В.В.Самойлов

Агрегирование многомерных объектных данных в задачах анализа ассоциаций[1]

Аннотация. В работе анализируются особенности обучающих данных в новых классах приложений, в которых для обнаружения знаний необходимо использовать методы анализа ассоциаций для случая, когда обучающие данные представлены в объектных базах данных, т.е. представлены множеством взаимосвязанных таблиц реляционной базы данных с онтологией на мета уровне. Предлагается алгоритм агрегирования "сырых" обучающих данных, который позволяет получить более экономное и более информативное их представление. В свою очередь, это позволяет повысить эффективность анализа данных при использовании "классических" методов, а также решать ряд новых задач, в частности, задач причинного анализа данных. Предложенный алгоритм демонстрируется на конкретном примере.

Ключевые слова: интеллектуальный анализ данных и обнаружение закономерностей, анализ ассоциаций, объектные базы данных, онтология, агрегирование обучающих данных.

Введение

Поиск ассоциаций является одним из быстроразвивающихся разделов интеллектуальной обработки данных. Исследования в этой области, которые насчитывают уже почти двадцатилетнюю историю, носят, главным образом, прикладной характер, поскольку они с самого начала были инициированы практическими потребностями и, фактически, до настоящего времени обслуживают потребности, которые возникают на практике. Традиционно этот подход используется в качестве теоретического базиса в области маркетинговых исследований, например, при изучении и прогнозировании спроса на товары в различных регионах. Другие приложения из этой же области – это анализ вариантов покупательских корзин, что необходимо для улучшения организации торговли внутри конкретного магазина и др. Широкое применение получили модели ассоциаций в рекомендующих системах, в частности, в электронной коммерции. Например, любой пользователь электронного книжного магазина Amazon, просматривающий печатные издания по интересующей его тематике, всегда получает некоторые рекомендации, которые генерируются по результатам анализа ассоциаций, извлеченных из предыстории покупок. Перспективной областью использования методов и моделей анализа ассоциаций являются гуманитарные науки. Они могут найти широкое применение при изучении и прогнозировании общественного мнения в социологии, политологии и в других гуманитарных областях, где семантическая насыщенность информации, которой приходится оперировать, обуславливает использование современных интеллектуальных методов и подходов. Перспективно использование анализа ассоциаций для изучения различных виртуальных Интернет–сообществ, в том числе, международных, формирующихся при возникновении многочисленных социальных сетей, форумов и т.п. В частности, такой

[1] Работа частично поддерживается Программой «Информационные технологии и методы анализа сложных систем» Отделения нанотехнологий и информационных технологий Российской академии наук, проект 1.12.

анализ может дать весьма полезные результаты для обнаружения общественно опасных сообществ, например, сообществ, имеющих террористическую направленность, молодежных сообществ фашистского толка и т.п. Эти и многие другие классы приложений могут получить новый импульс благодаря использованию моделей и методов анализа ассоциаций.

Новые приложения диктуют и новые требования к методам интеллектуального анализа данных. Традиционные подходы и алгоритмы, разработанные в области обнаружения закономерностей, в частности, в области анализа ассоциаций, например, для обнаружения часто встречающихся паттернов и ассоциативных правил, зачастую оказываются неэффективными в новых приложениях.

Основные трудности при этом обусловлены типами данных и их размерностью, а также сложностью их структуры.

В данной работе анализируются особенности современных приложений, где перспективно использование методов ассоциативного анализа, а также предлагается алгоритм агрегирования "сырых" данных, который позволяет получить более экономное и более информативное их представление. В свою очередь, это позволяет повысить эффективность анализа данных при использовании "классических" методов, а также решать ряд новых задач, в частности, задач причинного анализа данных [2].

Далее в разделе 1 анализируются особенности современных приложений, в которых анализ ассоциаций является важным этапом обнаружения знаний. В разделах 2 и 3, соответственно, формулируется постановка задачи и дается описание базовой идеи разработанного подхода и алгоритма агрегирования, который выполняет преобразование анализируемых данных к новой модели их представления. В разделе 4 разработанный алгоритм предобработки сырых данных демонстрируется на примере формирования агрегированных признаков в задаче предсказания оценки ("рейтинга") некоторого фильма на основе "предыстории" оценок, которые давались зрителями различным фильмам. В заключении (раздел 5) резюмируются основные результаты работы, а также намечаются пути дальнейших исследований в области анализа ассоциаций, в частности, в области построения эффективных методов выявления причин тех или иных явлений.

1. Анализ особенностей современных задач анализа ассоциаций

Современные методы анализа ассоциаций [3, 6, 13] обычно рассматривают транзакционную исходную модель обучающих (экспериментальных) данных в виде таблицы транзакций/фактов и связанных таблиц объектов/справочников. Эту модель можно привести к виду плоской таблицы с большим количеством пропусков, поскольку один и тот же объект/наименование принимает участие только в ограниченном наборе транзакций. Столбцы этой таблицы идентифицируются именами объектов или идентификаторами их атрибутов/свойств (например, имя типа объекта и его количество), а строки соответствуют "примерам", например конкретным транзакциям, в которых тот или иной объект или некоторый объект с заданным свойством появляется или не появляется. Элементы такой таблицы могут быть представлены в различных шкалах, например, в категориальной шкале (в шкале наименований), числовой или линейно упорядоченной шкалах. Если некоторый объект, которому соответствует определенный столбец в таблице базы данных, присутствует в примере (транзакции), то этот факт кодируется некоторым образом в соответствующей ячейке таблицы, его отсутствие кодируется иным образом.

В классических задачах анализа ассоциаций число столбцов может исчисляться тысячами и более, но число примеров, как правило, намного больше: оно может исчисляться десятками миллионов. Например, если одна строка кодирует набор товаров, который оплатил покупатель в крупном супермаркете, то за сутки в этом супермаркете могут быть накоплены данные, состоящие из десятков тысяч строк. Соответственно, если речь идет о накоплении данных для маркетингового исследования в масштабе региона, притом за период, исчисляемый месяцами, то общее число таких записей может быть порядка сотен миллионов и на порядок больше. Поэтому большинство существующих методов нацелены, прежде всего, на снижение перебора по множеству записей базы данных.

Другая особенность существующих подходов и методов анализа ассоциаций состоит в том, что они, как правило, в качестве модели данных используют их представление в виде

плоской таблицы. Совсем иная ситуация имеет место в новых приложениях, например, в приложениях, имеющих целью постоянный мониторинг динамический пересчет профилей пользователей в рекомендующих системах, в области анализа социальных сетей и виртуальных сообществ, в политологии, социологии, в области семантического поиска веб-ресурсов, извлечении знаний из текстов и тому подобное. Первое принципиальное отличие таких задач от "классических" состоит в том, что в них данные обычно задаются десятками и более различных таблиц и могут быть представлены[2] только в *реляционной базе данных* достаточно сложной структуры. Этот факт вызывает серьезные проблемы, делая методы анализа ассоциаций существенно более громоздкими и менее эффективными. При анализе данных в новых классах приложений это свойство обучающих данных необходимо иметь в виду. Однако исследования последних лет привели к тому, что с этой проблемой "классические" методы более или менее справляются [5].

Есть еще одна особенность данных, более трудная для преодоления, связана с представлением обучающих данных в объектных базах. В объектных базах реляционные данные на мета – уровне дополнительно структурированы онтологией (этот уровень иногда называют также уровнем метазнаний). Как известно, понимание важности использования онтологии в модели представления знаний в интеллектуальных системах является одним из важнейших достижений искусственного интеллекта последнего десятилетия. Именно онтология дала толчок к появлению методов принятия решений, зависящих от текущего контекста, одному из новых направлений исследований и разработок в области искусственного интеллекта и принятия решений. Современные методы интеллектуального анализа данных активно используют новые возможности, которые появляются при этом. Но совместное использование реляционной структуры данных и онтологии, задающей дополнительные *отношения* на множестве понятий предметной области, выдвигает и новые непростые проблемы.

Однако эти проблемы носят скорее технический характер и могут быть преодолены более или менее эффективно благодаря использованию специальных языков запросов к реляционной базе данных, которые формулируются в терминах онтологии. Разработка таких языков является в настоящее время предметом самостоятельных интенсивных исследований [4, 7, 8]. Современные языки запросов такого типа способны осуществлять эффективный поиск. Поскольку анализ таких языков не является целью данной работы, а приложение, рассматриваемое далее, носит демонстрационный характер, то в ней используется стандартный языка SQL, хотя в этом случае запросы получаются более сложными, а эффективность поиска ниже.

Наиболее сложные проблемы анализа ассоциаций в современных приложениях определяются особенностями типов данных, представленных в обучающих данных, а также их размерностью. (Напомним, что под размерностью обучающих данных понимается число различных атрибутов, используемых для их описания, и количество их значений.) Для упомянутого ранее типа задач характерны значительные мощности множеств экземпляров понятий, для которых необходимо отыскивать ассоциации, причем характерно, что это может иметь место даже для онтологий с небольшим множеством базовых понятий. Проиллюстрируем это на одном из достаточно простых примеров, и именно, на примере данных, накопленных в базе данных MovieLense на сайте [11], который предназначен для помощи посетителям сайта в выборе фильмов в соответствии с их предпочтениями. В этой базе данных хранится история посещений большого количества пользователей с информацией об их выборе и последующих оценках просмотренных фильмов. Для каждого фильма оцененного посетителем сайта хранится информация, представленная тремя *базовыми* атрибутами. Первый из них – это *жанр* фильма, *Genre,* который принимает значения из множества $\boldsymbol{G}$ ={*Action, Adult, Adventure, Animation, Biography, Crime, Comedy, Documentary, Family, Fantasy, Horror, Music, Musical, Mystery, Sci-Fi, Short, Sport, Thriller, War*}. Второй признак – это год выпуска фильма в прокат в США, *ReleaseYear*, $\boldsymbol{J}$ =[*1976;2005*]. Третий признак – это оценка (рейтинг), которую посетитель дал фильму.

[2] Часто в "сыром виде" они представляются просто текстовыми файлами, которые еще предстоит декодировать и преобразовать к реляционной форме.

Имеется всего пять вариантов оценки фильма $\Omega = \{1, 2, \ldots, 5\}$, причем рейтинг "1" соответствует самой низкой оценке. Основная задача – это выдача конкретному посетителю рекомендаций по новым фильмам для просмотра. Поэтому система должна обучаться прогнозированию предпочтений посетителей, что делается на основании обучения, для чего может использоваться анализ ассоциаций.

Хотя структура данных MovieLense достаточно проста по сравнению со структурами данных многих других рекомендующих систем коммерческого уровня (см., например, www.netflix.com, www.last.fm), однако множество значений, которые может принимать признак *жанр* (учитывая тот факт, что один и тот же фильм одновременно может характеризоваться несколькими значениями атрибута) для различных пользователей равно 2^{19}. Для таких размерностей данных использование традиционных подходов к обучению, в том числе, с использованием анализа ассоциаций нереально. Кроме того, попытка представить данные в плоской таблице, используя тот или иной язык запросов, приведет в результате к слабо заполненной таблице с огромным числом пропущенных значений, а с такими данными традиционные подходы работают очень плохо. Даже само понятие "множества признаков" по этим данным сложно определить.

Очевидно, что проблема предобработки данных рассмотренного типа является далеко не тривиальной. Она представляет собой самостоятельную проблему, без решения которой задача обнаружения знаний в данных описанного типа для последующего принятия решений вряд ли разрешима. Очевидно также, что не все атрибуты данных являются одинаково информативными, а потому если их использовать напрямую в качестве признаков, то много шансов потерпеть неудачу.

Конечно, имеются и другие проблемы, например, связанные с алгоритмической поддержкой задачи анализа ассоциаций применительно к рассматриваемому типу приложений. Некоторые из проблем проанализированы в работе [2].

В данной работе основное внимание уделяется сформулированной выше задаче предобработки обучающих данных, представленных в объектных базах данных, обладающих описанными выше свойствами. Подчеркнем, что эта проблема в настоящее время является одной их ключевых в рассматриваемом классе приложений.

2. Постановка задачи анализа ассоциаций

Рассмотрим кратко более формальную постановку задачи, для которой решается задача предобработки данных. Предполагается, что обучающие данные заданы множеством объектов в объектной базе данных, т.е. множеством взаимосвязанных таблиц реляционной базы данных, структурированных онтологией, которая задает структуру метазнаний о предметной области. Обучающие данные, подлежащие анализу, задают в неявной форме интерпретированное множество примеров (объектов), принадлежащих одному классифицируемому понятию, но имеющих в качестве атрибута различные метки классов ω_k, $\omega_k \in \Omega = \{\omega_1, \omega_2, \ldots, \omega_q\}$. Заметим, что символы $\omega_k \in \Omega$ могут быть категориальными метками (именами) классов, когда на их множестве не задано какой–либо структуры. Но они также могут быть и упорядоченными, например, частично или линейно. Примером последнего варианта являются значения рейтингов фильмов в базе данных MovieLense, которые принимают значения из множества целых чисел {1, 2, 3, 4, 5}. Очевидно, что в задаче предсказания рейтинга в такой постановке может дополнительно использоваться свойство линейного порядка на множестве классов.

Будем полагать, что значения атрибутов, характеризующих экземпляры понятий, могут быть представлены в категориальной, булевой и/или числовой шкалах. Числовые атрибуты также будем рассматривать как дискретные линейно упорядоченные атрибуты, поскольку они могут быть квантованы некоторым образом. Принимая во внимание тот факт, что категориальные признаки могут принимать значения из множеств большой мощности (число значений отдельных признаков может исчисляться тысячами и более–см. пример для данных относительно простой базы данных MovieLense в предыдущем разделе), использовать их напрямую в качестве категориальных признаков нереально с позиций вычислительной сложности. То же самое относится и к целочисленным призна-

кам типа $\boldsymbol{J}$ =[*1976; 2005*]. Для эффективного поиска ассоциативных (и других) закономерностей необходимо найти их более компактное представление. Иначе говоря, необходимо найти эффективный и рациональный путь их агрегирования. Этот вопрос рассматривается в следующем разделе.

3. Алгоритм агрегирования

Заметим, что описываемый далее подход к построению агрегатов понятий может использоваться не только в задачах ассоциативного анализа. Ассоциативный анализ зачастую является предварительной фазой исследования данных с дальнейшим использованием результатов, например, для построения правил для принятия решений. Здесь мы говорим об анализе ассоциаций по той причине, что предлагаемый подход опробован и практически использовался пока только в задачах анализа ассоциаций и оценки силы причинной связи, описываемой ассоциативным правилом [2], и в этом классе задач обнаружения закономерностей он оказался весьма продуктивным.

Существо разработанного алгоритма состоит в следующем. В соответствии с постановкой задачи $\Omega=\{\omega_1,...,\omega_m\}$ есть множество возможных классов решений, а обучающие данные представлены в объектной базе данных с онтологией на мета–уровне. Пусть $\boldsymbol{X}=\{X_1,...,X_t\}$ есть множество понятий онтологии, связанных с классифицируемым понятием, где X_u, $u\in\{1,...,t\}$, есть символ, используемый для обозначения имени понятия, а также и имени множества экземпляров этого понятия.

Используя некоторый язык запросов, для каждого понятия онтологии можно перечислить все его примеры, в том числе и примеры, связанные с экземплярами классифицируемого понятия, имеющим в качестве атрибута заданное значение метки класса. Пусть $a_s\in X_u$ есть некоторый экземпляр понятия X_u, $u\in\{1,...,t\}$, а $X_u^{(i)}\subseteq X_u$, $i=1,...,m$, есть некоторое *подмножество* таких экземпляров, которое ставится в соответствие классу $\omega_i\in\Omega$. Введем унарный предикат $B_u^{(i)}(a_s)$ таким образом, что он принимает значение "*истина*" тогда и только тогда, когда $a_s\in X_u^{(i)}$. Таким образом, $X_u^{(i)}$ есть область истинности унарного предиката $B_u^{(i)}$, т.е. $B_u^{(i)}$ принимает значение "*true*" для любого a_s: $a_s\in X_u^{(i)}$, и значение "*false*" – в противном случае.

Общая идея предобработки данных с целью агрегирование отдельных значений признаков состоит в следующем. Предполагается, что для любого $a_s\in X_u^{(i)}$ и $X_u^{(i)}\subseteq X_u$ экземпляр понятия $a_s\in X_u^{(i)}$ тогда и только тогда, когда для заданного $\omega_i\in\Omega$ и для любого $\omega_v\in\Omega$, $i\neq v$, выполняется неравенство

$$p(\omega_i / a_s) > p(\omega_v / a_s). \qquad (1)$$

Неравенство (1) выражает тот факт, что условная вероятность $p(\omega_i / a_s)$ класса ω_i при условии $X_u^{(i)}=a_s$ больше, чем условная вероятность любого иного класса при $X_u=a_s$.

Чтобы вычислить все элементы агрегата $X_u^{(i)}$, необходимо проверить неравенство (1) для всех $a_s\in X_u$ и для всех $\omega_v\in\Omega$ при $i\neq v$. Поскольку найти все экземпляры понятия X_u, заданные в обучающих данных, можно с помощью подходящего языка запросов к объектной базе данных [9], то неравенство можно проверить всех u, v, i и s. Таким образом можно построить все агрегаты $X_u^{(i)}$ и, соответственно, области истинности предикатов $B_u^{(i)}$.

Естественно при использовании неравенства (1) необходимо задавать некоторый порог. Тогда правило построения агрегатов $X_u^{(i)}$ (и соответственно, предикатов $B_u^{(i)}$) будет выглядеть таким образом

Правило 1.

$$\text{Если } p(\omega_k / a_s) > p(\omega_l / a_s) + \Delta \qquad (2)$$

(Δ–положительное число), то $a_s\in X_u^{(i)}$

Однако этим правилом можно воспользоваться только тогда, когда известны априорные

вероятности классов $p(\omega_i)$, $\omega_i \in \Omega$. Но во многих случаях эти вероятности неизвестны или не могут быть достоверно оценены по обучающим данным. В этом случае лучше всего предположить равновероятность классов [10] (см. также [1]). Тогда правило формирования агрегатов $X_u^{(i)}$ (и соответственно, предикатов $B_u^{(i)}$) будет выглядеть следующим образом:

Правило 2.

$$\text{Если } p(a_s/\omega_k) \geq p(a_s/\omega_l) + \Delta \quad (3)$$

(Δ–положительное число), то $a_s \in X_u^{(i)}$

Описанный выше подход применим к понятиям, связанным с классифицируемым понятием в отношении *1 – 0...N*, которые описываются множеством экземпляров, представленных как в категориальной шкале, так и в линейно упорядоченной (целочисленной, дискретной числовой) шкале. Заметим, что в случае его использования для понятий с атрибутами в числовой шкале данный алгоритм является более общим, чем обычно используемые алгоритмы, эксплуатирующие разбиение числового интервала значений атрибута на множество непересекающихся подынтервалов, поскольку результирующий агрегат может состоять из несвязного множества интервалов.

Заметим, что в порядке исследования алгоритма агрегирования применительно к числовым признакам был проанализирован алгоритм оптимального разделения числового интервала, отвечающего области значений некоторого понятия (или его атрибута) по критерию максимизации меры *информационный выигрыш* (*information gain*), предложенной Дж. Квинланом (J.R.Quinlan) в его методе C4.5 [12]. Однако оказалось, что он работает не очень удачно. Например, результаты, полученные по этому методу, могут не удовлетворять условиям, сформулированным в правилах (2), (3). Это можно видеть из примера, который приводится в следующем разделе.

Подчеркнем одну важную особенность решения, которое заменяет множество экземпляров понятий их агрегатами. По существу эта процедура носит характер первичного обучения. Она с самого начала направлена на построение более сильных ассоциаций между значениями одного понятия и связанного с ним другого понятий в каждом классе. По этой причине можно ожидать, что ассоциативные правила, построенные на таких пропозициях, будут обладать хорошими характеристиками по мере уверенности, приписываемой правилам [3]. Практика решения нескольких задач подтверждает это предположение. Действительно, постановка и используемая идея решения задачи агрегирования имеют под собой достаточно ясную мотивацию, хотя используемый алгоритм носит эвристический характер. Заметим, однако, что в практических задачах интеллектуальной обработки данных строгая постановка задачи, как правило, вообще невозможна, а соответствующие процедуры интеллектуального анализа данных базируются на интуитивных понятиях типа "информативность" признаков, "хорошее продукционное правило", "часто встречающийся паттерн", "хорошее ассоциативное правило" и тому подобное.

Следует обратить внимание на одну важную особенность результата, получаемого описанным выше образом. Типы исходных атрибутов могут быть любыми, однако результирующее пространство признаком содержит только однотипные признаки, находящиеся в отношении *1 – 1* с классифицируемым понятием, что позволяет свести многомерные обучающие данные к плоской таблице без пропущенных значений, т.е получить задачу обучения в традиционной форме. Эти признаки могут быть либо бинарными, если в дальнейшем в качестве механизма принятия решений предполагается использовать логические модели, либо бинарными с приписанной им вероятностной мерой, если далее будет использоваться логико–вероятностная модель принятия решений, либо только вероятностные меры множеств $X_u^{(i)}$, $u \in \{1,...,t\}$, $i = 1,...,m$, если предполагается использовать вероятностные модели принятия решений. Важно подчеркнуть, что последние две модели результата могут использоваться совместно, поскольку модель с вероятностной мерой множеств $X_u^{(i)}$, $u \in \{1,...,t\}$, $i = 1,...,m$, и модель с использованием бинарных предикатов $B_u^{(i)}$, $u \in \{1,...,t\}$, $i = 1,...,m$, могут быть использованы для построения двух изоморфных нормированных булевых алгебр, а именно, нормированной булевой алгебры множеств (алгебры событий–в соответствии с терминологией, принятой в теории вероятностей) и норми-

рованной булевой алгебры пропозициональных формул. При этом их совместное использование позволит корректным образом использовать вывод в вероятностной логике, вычисляя вероятности заключений с помощью операций алгебры событий. Теоретическое обоснование, а также детали такого подхода изложены в работе [2].

4. Экспериментальные результаты

Продемонстрируем предложенный алгоритм агрегирования данных, заданных объектной базой данных, на примере задачи предсказания рейтинга фильма по данным базы данных *MovieLense* [11]. Рейтинг фильма оценивается по пятибалльной шкале, поэтому полагаем, что имеется 5 классов решений $\Omega = \{1,2, \dots ,5\}$, причем рейтинг "1" соответствует самой низкой оценке. Для выбранного зрителя объемы использованных обучающих выборок в базе данных *MovieLense* для каждого из классов таковы: $N_1 = 2$, $N_2 = 15$, $N_3 = 99$, $N_4 = 209$, $N_5 = 65$. Всего, таким образом, обучающая выборка содержит 390 примеров.

По правилу 2 (раздел 3) построены агрегаты для всех классов для понятия жанр (*Genre)*, и год выпуска фильма (*ReleaseYear*), при этом они строились так, чтобы ими можно было воспользоваться для различения всех пар классов. Соответственно, обозначим агрегаты, которые должны быть построены, символами $X_i(\omega_k, \omega_l)$, $k,l = 1,...,m$, $i = 1,.2$, где первый аргумент указывает целевой класс, а второй – альтернативный класс, индекс $i = 1$ соответствует понятию *Genre* и $i = 2$ соответствует понятию *ReleaseYear*. Для понятия *жанр* для каждого класса таким способом было построено по четыре агрегата. Например, для класса ω_3 эти агрегаты получились такими:

$X_1(\omega_3, \omega_1) = \{$*Action, Adult, Adventure, Animation, Biography, Crime, Documentary, Family, Fantasy, Music, Musical, Mystery, Sci-Fi, Short, Sport, Thriller*$\}$,

$X_1(\omega_3, \omega_2) = \{$*Adult, Adventure, Animation, Biography, Documentary, Fantasy, Music, Musical, Mystery, Sci-Fi, Short, Sport*$\}$,

$X_1(\omega_3, \omega_4) = \{$*Adult, Music*$\}$,

$X_1(\omega_3, \omega_5) = \{$*Adult, Musical, Sci - Fi*$\}$,

Обратим внимание, что в общем случае $X_r(\omega_j, \omega_k) \neq X_r(\omega_k, \omega_j)$. В этом можно убедиться, анализируя содержание агрегатов, построенных для понятия *Genre*, в которых имя класса ω_3 выступает в качестве второго аргумента в других агрегатах, построенных для этого понятия:

$X_1(\omega_1, \omega_3) = \{$*Comedy*$\}$,

$X_1(\omega_2, \omega_3) = \{$*Horror*$\}$,

$X_1(\omega_4, \omega_3) = \{$*History, Horror, War*$\}$,

$X_1(\omega_5, \omega_3) = \{$*History, War*$\}$.

Приведем другие примеры агрегатов, построенных на множестве экземпляров понятия *Genre*, в которых имя класса ω_5 используется либо на первом, либо на втором месте:

$X_1(\omega_1, \omega_5) = \{$*Comedy*$\}$,

$X_1(\omega_2, \omega_5) = \{$*Horror* $\}$,

$X_1(\omega_3, \omega_5) = \{$*Adult, Musical, Sci - Fi*$\}$,

$X_1(\omega_4, \omega_5) = \{$*Horror, Musical, Sci - Fi*$\}$,

$X_1(\omega_5, \omega_1) = \{$*Action, Adventure, Animation, Biography, Crime, Documentary, Family, Fantasy, History, Music, Mystery, Short, Sport, Thriller, War*$\}$,

$X_1(\omega_5, \omega_2) = \{$*Adventure, Animation, Biography, Documentary, Fantasy, History, Music, Mystery, Short, Sport, War*$\}$,

$X_1(\omega_5, \omega_3) = \{$*History, War*$\}$,

$X_1(\omega_5, \omega_4) = \{$*Music*$\}$.

Можно видеть, что агрегаты, построенные на том же множестве экземпляров понятия *Genre* для различных целевых классов, и классов, выступающих в качестве вторых аргументов, сильно отличаются. Обратим внимание, что общее число признаков, сформированных согласно предложенному алгоритму, равно 20 вместо потенциально возможного их числа, равного 2^{19}.

То же самое было проделано для числового понятия (признака) *ReleaseYear*. Однако в отличие от понятия *Genre*, экземпляры которого представлены в категориальной шкале, при построении агрегатов для него использовался алгоритм разделения значений числового интервала путем максимизации меры Квинлана

информационный выигрыш [12], что уже отмечалось в предыдущем разделе. Отметим, что для числовых атрибутов и понятий можно воспользоваться также и любым из вышеперечисленных вариантов. В рассматриваемом примере для числового атрибута были построены следующие агрегаты:

$X_2(\omega_3, \omega_1) = \{J \geq 1992{,}5\}$, $X_2(\omega_3, \omega_2) = \{J \geq 1996{,}5\}$,

$X_2(\omega_3, \omega_4) = \{J \geq 1983{,}5\}$, $X_2(\omega_3, \omega_5) = \{J \geq 1980\}$.

Обратим внимание еще раз на важное достоинство предложенного алгоритма агрегирования. Оно состоит в том, что все результирующие признаки оказываются заданными в единой булевой шкале, хотя исходные понятия имеют целочисленный и категориальный типы.

Ранее было высказано предположение о том, что признаки, которые строятся с помощью предложенного алгоритма, могут уже обладать хорошей разделительной силой. В Табл. 1–2 приведены результаты тестирования агрегатов понятий *Genre и ReleaseYear* (они играют роль признаков в новом, "агрегированном" пространстве), построенных для целевого класса ω_3. Этим агрегатам соответствуют восемь признаков – предикатов $B_1(\omega_3, \omega_i)$ и $B_2(\omega_3, \omega_i)$, другого класса, выступающего в роли второго аргумента предиката. В каждой ячейке обеих таблиц приведены два числа. Первое из них соответствует числу примеров обучающих данных, которые попадают в ячейку таблицы, а второе число является оценкой вероятности правильного решения(соответствующие ячейки

Табл. 1. Результаты тестирования агрегата, построенного для признака *Genre.*

Истинный класс	Истинностное значение предиката $B_1(\omega_3, \omega_1)$, построенного для агрегата $X_1(\omega_3, \omega_1)$	
	" *истина* " (класс 3)	"*ложь*" (альтернативный класс)
1	0/0	2/0,020
3	61/0,604	38/0,376
Истинный класс	Истинностное значение предиката $B_1(\omega_3, \omega_2)$, построенного для агрегата $X_1(\omega_3, \omega_2)$	
	" *истина* " (класс 3)	"*ложь*" (альтернативный класс)
2	4/0,036	11/0,096
3	47/0,412	52/0,456
Истинный класс	Истинностное значение предиката $B_1(\omega_3, \omega_4)$, построенного для агрегата $X_1(\omega_3, \omega_4)$	
	" *истина* " (класс 3)	"*ложь*" (альтернативный класс)
3	6/0,019	93/0,302
4	13/0,042	196/0,637
Истинный класс	Истинностное значение предиката $B_1(\omega_3, \omega_5)$, построенного для агрегата $X_1(\omega_3, \omega_5)$	
	" *истина* " (класс 3)	"*ложь*" (альтернативный класс)
3	7/0,043	92/0,561
5	5//0,030	60/0,366

Табл. 2. Результаты тестирования агрегата, построенного для признака *ReleaseYear*

Истинный класс	Истинностное значение предиката $B_2(\omega_3, \omega_1)$, построенного для агрегата $X_2(\omega_3, \omega_1)$	
	" *истина* " (класс 3)	"*ложь*" (альтернативный класс)
1	0/0	2/0,020
3	12/0,119	87/0,861
Истинный класс	Истинностное значение предиката $B_2(\omega_3, \omega_2)$, построенного для агрегата $X_2(\omega_3, \omega_2)$	
	" *истина* " (класс 3)	"*ложь*" (альтернативный класс)
2	6/0,053	9/0,079
3	80/0,701	19/0,167
Истинный класс	Истинностное значение предиката $B_2(\omega_3, \omega_4)$, построенного для агрегата $X_2(\omega_3, \omega_4)$	
	" *истина* " (класс 3)	"*ложь*" (альтернативный класс)
3	99/0,322	0/0
4	203/0,659	6/0.019
Истинный класс	Истинностное значение предиката $B_3(\omega_3, \omega_5)$, построенного для агрегата $X_2(\omega_3, \omega_5)$	
	" *истина* " (класс 3)	"*ложь*" (альтернативный класс)
3	99/0,604	0/0
5	62/0,378	3/0,018

окрашены в серый цвет) и вероятностям ошибок (*false positives* и *false negatives*). Анализ содержания ячеек, в которых даны оценки вероятностей правильного принятия решений в случае, если соответствующий агрегат используется в качестве единственного правила классификации, показывает, что более чем в половине случаев вероятность правильного разделения классов (она равна сумме оценок вероятностей, представленных в ячейках серого цвета) больше 0,5. Этот результат следует считать очень хорошим, поскольку агрегаты такого качестве далее используются как компоненты нового признакового пространства. Далее они должны использоваться для построения решающих правил. Решение этой задачи представлено в работе [2].

В целом, экспериментальные результаты, полученные в данном примере, а также результаты, полученные в ряде других примеров, подтвердили работоспособность используемых эвристик, положенных в основу алгоритма агрегирования экземпляров понятий, представленных в объектных базах данных.

Заключение

Главные цели данной работы: во–первых, обратить внимание исследователей и разработчиков на особенности многомерных данных, используемых в процессах извлечения знаний, в частности, в задачах анализа ассоциаций в современных приложениях, которые обычно представляются в объектных базах данных, а также на важность тщательного подхода к проблеме предварительной их обработки, во–вторых, предложить свое решение для одной из ключевых задач предобработки данных, а именно, задачи формирования признакового пространства с помощью агрегирования многомерных гетерогенных признаков (категориальных, числовых, упорядоченных), когда области определения соответствующих переменных имеют достаточно большую мощность. Задачи такого рода являются сложными при построении моделей принятия решений, в частности, моделей классификации и прогнозирования.

По существу, основная идея данной работы состоит в том, что предварительная обработка данных для последующего извлечения знаний, закономерностей, сама по себе должна быть "интеллектуальной" и многоэтапной. Заметим, что в настоящее время уже широко используются различные методы интеллектуальной предобработки данных. Типичным примером является задача обнаружения и фильтрации ошибочных данных, которые "выпадают" из генеральной совокупности, а по сути – ей не принадлежат. Среди методов обнаружения таких аномалий в данных наиболее эффективными являются методы кластерного анализа, которые хорошо работают, например, тогда, когда приходится иметь дело с выборками малого объема.

В данной работе задача формирования признакового пространства решается как задача построения простых правил классификации. Такой подход позволяет преодолеть трудности, связанные с большой размерностью задачи, а также трудности, возникающие при совместной обработке гетерогенных признаков. Работоспособность разработанного подхода подтверждена экспериментальными результатами.

Будущие исследования и разработки планируется проводить в направлении расширения области практических приложений предложенного подхода, а также в области разработки новых методов агрегирования данных, представленных с помощью объектных баз данных. Использование онтологии предоставляет такие возможности.

Литература

1. Городецкий В. И., Серебряков С.В. Методы и алгоритмы коллективного распознавания //Автоматика и Телемеханика. 2008. № 11. С. 3–40.
2. Городецкий В.И., Самойлов В.В. Ассоциативный и причинный анализ и ассоциативные байесовские сети. Сборник Трудов СПИИРАН, выпуск 9, стр. 13–65.
3. J-M.Adamo. Data Mining for Association Rules and Sequential Patterns. Springer, 2000.
4. T.Calders, B., A.Prado. Integrating Pattern Mining in Relational Databases Lecture notes in computer science, vol. 4213,
5. S.Dzeroski, N.Lavrac (Eds.) Relational Data Mining. Springer, 2001, 398 p.
6. J.Han, M.Kamber. Data Mining. Concept and Techniques. Morgan Kaufman Publishers, 2000. 454-461.
7. J. Han, Y. Fu, W. Wang, K. Koperski and O.Zaiane. DMQL: A Data Mining Query Language for Relational Databases. In SIGMOD DMKD Workshop, 1996.
8. T. Imielinski and A. Virmani. MSQL: A Query Language for Database Mining. Data Mining and Knowledge Discovery, Vol. 3, 1999, 373–408.
9. S.Jean, Y.Aït-Ameur, G.Pierra. Querying Ontology Based Database Using OntoQL (An Ontology Query Language),,Lecture Notes in Computer Science, Vol. 4275, 2006, 704-721.

10. J. Kittler, M. Hatef, P. W. Duin. On combining classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence, № 20(3), 1998, 226–239.
11. http://movielense.com.
12. J. Quinlan. C4.5: Programs for Machine Learning. Elsevier Science & Technology Books, 1992. 316 p.
13. C.Zhang, S.Zhang: Association Rule Mining, Models and Algorithms Springer 2002, 240 p.

Самойлов Владимир Владимирович. Научный сотрудник Санкт-Петербургского института информатики и автоматизации РАН. Окончил Высшее военно-морское инженерное училище им. Ф.Э. Дзержинского в 1993 году. Опубликовал более 30 работ в области: интеллектуальный анализ данных и обнаружение закономерностей, многоагентные системы и т.д. E-mail: samovl@iias.spb.su.