеров, равно нулю.

На Рис. 3 цифрами помечены области с соответствующим этим цифрам расстоянием между текущим случаем и прецедентами из этой области.

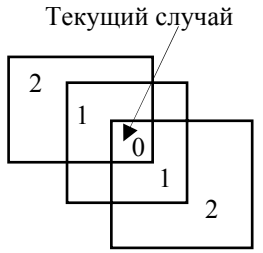

Рис. 3. Степени близости прецедентов.

Предложенная локальная метрика не является метрикой в классическом понимании, а только имеет интерпретацию расстояния. Для нее не гарантируется выполнение правила симметричности, потому что она привязана к объекту, и при переходе к другому объекту будет рассматриваться уже в его

пространстве признаков. По этой же причине не гарантируется выполнение правила треугольника. Однако она позволяет учитывать контекст взаимоотношений объекта с окружающими, особенно в непосредственной близости от него.

8. Заключение

В работе предложен метод принятия решений, основанный на совместном применении ранее не комбинировавшихся (по крайней мере, в отечественных разработках) методов извлечения знаний и вывода по прецедентам, где методы добычи данных используются для автоматического отбора из большой базы прецедентов.

На данный момент все существующие интегрированные системы подобного рода пытаются строить для себя модель данных как способ получения фонового знания. Например, системы, использующие байесовские сети – причинно-следственную модель, системы, использующие предварительную кластеризацию – понятийную модель.

Основная цель привлечения фонового знания в системах вывода по прецедентам – получение сведений для разумного выбора наиболее подходящих прецедентов и адаптации найденного решения. А это, в свою очередь, в большой степени зависит от выбранной меры близости. Наиболее часто используемым методом в выборе прецедентов является метод "ближайшего соседа". В произволе, который допускают системы при выборе меры близости в этом методе, и заключается их главный недостаток.

Предложенная в работе локальная контекстно-зависимая метрика имеет интерпретацию расстояния и позволяет ранжировать объекты, по отношению к исследуемому, целыми числами. При ее построении может быть использована как предварительная кластеризация базы прецедентов, так и разбиение базы на классы эквивалентности с привлечением экспертного знания.

Неполное описания объектов и попадание текущего случая в пересечение понятий (что часто встречается на практике) также не являются препятствием. Более того, сам факт такого пересечения используется в предлагаемой метрике.

При построении метрики используется предложенный авторами модифицированный метод кластерного анализа, ориентированный на распознавание объектов в ситуациях, когда объекты и кластеры имеют не полностью совпадающие наборы признаков. Эта метрика применима к широкому кругу приложений и не накладывает ограничений на типы используемых атрибутов.

В нашей стране такой подход еще не получил должного развития. По сравнению с упомянутой ранее зарубежной разработкой (система M^2), где используется предварительная кластеризация прецедентов, предлагаемый подход позволяет работать в условиях нефиксированного набора атрибутов, что часто встречается в различных приложениях в ситуациях, когда текущий

случай попадает в смешение различных понятий из-за того, что он не полностью описан.

Что касается адаптации решения – предлагаемый метод позволяет сделать эту проблему более формализуемой. Хотя в общем случае проблема адаптации остается зависимой от предметной области, предложенный подход значительно упрощает эту задачу, так как учитывает фоновое знание.

Методы CBR уже применяются во множестве прикладных задач – в медицине, управлении проектами, для анализа и реорганизации среды, разработки товаров массового спроса с учетом предпочтений разных групп потребителей и т. д. Следует ожидать приложений методов CBR к задачам интеллектуального поиска информации, электронной коммерции (предложение товаров, создание виртуальных торговых агентств), планирования поведения в динамических средах, компоновки, конструирования, синтеза программ.

Библиография

Aamodt 94 — Agnar Aamodt and Enric Plaza. "Case-based reasoning: Foundational issues, methodological variations, and system approaches". *AI Communications*, 7(1):39-59, 1994.

Acorn 92 — Acorn T., Walden S. (1992). "SMART: Support management cultivated reasoning technology for Compaq customer service". In Proceedings of AAAI92. Cambridge, MA: AAAI Press/MIT Press.

Aha 93 — Aha D. W., Salzberg S. L. "Learning to Catch: Applying Nearest Neighbor Algorithms to Dynamic Control Tasks". In P. Cheeseman & R. W. Oldford (Eds.) Selecting Models from Data: Artificial Intelligence and Statistics. - New York, NY: Springer-Verlag, 1993.

Aha 95 — Aha D. W. "An Implementation and Experiment with the Nested Generalized Exemplars Algorithm". Technical Report AIC-95-003. - Washington, DC: Naval Research Laboratory, Navy Center for Applied Research in Artificial Intelligence, 1995

Aha 96 — David W. Aha and Li Wu Chang. "Cooperative bayesian and case-based reasoning for solving multiagent planning tasks". Technical report, Navy Center for Applied Research in AI, Naval Research Laboratory, Washington, DC, USA, 1996.

Althof 95/1 — Klaus-Dieter Althof, Eric Auriol, Ralph Barlette, and Michel Manago. *A Review of Industrial Case-Based Reasoning Tools*. AI Intelligence, 1995.

Althof 95/2 — Klaus-Dieter Althof, Eric Auriol, Ralph Traphöner, and Stefan Wess. "Inreca – a seamlessly integrated system based on inductive inference and case-based reasoning". In Agnar Aamodt and Manuela Veloso, editors, *Case-Based Reasoning Research and Development, ICCBR-95*, pages 371-380, 1995.

Althof 96 — Althoff, K.D., Auriol, E., Barletta, R., and Manago, M. "A Review of Industrial Case-based Reasoning Tools". An AI Perspectives Report. Series Editor: Alex Goodall. 1996.

Anand 97/1 — Sarabjot Anand, Bryan Scotney, Mee Tan, Sally McClean, David Bell, John Hughes, and Ian Magill. "Designing a kernel for data mining". *IEEE Expert*, 12(2):65-74, 1997.

Anand 97/2 S. S. Anand, W. Dubitzky, D. Patterson, A. Schuster, J. G. Hughes. "M^2: A First Step Towards Automated Generation and Updating of Case-Knowledge from Databases", Internal Report, Faculty of Informatics, University of Ulster, 1997 (available from http://iserve1.infj.ulst.ac.uk:8080/m2.ps).

Anand 98 Anand S. S., Patterson D. W., Hughes J. G., Bell D. A. "Discovering Case Knowledge Using Data Mining". 2nd Pacific-Asia Conference in Knowledge Discovery in Databases (PAKDD-98), Australia, pp25-35, 1998.

Anand 99 Anand, S.S., Hughes, J.G., Bell, D.A. and Hamilton, P. "Utilising Censored Neighbours in Prognostication", Workshop on Prognostic Models in Medicine, Eds. Ameen Abu-Hanna and Peter Lucas, Aalborg, (AIMDM'99), Denmark, pp15-20, 1999.

Ashley 88 Ashley, K.D. and Rissland, E.L. (1988). "Waiting on Weighting: A Symbolic Least Committment Approach". In Proceedings of the Seventh National Conference on Artificial Intelligence, pp. 239-244.

Ashley 90 Ashley, Kevin D. (1990). "Modelling Legal Arguments: Reasoning with Cases and Hypotheticals", Cambridge, MIT Press

Bareiss 88 Bareiss, E., Porter, B., and Wier, C. (1988). PROTOS: An Exemplar-based Learning Apprentice. International Journal of Man-Machine Studies, 29: 549-561.

Bareiss 91 Bareiss, R., ed. 1991. Proceedings of the DARPA Case-Based Reasoning Workshop. San Francisco, Calif.: Morgan Kaufmann.

Blanzieri 00 Enrico Blanzieri, Luigi Portinale (Eds.): Advances in Case-Based Reasoning, 5th European Workshop, EWCBR 2000, Trento, Italy, September 6-9, 2000, Proceedings. Lecture Notes in Computer Science 1898 Springer 2000, ISBN 3-540-67933-2 Contents BibTeX - EWCBR 2000 Home Page

Brand 98/1 Brand E., Gerritsen R. "Decision Trees". DBMS. - 1998. - № 7.

Brand 98/2 Brand E., Gerritsen R. "Naive-Bayes and Nearest Neighbor". DBMS. - 1998. - № 7.

Breese 95 John S. Breese and David Heckerman. "Decision-theoretic case-based reasoning". In *Proceedings of the Fifth International Workshop on Artificial Intelligence and Statistics*, pages 56-63, 1995.

Breiman 84 Breiman L., Friedman J. H., Olshen R. A., & Stone, C. J. (1984). "Classification and regression trees". Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

Bull 97 M. Bull, G. Kundt, and L. Gierl. "Discovering of health risks and case-based forecasting of epidemics in a health surveillance system". In Jan Komorowski and Jan Zytkow, editors, *Principles of Data Mining and Knowledge Discovery. Proceedings*, pages 68-77, 1997.

Bundy 97 Alan Bundy, editor. *Artificial Intelligence Techniques*. Springer Verlag, 1997.

Brand 98 Brand E., Gerritsen R. "Decision Trees", DBMS. - 1998. - № 7.

Buntine 92 W. Buntine. "A theory of classification rules". 1992

Craw 02 Susan Craw, Alun D. Preece (Eds.): Advances in Case-Based Reasoning, 6th European Conference, ECCBR 2002 Aberdeen, Scotland, UK, September 4-7,

2002, Proceedings. Lecture Notes in Computer Science 2416 Springer 2002, ISBN 3-540-44109-3 Contents BibTeX - ECCBR 2002 Home Page

Curet 96 O. Curet, J. Elliott, M. Jackson. "Designing knowledge discovery based systems in business, finance and accounting with a case-based approach: two case studies", IEEE Colloquium on Knowledge Discovery and Data Mining, 1996

Dingsøyr 98 Torgeir Dingsøyr "Integration of Data Mining and Case-Based Reasoning" http://www.idi.ntnu.no/~dingsoyr/diploma/

Fayyad 96 U. Fayyad, G. Piatetsky-Shapiro, P. Smyth. "From Data Mining to Knowledge Discovery: An Overview". In Advances in Knowledge Discovery and Data Mining (Eds. U.M.Fayyad, G.Piatetsky-Shapiro, P.Smyth), Cambridge, Mass: MIT Press, 1996, pp. 1-34.

Fuernkranz 96 Fuernkranz J. "Separate-and-Conquer Rule Learning". - Vienna: Austrian Research Institute for Artificial Intelligence, Technical Report OEFAI-TR-96-25, 1996.

Funk 04 Peter Funk, Pedro A. González-Calero (Eds.): Advances in Case-Based Reasoning, 7th European Conference, ECCBR 2004, Madrid, Spain, August 30 - September 2, 2004, Proceedings. Lecture Notes in Computer Science 3155 Springer 2004, ISBN 3-540-22882-9 Contents BibTeX - ECCBR 2004 Home Page

Gruvaeus 72 Gruvaeus, G., & Wainer, H. "Two additions to hierarchical cluster analysis". The British Journal of Mathematical and Statistical Psychology, 25, 200-206, 1972.

Hammond 86 Hammond, K. "A model of case-based planning", in Proceedings of the Fifth National Conference on Artificial Intelligence, 65-95. Menlo Park, Calif.: American Associan for Artificial Intelligence. 1986.

Hammond 89 Hammond, K., ed. 1989. Proceedings of the DARPA Case-Based Reasoning Workshop. San Francisco, Calif.: Morgan Kaufmann.

Hartigan 75 Hartigan, J. A. "Clustering algorithms". New York: Wiley, 1975.

Hartigan 78 Hartigan, J. A. & Wong, M. A.. "Algorithm 136. A k-means clustering algorithm". Applied Statistics, 28, 100, 1978.

Haton 94 Jean Paul Haton, Mark T. Keane, Michel Manago (Eds.): Advances in Case-Based Reasoning, Second European Workshop, EWCBR-94, Chantilly, France, November 7-10, 1994, Selected Papers. Lecture Notes in Computer Science 984 Springer 1995, ISBN 3-540-60364-6

Heckerman 95 Heckerman D., Geiger D., Chickering D. "Learning Bayesian networks: The combination of knowledge and statistical data". Machine Learning. - 1995. - 20. - P. 197-243.

Heckerman 97 Heckerman D. "Bayesian Networks for Data Mining". Data Mining and Knowledge Discovery. - 1997. - № 1. - P. 79-119.

HTTP/1 http://is1.cemi.rssi.ru/ruswin/index.htm

HTTP/2 http://www.wizsoft.com

Johnson 67 Johnson, S. C. Hierarchical clustering schemes. Psychometrika, 32, 241-254, 1967.

Kitano 96 Hiroaki Kitano, Hideo Shimazu, and Akihiro Shibata. "Case-method: A methodology for building large-scale case-based systems". In *Proceedings of the AAAI*, pages 303-308, 1993.

Kolodner 83 Kolodner, J.L. "Maintaining Organization in a Dynamic Long-term memory". Cognitive Science, 7(4): 243-280, 1983.

Kolodner 88 Kolodner, J., ed. Proceedings of the DARPA Case-Based Reasoning Workshop. San Francisco, Calif.: Morgan Kaufmann. 1988.

Leake 96 David B. Leake. "Case-Based Reasoning - Experiences, Lessons and Future Directions". AAAI/MIT Press, 1996.

Murthy 97 S. Murthy. "Automatic construction of decision trees from data: A Multi-disciplinary survey", 1997.

Nguyen 93 Nguyen, T., Czerwinski, M., and Lee, D. (1993). "Compaq QUICKSOURCE: Providing the Consumer with the power of AI". AI Magazine, Fall 1993, pp. 50-60.

Parsaye 97 Parsaye K. "Rules are Much More than Decision Trees". The Journal of Data Warehousing. - 1997. - № 1.

Quinlan 87 Quinlan J. R. "Generating production rules from decision trees". In Proceedings of the 10th International Joint Conference on Artificial Intelligence (IJCAI-87). - Morgan Kaufmann, 1987. - P. 304-307.

Quinlan 93 Quinlan J. R. "C4.5: Programs for Machine Learning". Morgan Kaufmann Publishers, San Mateo, CA, 1993.

Rodriguez 97 Andres F. Rodriguez, Sunil Vadera, and L. Enrique Sucar. "A probabilistic model for case-based reasoning". In I Smith and B Faltings, editors, *Case-Based Reasoning Research and Development, ICCBR-97. Proceedings*, pages 623-632, 1997.

Schank 82 Schank, R. (Ed.) (1982). "Dynamic Memory: A Theory of Learning in Computers and People". New York: Cambridge University Press.

Simpson 85 Simpson, R.L. (1985). "A computer model of case-based reasoning in problem solving: An investigation in the domain of dispute mediation". Ph.D. thesis, School of Information and Computer Science, Georgia Institute of Technology.

Smith 96 Ian F. C. Smith, Boi Faltings (Eds.): Advances in Case-Based Reasoning, Third European Workshop, EWCBR-96, Lausanne, Switzerland, November 14-16, 1996, Proceedings. Lecture Notes in Computer Science 1168 Springer 1996, ISBN 3-540-61955-0

Smyth 95 B. Smyth, M. T. Keane. "Remembering to Forget: A Competence-Preserving Case Deletion Policy for Case-Based Reasoning Systems", in Proc. of IJCAI-95, pp 337 - 382, 1995.

Smyth 98 Barry Smyth, Padraig Cunningham (Eds.): Advances in Case-Based Reasoning, 4th European Workshop, EWCBR-98, Dublin, Ireland, September 1998, Proceedings. Lecture Notes in Computer Science 1488 Springer 1998, ISBN 3-540-64990-5 Contents BibTeX

Tirri 96 Henry Tirri, Perti Kontkanen, and Petri Myllymäksi. "A bayesian framework for case-based reasoning". In I Smith and B Faltings, editors, *Advances in Case-Based Reasoning, EWCBR-96*, pages 413-427, 1996.

UKCBR 04 http://www.bcs-sgai.org/ai2004/

UKCBR 05 http://www.bcs-sgai.org/ai2005/

Wang 99 Wang, H., Dubitzky, W., Dьntsch, I., Bell, D.A., "A Lattice Machine Approach to Automated Case Base Design: Marrying Lazy and Eager Learning". Proc. 17th Int. Joint Conference on Artificial Intelligence (IJCAI-99), Sweden, 1999.

Wess 93 Stefan Wess, Klaus-Dieter Althoff, Michael M. Richter (Eds.): Topics in Case-Based Reasoning, First European Workshop, EWCBR-93, Kaiserslautern, Germany, November 1-5, 1993, Selected Papers. Lecture Notes in Computer Science 837 Springer 1994, ISBN 3-540-58330-0

Wettschereck Wettschereck D., Aha D. W., Mohri T. "A Review and Empirical Evaluation of Feature Weighting Methods for a Class of Lazy Learning Algorithms". Artificial Intelligence Review. - 11. - pp. 273-314.

Айвазян 89 Айвазян С. А., Бухштабер В. М., Юнюков И. С., Мешалкин Л. Д. "Прикладная статистика: Классификация и снижение размерности". М.: Финансы и статистика, 1989.

Гупал 93 Гупал А. М., Пономарев А. А., Цветков А. М. "Об одном методе индуктивного вывода с подрезанием деревьев решений". Кибернетика и системный анализ. - 1993. - № 5. - С. 174-178.

Дюк 94 Дюк В. А. "Компьютерная психодиагностика".- СПб: "Братство", 1994.

Дюк 96 Дюк В. А. "Формирование знаний в системах искусственного интеллекта: геометрический подход". Вестник Академии Технического Творчества. - СПб, 1996, №2. - с.46-67.

Дюк 01 Дюк В. А., Самойленко А. П. "Data Mining: учебный курс" – СПб: "Питер", 2001. – 368 с.

Каменнова 95 М.С. Каменнова. "Корпоративные информационные системы: технологии и решения". Системы Управления Базами Данных № 3/1995 стр. 88-99.

Киселев 97 М. Киселев, Е. Соломатин. "Средства добычи знаний в бизнесе и финансах". - Открытые системы, № 4, 1997, с.41-44.

Уоссермен 92 Уоссермен. Ф. "Нейрокомпьютерная техника: Теория и практика". - М.: Мир, 1992. - 240 с.

Цветков 93 Цветков А. М. "Разработка алгоритмов индуктивного вывода с использованием деревьев решений". Кибернетика и системный анализ. - 1993. - № 1. - С. 174-178.