УДК 004.853

В.И. Городецкий, О.Н. Тушканова

АССОЦИАТИВНАЯ КЛАССИФИКАЦИЯ: АНАЛИТИЧЕСКИЙ ОБЗОР. ЧАСТЬ 1

Городецкий В.И., Тушканова О.Н. **Ассоциативная классификация: аналитический обзор. Часть 1.**

Аннотация. В работе описаны основные результаты, модели и методы, разработанные в области ассоциативной классификации, ориентированные на обработку данных большого объема. В работе дается постановка задачи ассоциативной классификации, вводится необходимая терминология и формальные обозначения, используемые в ассоциативной классификации. Приводится описание и сравнительный анализ ранних подходов, методов и конкретных алгоритмов ассоциативной классификации. Дается оценка вклада первых работ, посвящённых ассоциативной классификации, в развитие этого направления.

Ключевые слова: большие данные, ассоциативное правило, ассоциативная классификация.

Gorodetsky V., Tushkanova O. **Associative classification: analytical overview. Part 1.**

Abstract. The paper topic is associative classification intended for processing of big data. It formulates corresponding problem statement and introduces basic concepts and formal notation used in associative classification. An extended overview and comparative analysis of the early approaches, models and algorithms for associative classification form the main paper contents. The paper assesses the contribution of the first papers devoted to associative classification to the development of this area and formulates goals of the further research.

Keywords: associative classification, emerging pattern, big data.

1. Введение. Среди множества современных моделей анализа данных особое место занимают модели, связанные с решением наиболее представительного класса задач принятия решений, а именно задач классификации. История исследований моделей классификации насчитывает уже несколько десятилетий, однако их актуальность не снижается ввиду регулярного появления новых классов задач и конкретных приложений, специфических типов данных, описывающих объекты классификации, новых требований к качеству решения задач и т.п. Современные приложения во многом не похожи на те, для которых развивались классические модели, методы и алгоритмы решения задач классификации. Типичными примерами приложений нового типа являются задачи интеллектуального управления бизнесом (англ. *business intelligence*), анализ социальных сетей и принятие решений по тем или иным аспектам их функционирования, приложения в области прогнозирования спроса и персонификации предложений в интернет–торговле, в области продвижения веб-сайтов и многие другие. Изменился также и характер самих задач классификации. Например, важ-

нейшую роль приобретают задачи типа *Что, ..., если?*, в которых нужно определять атрибуты, факторы, явления, процессы и т.п., которые связаны с наблюдаемыми на практике нежелательными отклонениями свойств тех или иных процессов от номинальных или от желаемых значений, чтобы затем управлять ими для изменения свойств процессов в нужном направлении. Эти и другие причины требуют разработки новых и новых моделей, методов и алгоритмов в области, которую принято называть интеллектуальным анализом данных, в частности, в области обучения классификации и синтеза классификаторов.

Специфика приложений, примеры которых приведены выше, проявляется, прежде всего, в особенностях данных, которые доступны в них для принятия решений. Новый класс задач подобного рода и новые непростые проблемы появились в последнее время в связи развитием концепции *больших данных*. Хотя понятие больших данных (англ. *Big Data*) было введено совсем недавно (2008 г.) [1], оно быстро вошло в обиход специалистов, и в настоящее время является общепринятым.

Естественно, что точного определения понятия больших данных не существует, и оно дается описательно. К большим данным относят не просто данные большого объема и/или высокой размерности. Большие данные обычно обладают сложной структурой. Как правило, они описываются разнородными атрибутами (числовыми, булевыми, ординальными, номинальными), содержат в себе тексты на естественном языке, изображения, а также данные некоторых специальных форматов. К ним относятся, например, веб–ссылки, адреса электронной почты, номера телефонов, адреса и имена организаций и людей, даты и т.п. Нередко такие данные имеют характер потоков во времени. Поэтому часто они представляются множеством транзакций и первично записываются в транзакционных базах данных. Такие данные могут иметь объемы, измеряемые терабайтами, петабайтами и более, и размерности, исчисляемые сотнями и тысячами атрибутов, каждый их которых, в свою очередь, может иметь сложную структуру или быть неструктурированным, например, может быть текстом на естественном языке. Как принято сейчас говорить, эти данные характеризуются "*тремя V*": *Volume*, *Velocity*, *Variety*, т.е. *объемом*, *скоростью* прироста и *разнообразием* шкал и структур представления [1, 2]. Типичными примерами больших данных являются распределенные потоки текстовых сообщений из социальных сетей, метеорологические, экологические и другие пространственно–временные данные исследований окружающей среды. Сюда же относятся потоки данных о соединениях абонентов сотовой связи и их местонахождениях, серверные данные

интернет–торговли, содержащие информацию о покупателях, товарах, динамике и структуре покупок, данные о финансовых потоках банков с офисами, распределенными глобально, потоки данных о дорожном движении в мегаполисах и т.п.

Появление таких данных потребовало не только новых методов и средств их хранения (NoSQL, Hadoop и др.), но также и создания специальных технологий их обработки и представления, в частности, визуализации. Отметим, что в литературе иногда к понятию *большие данные* относят не только сами данные с описанными выше свойствами, но также и все специфические методы, средства и технологии их хранения, обработки и представления результатов. Иначе говоря, этот термин иногда используется также и как название соответствующего научного направления.

В настоящее время для работы с большими данными, в частности, в интересах решения задач классификации, развиваются новые методы, модели и алгоритмы интеллектуальной обработки больших данных. Одним из таких новых направлений является *ассоциативная классификация*. Следует отметить, что хотя в большинстве исследований аспект, непосредственно связанный с особенностями ассоциативной классификации применительно к большим данным явно не обсуждается, тем не менее, все исследования в этой области акцентируют внимание на эффективной работе именно с данными большого объема и размерности.

В данной работе дается обзор алгоритмов, моделей и методов, разработанных в области ассоциативной классификации применительно к обработке данных большого объема на начальном этапе развития этого направления интеллектуального анализа данных. Задача ассоциативной классификации была впервые сформулирована в работе [3], так что данное направление активно развивается уже в течение более чем 15 лет. В разделе 2 дается постановка задачи ассоциативной классификации, вводится необходимая терминология и формальные обозначения, используемые в последующей части обзора. Раздел 3 представляет основное содержание данной работы. В нем приводится описание и сравнительный анализ первых результатов, полученных в области ассоциативного анализа. В заключении по работе дается оценка вклада работ, посвящённых ассоциативной классификации, в развитие этого направления на его начальном этапе, а также формулируются цели дальнейшего исследования.

2. Ассоциативная классификация: термины, обозначения и модели ассоциаций. В данном обзоре рассматриваются методы поиска ассоциативных правил, которые ориентированы на решение задач

классификации. При поиске ассоциативных правил конкретное приложение и задача, в которой далее предполагается использовать полученные правила, не конкретизируется. В отличие от этого, в ассоциативной классификации такая задача указывается явно, и потому в ней отыскиваются ассоциативные правила специального вида. Поэтому если задача поиска ассоциативных правил в общем случае относится к задачам поиска анализа связей в данных (англ. *data mining*), то задача поиска ассоциативных связей относится к области машинного обучения (англ. *machine lrarning*). В правой части этих правил может присутствовать только целевая переменная, а именно *метка класса*, что существенно сужает множество искомых правил и, следовательно, снижает сложность поиска. Такие правила принято называть *ассоциативными правилами класса* (*class associative rules, CARs*).

Задачи ассоциативной классификации обладают и рядом других особенностей, поэтому поиск классифицирующих ассоциативных правил имеет свою специфику. В отличие от классических постановок задач поиска ассоциативных правил, обучающие данные для синтеза ассоциативных классификаторов могут не являться транзакциями, что на практике может привести к значительному увеличению числа потенциальных правил. Например, обучающие данные могут быть гетерогенными, иметь сложную структуру и даже являться текстами. Это усложняет задачу предобработки данных, используемых для обучения. Существуют и другие особенности задач ассоциативной классификации, которые требуют введения некоторых специальных понятий, используемых авторами работ по данной тематике.

Рассмотрим формальную постановку задачи ассоциативной классификации, введем некоторые базовые термины и формализмы, используемые в существующей литературе по данной проблематике, в частности, авторами работ, анализируемых далее.

Пусть $\boldsymbol{D}$ – транзакционная база данных (множество данных), $D_i \in \boldsymbol{D}$ – произвольная транзакция, $\boldsymbol{X}$ – множество всех символов, которые используются для обозначения объектов (признаков, атрибутов) в транзакциях множества $\boldsymbol{D}$, $\boldsymbol{A}$ – подмножество символов из множества $\boldsymbol{X}$ и $\boldsymbol{D(A)}$ – подмножество множества транзакций из множества $\boldsymbol{D}$, каждая из которых содержит подмножество символов $\boldsymbol{A} \in \boldsymbol{X}$ в качестве подмножества. Для характеристики статистических свойств подмножества $\boldsymbol{A}$ в базе данных $\boldsymbol{D}$ используют отношение мощности n_A множества $\boldsymbol{D(A)}$ к мощности n всего множества транзакций $\boldsymbol{D}$. Эту величину принято называть *поддержкой* (*support*) подмножества $\boldsymbol{A}$ во множестве транзакций $\boldsymbol{D}$:

$$supp(\boldsymbol{A}) = n_A / n . \quad (1)$$

Пусть даны два набора символов (объектов) $\boldsymbol{A}\in\boldsymbol{X}$ и $\boldsymbol{B}\in\boldsymbol{X}$, причем $\boldsymbol{A}$ и $\boldsymbol{B}$ не имеют общих элементов, и пусть σ и γ – вещественные числа из интервала [*0, 1*]. Говорят [4, 5], что выражение вида $\boldsymbol{A}\rightarrow\boldsymbol{B}$ есть *ассоциативное правило* с *порогом уверенности* $conf(\boldsymbol{A}\rightarrow\boldsymbol{B})=\gamma$ и *порогом поддержки* $supp(\boldsymbol{A})=\sigma$ (σ, γ – ассоциативное правило), если справедливы следующие неравенства:

$$n_{AB}/n\geq\sigma, \tag{2}$$

$$n_{AB}/n_A\geq\gamma, \tag{3}$$

где n_{AB} – количество транзакций во множестве $\boldsymbol{D}$, которые содержат объединение множества символов подмножеств $\boldsymbol{A}$ и $\boldsymbol{B}$. Модель ассоциативного правила, заданную условиями (2), (3), принято называть моделью типа *поддержка–уверенность*.

Подмножество (последовательность) элементов $\boldsymbol{A}$ принято называть посылкой ассоциативного правила $\boldsymbol{A}\rightarrow\boldsymbol{B}$, а подмножество (последовательность) $\boldsymbol{B}$ – его следствием. Обычно эти последовательности называют паттернами (*patterns*). В задачах ассоциативной классификации заключение правила может содержать только однолитерный паттерн, который является именем одного из классов. Поэтому в общем случае основная подзадача задачи ассоциативной классификации сводится к поиску множества (σ, γ)–ассоциативных правил для каждого класса. Эта подзадача называется обычно задачей *обучения* классификатора. Другая подзадача – это синтез классификатора на множестве найденных ассоциативных правил.

Сделаем два важных замечания, касающихся понятия *ассоциативное правило*. Первое замечание касается задания линейного порядка на множестве символов $\boldsymbol{X}$. Множества $\boldsymbol{D(A)}$ и $\boldsymbol{D(B)}$– это множества транзакций, в которых каждая транзакция содержит подмножества $\boldsymbol{A}$ и $\boldsymbol{B}$, соответственно, в качестве подмножеств. В них конкретные символы могут следовать в любом порядке. Если наборы этих символов интерпретировать как компоненты векторов (это часто создает большие удобства с формальной точки зрения), то их следует рассматривать как упорядоченные последовательности (цепочки) символов. Зададим линейный порядок на множестве всех символов $\boldsymbol{X}$ и будем, где это удобно, рассматривать любое его подмножество как последовательность символов, упорядоченную в соответствии с введенным порядком на множестве $\boldsymbol{X}$.

Второе замечание касается интерпретации ассоциативного правила как некоторой статистической зависимости. Можно видеть, что ассоциативное правило задает *статистическую зависимость* между

посылкой правила A и его следствием B, а числа σ, γ являются статистическими оценками двух вероятностей. Величина σ является оценкой вероятности $p(A)$ появления последовательности A в транзакциях базы данных D, а величина γ является оценкой вероятности появления последовательности B в транзакциях этой базы, в которых появилась последовательность A, т.е. γ является оценкой условной вероятности $p(B/A)$. Даже в серьезной литературе часто можно встретить высказывания о том, что ассоциативное правило задает отношение импликации. Однако это серьезное заблуждение. Чтобы избежать в дальнейшем путаницы, подчеркнем, что семантика отношения, задаваемого ассоциативным правилом, является совершенно иной, чем семантика отношения, задаваемого импликацией в вероятностной пропозициональной логике или отношением выводимости, задаваемым в аналогичном пропозициональном исчислении. В этом контексте термин *ассоциативное правило* применительно к отношению $A \rightarrow B$ вряд ли является удачным, поскольку он может ввести в заблуждение относительно его семантики. Однако этот термин является общепринятым, а потому для этого отношения далее, несмотря на его неудачность, будет использоваться именно он.

Классические методы поиска ассоциативных правил используют модель, известную под названием *Apriori* [6]. Эта модель является переборной с механизмом отсечения, использующим свойство антимонотонности вероятности появления паттерна (его поддержки) по мере увеличения его длины. Существенно более эффективным методом является группа алгоритмов, известная под названием *FP-growth* [7], однако он пользуется меньшей популярностью у прикладников ввиду его большей сложности. Заметим, что если база данных не является транзакционной, то для использования названных алгоритмов потребуется некоторое преобразование данных.

Понятие ассоциативного правила, введенное условиями (2) и (3), обладает большим недостатком, а именно, оно не учитывает возможную вероятностную независимость, которая может существовать между паттернами A и B, когда говорить о существовании ассоциации не имеет смысла. Действительно, можно столкнуться с ситуацией, в которой меры поддержки и уверенности будут достаточно большими просто за счет больших значений вероятностей компонент паттернов, и тогда может быть сделано ошибочное заключение о существовании ассоциативной связи между этими паттернами.

Попытка ослабить действие отмеченного недостатка модели ассоциативного правила вида *поддержка–уверенность* (2), (3) была

предпринята в модели Г.Пятецкого–Шапиро, предложенной в [8]. Дополнительно к мерам поддержки и уверенности, в ней вводится еще один параметр для выбора или отклонения правила. Этот параметр использует известную точечную статистическую меру зависимости между случайными величинами, представленную нижеследующей формулой:

$$I = (n \times n_{AB}) / (n_A \times n_B), \tag{4}$$

которая является статистической оценкой величины

$$I = \frac{P(\boldsymbol{A}, \boldsymbol{B})}{P(\boldsymbol{A})P(\boldsymbol{B})}. \tag{5}$$

В формуле (5) величина $\boldsymbol{P(A,B)}$ есть вероятность совместного появления паттернов $\boldsymbol{A}$ и $\boldsymbol{B}$, а величины $\boldsymbol{P(A)}$ и $\boldsymbol{P(B)}$ – это вероятности появления паттернов $\boldsymbol{A}$ и $\boldsymbol{B}$ в этой же выборке. Близость этой величины к единице свидетельствует о слабой статистической зависимости между паттернами $\boldsymbol{A}$ и $\boldsymbol{B}$. Эта величина в модели Г. Пятецкого–Шапиро используется для задания пороговой характеристики

$$\left| \frac{P(\boldsymbol{A}, \boldsymbol{B})}{P(\boldsymbol{A})P(\boldsymbol{B})} - 1 \right| \geq \delta_{\min}, \tag{6}$$

при этом величина δ_{min} названа автором *минимальным интересом*.

Соответственно, в данной модели определение ассоциативного правила, дополнительно к требованиям (2) и (3), расширено требованием (6), которое позволяет отличить зависимые паттерны $\boldsymbol{A}$ и $\boldsymbol{B}$ от независимых. Эту модель называют моделью *поддержка–уверенность–зависимость*. Параметрами алгоритмов для поиска ассоциативных правил в этом случае являются значения минимальной поддержки $\sigma_{\min}$, минимальной уверенности $\gamma_{\min}$ и минимального интереса δ_{min}. С помощью выбора их значений можно управлять числом ассоциативных правил, которые будут генерироваться соответствующей программой, а также их качеством.

Более строго эта же модель оценки зависимости между компонентами паттерна введена в работе [5]. В ней для проверки зависимости паттернов $\boldsymbol{A}$ и $\boldsymbol{B}$ используется классический χ^2–тест математической статистики, проверяющий значимость гипотезы о равенстве слу-

чайных величин (точечных оценок вероятностей), присутствующих в числителе и знаменателе метрики Г. Пятецкого–Шапиро (5):

$$H_0 : P(\boldsymbol{A},\boldsymbol{B}) - P(\boldsymbol{A})\,\boldsymbol{P}(\boldsymbol{B}) = 0. \qquad (7)$$

Как обычно, алгоритм проверки этой гипотезы состоит в том, чтобы сосчитать оценки вероятностей отдельных паттернов по выборке, сосчитать оценку вероятности их совместного появления в выборке и оценить по критерию χ^2 значимость различия между этими величинами для заданного объема выборки и заданного порога отсечки. Напомним, что значение порога отсечки задает уровень значимости этого различия. Обычно уровень значимости должен быть не меньше, чем 0,95. Заметим, что в работе [8] какая-либо оценка разброса случайной величины (7) не рассматривается.

Обратим внимание на следующее свойство описанной здесь модели ассоциативного правила. Ассоциативная связь, отвечающая модели *поддержка–уверенность–зависимость*, является симметричной относительно посылки и заключения. Другими словами, она не дает информации о направлении этой связи, утверждая только, что или $\boldsymbol{A} \rightarrow \boldsymbol{B}$, или $\boldsymbol{A} \leftarrow \boldsymbol{B}$, или эта связь двухсторонняя, т.е. $\boldsymbol{A} \leftrightarrow \boldsymbol{B}$. Естественно, что отсутствие информации о направлении ассоциативной связи в рассматриваемой здесь модели ассоциативного правила является ее недостатком. Этот недостаток преодолевается в моделях ассоциаций причинного типа, в которых рассматривается направленная статистическая связь вида $\boldsymbol{A} \rightarrow \boldsymbol{B}$. Построение формальной модели причинной связи, методы поиска и использования ассоциаций причинного типа в задачах принятия решений – это достаточно актуальная и практически важная тема, которая активно исследуется уже в течение почти трех десятилетий. Она развивалась сначала в рамках модели байесовских и причинных сетей доверия, в которых напрямую понятие ассоциативной связи не используется. Однако в последнее десятилетие намечается достаточно сильная конвергенция идей причинных сетей доверия и ассоциативных правил классификации в рамках направления, которое называется ассоциативно–причинная классификация [9, 10]. Однако это уже специальная тема, которая требует отдельного рассмотрения.

В последующем материале данной работы описываются основные результаты, модели, методы, и алгоритмы, разработанные в области ассоциативной классификации, а также приводится их сравнительный анализ применительно к работе с данными большого объема.

3. Ассоциативная классификация: Начальные модели и методы. Работа [3] была первой работой, в которой сформулирована за-

дача ассоциативной классификации. Во многом эта работа выполнена по аналогии с другими моделями поиска правил классификации, которые были разработаны ко времени ее публикации.

В этой работе в постановке задачи рассматриваются обучающие данные, заданные в форме обычной таблицы *объект–признак* с N примерами (строками), l атрибутами и q классами. Полагается, что атрибуты данных являются либо категориальными, либо целочисленными, либо числовыми. Категориальные атрибуты просто нумеруются (точнее–заменяются последовательными целочисленными значениям, однако это не вполне корректно, т.к. при этом на значениях атрибутов искусственно вводится некоторый порядок, реально несуществующий), непрерывные атрибуты заменяются набором дискретных значений и также заменяются нумерованными атрибутами. Таким способом каждый пример выборки трансформируются во множество пар <*атрибут, целочисленное значение*>, которому ставится в соответствие метка класса. В работе рассматривается модель ассоциативного правила в стандартной форме *поддержка–уверенность* вида $\boldsymbol{A_i} \rightarrow \boldsymbol{B_k}$, где посылка $\boldsymbol{A_i} \in \boldsymbol{X}$ есть последовательность пар <*атрибут, целочисленное значение*>, а $\boldsymbol{B_k}$ есть метка класса, $k \in \{1,\ldots,q\}$. Заметим, что авторы ошибочно называют такое правило импликацией (см. по этому поводу замечание в разделе 2).

Предложенный в работе алгоритм поиска ассоциативных правил назван *CBA* (*Classification Based on Associations*). Он состоит из трех шагов, среди которых *первым* является уже описанный алгоритм приведения данных к квази–целочисленной форме. На *втором* шаге поочередно для каждой метки класса генерируется все множество ассоциативных правил, удовлетворяющих заданным ограничениям на минимальные значения мер поддержки и уверенности. Этот шаг реализуется с использованием стандартного алгоритма *Apriori* [6]. Поиск правил для ассоциативной классификации применительно к конкретному классу реализуется с помощью эвристической процедуры на третьем шаге. Для этого используются слегка модифицированные идеи бустинга [11]. Сначала на множестве всех правил класса определяется линейный (тотальный) порядок таким образом:

Правило $r_i \succ r_j$ (первое правило предшествует второму), если

1. Мера уверенности *conf* правила r_i больше, чем правила r_j.

2. Меры уверенности обоих правил одинаковы, но правило r_i имеет большее значение меры поддержки.

3. Обе меры имеют одинаковые значения для обоих правил, но первое правило сгенерировано раньше второго.

Опишем общую идею *третьего* шага предложенного алгоритма для поиска множества правил классификации для конкретного класса, например, для класса $\boldsymbol{B_k}$. Сначала из множества правил, упорядоченного по отношению $\succ$ и имеющего в качестве заключениях имя класса $\boldsymbol{B_k}$, выбирается правило с наименьшим номером. Далее для этого правила находятся все примеры, которые этим правилом *покрываются*. Из обучающей выборки все примеры, найденные таким образом, удаляются, и далее они в процессе выбора правил для класса $\boldsymbol{B_k}$ участия не принимают. Далее аналогичные действия выполняются для следующего (по порядку) правила и т.д. Если очередное правило покрывает хотя бы один новый пример, то оно рассматривается как потенциальное правило классификации и помечается соответствующим образом. После каждого шага вычисляется относительное число ошибок классификации, достигаемое при использовании сформированного множества правил. Если эта величина превышает заданный порог, то процесс формирования итогового множества правил классификации продолжается до тех пор, пока на текущем шаге остаются примеры, которые еще не покрыты ни одним из выбранных правил. Заметим, что примерно так же строится обычная процедура обучения на основе бустинга [11]. После останова процедуры отбора правил классификации из итогового множества удаляются те правила, которые не улучшают точность классификации.

Описанный алгоритм выбора правил удовлетворяет следующим двум условиям:

1. Каждый пример в данных, используемых для обучения, будет покрыт хотя бы одним правилом, и это правило имеет наименьший номер в построенной последовательности правил.

2. Каждое правило в выбранном множестве покрывает хотя бы один пример данных, который не покрывается другим правилом.

Очевидно, что такой алгоритм не является эффективным. Авторы это понимают и предлагают его эвристическую модификацию. Суть этой модификации алгоритма состоит в том, что в нем наилучшее правило отыскивается поочередно для каждого примера. Поэтому эвристический вариант алгоритма является многопроходным по множеству данных. Алгоритм состоит из трех этапов.

На первом этапе для каждого примера $d \in \boldsymbol{D}$ класса $\boldsymbol{B_k}$ находятся два правила, одно из которых правильно классифицирует этот пример и имеет наименьший номер в последовательности правил для этого класса (оно обозначается *cRule*). Второе, аналогичное первому, имеет наименьший номер, но классифицирует этот пример неверно (*wRule*). Если $cRule \succ wRule$, то *cRule* включается во множество потен-

циальных правил классификации для класса $\boldsymbol{B_k}$. В противном случае ситуация обрабатывается несколько более сложным образом. Это связано с тем, что на текущем этапе пока неясно, как правило *cRule* ведет себя по отношению к примерам других классов, и следует ли его включать в потенциальное множество правил для класса $\boldsymbol{B_k}$. Для каждого *cRule* определяется также (и хранится в некоторой структуре данных) то множество примеров, которое этим правилом покрывается.

На втором этапе для тех примеров, для которых выбор правила на этапе 1 сделан не был, выполняется повторный проход по данным. На этом проходе отыскиваются все те *wRule*–правила, которые предшествуют *cRule,* построенному для соответствующего примера. Далее каждое *wRule*– правило, найденное для примера, анализируется, и если оно помечено как *cRule*–правило для некоторого другого примера данного класса, то оно оставляется в найденном множестве, в противном случае – удаляется.

На третьем этапе выполняется финальный выбор правил для класса B_k. Он реализуется в два шага. На первом шаге найденное множество правил упорядочивается и далее вычисляется ошибка классификации по мере увеличения числа правил в соответствии с установленным их порядком. При этом может оказаться, что некоторые правила не покрывают новых примеров (не увеличивают точность классификации). Тогда такие правила удаляются. На втором шаге удаляются правила, которые вносят наибольшие ошибки. В работе [3] приведен псевдокод этого алгоритма, который дополняет его деталями, опущенными здесь.

Авторы сравнивают свой алгоритм с классическим алгоритмом *C4.5* на основе вычислительных экспериментов с 26 наборами данных из *UCI ML Repository* [12] и делают вывод о том, что предложенный ими алгоритм ассоциативного поиска правил классификации демонстрирует более высокую точность.

Ценность этой работы состоит, по-видимому, в том, что она сформулировала проблему ассоциативной классификации, очертила ее особенности и предложила эвристический алгоритм классификации *CBA*. Этот алгоритм аналогичен обычному алгоритму поиска ассоциативных правил с добавлением идей бустинга, разработанного ранее для машинного обучения. Однако более поздние работы показали, что задача поиска ассоциативных правил для решения задач классификации значительно своеобразнее и намного сложнее, чем это может показаться на основании работы [3].

Развитием алгоритма *CBA* является алгоритм *CMAR* (***C****lassification based on* ***M****ultiple* ***A****ssociation* ***R****ules*), предложенный в работе [13].

В этой работе задача классификации формулируется аналогично тому, как она сформулирована в [3]. Ее принципиальное отличие от последней работы состоит в том, что в ней делается акцент на повышение эффективности, причем как процессов генерации ассоциативных правил, так и самих алгоритмов классификации. Заметим, что обеспечение вычислительной эффективности ассоциативной классификации – это ключевая проблема, которая плохо решается большинством методов, предложенных для этих целей. Эта проблема особенно остро проявляется при работе с большими данными.

Для обеспечения высокой эффективности авторы вводят в алгоритм *CBA* два новшества. Первое состоит в том, что они отказываются от использования переборных алгоритмов типа *Apriori* для поиска ассоциативных правил. Вместо него они используют свой метод, разработанный ими двумя годами раньше, который хорошо известен в литературе под названием *метод возрастающих паттернов* (англ. *Frequent Pattern growth, FP-growth*) [7]. Основное новшество этого алгоритма, которое и является источником его эффективности, состоит в том, что все последовательности–кандидаты на включение в искомое множество часто встречающихся паттернов, формирующих посылки правил, представляются в виде *префиксного дерева последовательностей* (*FP-tree*). В этом дереве каждый узел соответствует некоторому символу множества $\boldsymbol{X}$, а последовательности символов с общим префиксом представляются общей последовательностью узлов дерева с началом в его корне. После генерации множества часто встречающихся паттернов дерево типа *FP-tree* оказывается очень удобной структурой их представления. После некоторой модификации оно используется как для реализации процедур отсечения “плохих” правил, так и для просмотра и поиска правил в процессе классификации новых примеров, когда требуется выполнять мэтчинг (сравнение с образцом, от англ. *matching*) тестируемого примера с большим числом ассоциативных правил. Получаемая в итоге структура для представления ассоциативных правил называется авторами *CR-tree*.

Дадим описание этого алгоритма и дерева *CR-tree* более детально, поскольку генерация ассоциативных правил в алгоритме *CMAR* несколько отличается от аналогичного процесса в общем случае алгоритма *FP-growth*. По сравнению с алгоритмом *FP-growth* алгоритм поиска часто встречающихся паттернов в *CMAR* имеет два основных отличия. Если алгоритм *FP-growth* выполняется в два шага, на которых сначала строится дерево часто встречающихся последовательностей, которые имеют значение поддержки больше заданного порогового значения, а затем генерируются ассоциативные правила с требуе-

мым значением меры уверенности, то в алгоритме *CMAR* часто встречающиеся паттерны и правила генерируются за один шаг. Второе отличие состоит в том, что алгоритм *CMAR* для каждого правила запоминает распределение значений поддержки на множестве всех классов, для которых данное правило имеет ненулевое ее значение. Последнее не требует дополнительных вычислений, поскольку значения поддержки для классов в *CMAR* вычисляются в любом случае.

Множество сгенерированных ассоциативных правил, каждое из которых имеет три атрибута – метка класса, значение поддержки и меры уверенности, хранится в структуре дерева *CR-tree.* Пример такого дерева представлен на рисунке 1 (рисунок заимствован из работы [13]). Можно видеть, что *CR-tree* является достаточно компактной структурой. Оно уже хранит индексы для доступа к правилам. Очевидно, что просмотр правил при таком их представлении является эффективной процедурой.

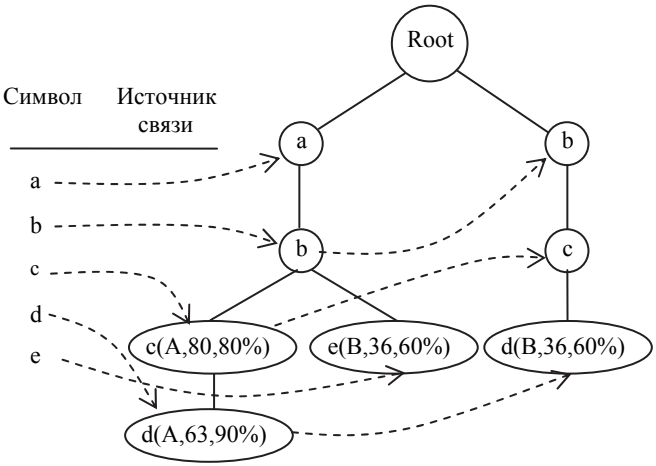

Рис. 1. Пример *CR-tree.* A, B–метка класса, 80–поддержка (число примеров класса, содержащих паттерн), 80%–значение уверенности

Однако не все правила построенного таким способом дерева используются в алгоритме классификации. Наименее эффективные правила удаляются на этапе отсечения. Можно видеть, что и процедура отсечения может быть эффективно реализована с помощью того же дерева *CR-tree* следующим образом. Сначала на множестве правил задается глобальный порядок, в соответствии с которым далее упоря-

дочиваются все правила класса. Порядок этот строится следующим образом.

Говорят, что из двух правил r_1 и r_2 правило r_1 имеет больший ранг, чем правило r_2, иначе, $r_1 \succ r_2$, если и только если (1) $conf(r_1)>conf(r_2)$; или (2) если $conf(r_1)=conf(r_2)$, но $supp(r_1)>supp(r_2)$; или (3) если $conf(r_1)=conf(r_2)$ и $supp(r_1)=supp(r_2)$, но правило r_1 в левой части имеет меньшее число символов, чем правило r_2. Говорят также, что правило r_1: $\boldsymbol{P}\rightarrow\boldsymbol{C}$ является более общим по отношению к правилу r_2: $\boldsymbol{P}'\rightarrow\boldsymbol{C}'$ тогда и только тогда, когда посылка $\boldsymbol{P}$ первого правила $\boldsymbol{P}$ является подмножеством посылки $\boldsymbol{P}'$ второго правила. Введенный порядок используется алгоритмом *CMAR* для отсечения правил из дерева *CR-tree.*

Процедура отсечения выполняется за три шага:

Шаг 1. Отсекаются менее общие правила с низким значением меры уверенности. Это отсечение выполняется каждый раз, когда правило вставляется в *CR-tree* в процессе генерации правил.

Шаг 2. После построения дерева *CR-tree* из него удаляются правила, в которых посылка и заключение имеют отрицательную корреляцию. Этот факт определяется применением χ^2–теста. Заметим, что с необходимостью удаления таких правил нельзя согласиться безоговорочно, поскольку правила с отрицательной корреляцией в некоторых случаях несут очень полезную информацию для классификации, в особенности, если связь имеет причинный характер. Например, если некоторый паттерн имеет значение коэффициента корреляции с заключением близкое к величине *–1,* то такое правило является *запретом* для данного класса. А запрет представляет собой очень сильную закономерность. Его удаление из множества классификационных правил вряд ли оправдано.

Шаг 3. Отсекаются правила, имеющие значение фактора покрытия примеров класса в обучающей выборке меньшее, чем заданный порог.

Что касается алгоритма классификации, то он строится на основе полученного множества классификационных правил аналогично тому, как это делается в алгоритме *CBA* с некоторым отличием в модели объединения решений, выдаваемых различными правилами. В алгоритме *CMAR* сначала все правила, оставленные в дереве *CR-tree* после отсечения разбиваются на группы, каждая их которых в заключении содержит метку одного и того же класса. При тестировании примера для каждой такой группы правил в *CMAR* вычисляется значение веса. Этот вес является некоторой эвристически выбранной функцией, которую авторы называют *взвешенной χ^2–статистикой* [13]. Заметим,

что эта мера имеет достаточно сложное выражение. Авторы признают, что она не имеет никакого теоретического обоснования или содержательной интерпретации. Мотивацией для ее использования в алгоритме *CMAR* являются только результаты экспериментальных исследований. Решение принимается в пользу того класса, для которого взвешенная χ^2–статистика принимает наибольшее значение.

Свойства алгоритма *CMAR* исследованы на 26 наборах тестовых данных из UCI репозитория [12]. На основании экспериментальных результатов, приведенных в работе, авторы делают вывод о том, что алгоритм *CMAR* обладает существенно лучшими характеристиками по вычислительной эффективности по сравнению с методами *CBA* и *C4.5* [14]. Он обладает также лучшими показателями по точности решения задач классификации, однако в этом отношении его преимущества не столь существенны.

Для генерации правил ассоциативной классификации в работе [15] предлагается использовать идеи метода ID3 [16]. Этот метод был очень популярен в период с 1980 по 2000 г.г. Хотя предложенный метод, для которого авторы используют название *CPAR (**C**lassification based on **P**redictive **A**ssociation **R**ules),* эксплуатирует довольно старую идею, он имеет некоторые новые свойства, заслуживающие упоминания.

В основу метода положен алгоритм *FOIL* [17]. Этот алгоритм рекурсивно отыскивает атрибут (признак), добавляемый к последовательности–потенциальной посылке формируемого правила, который максимизирует метрику, называемую *информационным выигрышем* (*information gain*). Максимизация этой метрики ведет к наилучшему покрытию текущего множества обучающих данных. Данные выборки, покрытые найденным правилом, удаляются из обучающего множества, и далее поиск атрибутов, обеспечивающих максимизацию информационного выигрыша, ведется по отношению к новому, сокращенному множеству обучающих данных. Как хорошо известно, этот метод работает с парой классов. Если классов много, то метод *FOIL* использует хорошо известную схему “выбранный класс” – “все другие классы” для сведения задачи с множеством классов к последовательности задач бинарной классификации. Алгоритм CWAC [18], по сути, аналогичен алгоритму FOIL. Отличие состоит только в том, что авторы [18], кроме *информационного выигрыша,* используют для оценки правил *взвешенную поддержку* и *взвешенную уверенность*.

Модификация же метода CPAR применительно к задаче ассоциативной классификации состоит в следующем. В отличие от *FOIL*, который на каждом шаге рекурсивного формирования посылки прави-

ла допускает только одно ее продолжение за счет добавления нового атрибута, алгоритм *CPAR* рассматривает несколько таких продолжений. В качестве вариантов продолжений он рассматривает все те атрибуты, которые имеют одинаковые или близкие значения *информационного выигрыша*. Эту часть алгоритма *CPAR* авторы называют *PRM*–алгоритмом, *Predictive Rule Mining.* Далее, в отличие от того, как это делается в некоторых вариантах алгоритмов генерации правил классификации с использованием идей бустинга [11], пример обучающих данных, покрытый вновь сгенерированным правилом, не удаляется из процесса обучения. Он используется и на последующих шагах поиска, но с меньшим весом. Каждое сгенерированное правило оценивается по некоторой метрике, значение которой используется в дальнейшем для принятия решения о том, использовать ли то или иное правило в алгоритме классификации или нет. Для каждого класса оставляется k наилучших (по упомянутой метрике) правил, на базе которых и строится алгоритм классификации новых примеров. Заметим, что механизм классификации на основе множества построенных правил не отличается оригинальностью. В нем для классифицируемого примера оценивается среднее значение вероятности его принадлежности к каждому классу по множеству всех правил. Предпочтение отдается тому из классов, для которого эта точность наибольшая. Детали алгоритма, как и доказательства корректности различных его шагов применительно к задачам ассоциативной классификации, могут быть найдены в работе [15].

Метод *CPAR* был исследован экспериментально на 26 наборах данных их UCI–репозитория [12]. По утверждению авторов, он превзошел по точности предсказания класса другие методы, которые на момент публикации работы [15] рассматривались как наилучшие. К ним относятся, в частности, такие методы, как *C4.5* [16], *RIPPER* [19], *CBA* [3], *CMAR* [13] и ACAC [14, 20].

Заметим, что в число алгоритмов, с которыми авторы сравнивали метод *CPAR*, не вошли алгоритмы, основанные на использовании понятия эмерджентных паттернов, хотя эти алгоритмы были опубликованы за несколько лет до публикации алгоритма *CPAR*.

4. Заключение. Модели ассоциативной классификации, получившие основное развитие в течение последних пятнадцати лет, предлагают подход, который пытается интегрировать в себе некоторые базовые результаты теории и практики классического индуктивного обучения и механизмы поиска ассоциативных правил. Целью такой интеграции является повышение вычислительной эффективности и точности решения традиционных задач классификации с ориентацией на

использование полученных моделей для анализа больших данных. Первые модели ассоциативной классификации рассматривали задачу ассоциативной классификации просто как частный случай задачи поиска ассоциативных правил, в котором правая часть правила может принимать значения из фиксированного множества меток (идентификаторов классов). Рассмотренные в работе алгоритмы *CBA* [3], *CMAR* [13] и *CPAR* [15], являются примерами такого прямолинейного подхода. Однако именно в этих работах была явно сформулирована проблема ассоциативной классификации, очерчены ее особенности. Эти работы задали исходный уровень эффективности алгоритмов ассоциативной классификации и выявили главные проблемы, которые возникают при обучении и использовании моделей ассоциативной классификации, определив тем самым направления дальнейшего развития этих моделей.

Отметим, что более поздние работы показали, что задача поиска ассоциативных правил для решения задач классификации значительно своеобразнее и намного сложнее, чем это может показаться на основании работ [3, 13, 15].

Во второй части текущей работы будет более подробно рассмотрена группа методов и алгоритмов, предназначенных для поиска ассоциативных правил классификации с помощью эмерджентных паттернов. Этот подход, по существу, определил новое направление в области ассоциативной классификации. Это направление активно развивается вплоть до настоящего времени. Его результаты позволяют устранить некоторые недостатки начальных моделей и алгоритмов ассоциативной классификации, описанных в данной работе. Работы в этом направлении [21-24] во многом способствовали более глубокому пониманию специфики задач ассоциативной классификации и путей ее эффективной алгоритмизации.

Литература

1. Wikipedia.org: the free encyclopedia // URL: http://en.wikipedia.org/w/index.php?title=Big_data&oldid=556537897 (дата обращения 20.06.2014 г.).
2. *Douglas L.* 3D Data Management: Controlling Data Volume, Velocity and Variety // URL: http://blogs. gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf (дата обращения 20.06.2014 г.).
3. *Liu B., Hsu W., Ma Y.* Integrating classification and association rule mining // Proceedings of the KDD'98, New York, NY, Aug. 1998. pp. 80–86.
4. *Городецкий В.И., Самойлов В.В.* Ассоциативный и причинный анализ и ассоциативные байесовские сети // Труды СПИИРАН. 2009. №9. С. 13-65.
5. *Brin S., Motwani R., Silverstein C.* Beyond market baskets: generalizing association rules to correlations // Proceedings of the ACM SIGMOD Intern. Conf. on Management of Data. 1997. pp. 255–264.

6. *Agrawal R., Sricant R.* Fast Algorithm for Mining Association rules // Proceedings of the 20th Intern. Conference on Very Large Databases, Santiago, Chile. 1994. pp. 68-77.
7. *Han J., Pei J., Yin Y.* Mining frequent patterns without candidate generation // Proceedings of the ACM SIGMOD Intern. Conf. on Management of Data. 2000. pp. 1–12.
8. *Piatetsky–Shapiro G.* Discover, analysis, and presentation of strong rules // Knowledge discovery from Databases. G. Piatetsky–Shapiro and W.Frawley (Eds.). AAAI Press/MIT Press. 1991. pp. 229-248.
9. *Aliferis C.F., Statnikov A., et al.* Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification Part I: Algorithms and Empirical Evaluation // Journal of Machine Learning Research. 2010. no. 11. pp. 171-234.
10. *Aliferis C.F., Statnikov A.*, et al. Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification Part II: Analysis and Extensions // Journal of Machine Learning Research. 2010. No. 11. pp. 235 – 299.
11. *Schapire R.E.* The Boosting Approach to Machine Learning. An Overview Nonlinear Estimation and Classification. Springer. Lecture Notes in Statistics. vol. 171. Denison D.D., Hansen M.H, Holmes C.C., Mallick B., Yu B. (Eds). 2003. pp. 149–172.
12. *Blake C.L., Murphy P.M.* UCI Repository of machine learning database. University of California, Department of Information and Computer Science. Irvine, CA. 1998 // URL: http://www.cs.uci.edu/ mleam/mlrepository.html (дата обращения 20.06.2014).
13. *Li W., Han J., Pei J.* CMAR: Accurate and efficient classification based on multiple class-association rules // Proceedings of the ICDM'01, San Jose, CA, Nov. 2001. pp. 369–376.
14. *Wedyan S.* Review and Comparison of Associative Classification Data Mining Approaches // International Journal of Computer, Information, Systems and Control Engineering. 2014. vol. 8 no.1. pp. 34–45.
15. *Yin X., Han J.* CPAR: Classification Based on Predictive Association Rule // Proceedings of the SDM'03. 2003. pp. 369–376.
16. *Quinlan J.R.* C4.5: Programs for Machine Learning. Morgan Kaufmann, 1993.
17. *Quinlan J.R., Cameron-Jones R.M.* FOIL: A midterm report // Proceedings of the European Conference on Machine Learning. Vienna, Austria. 1993. pp. 3–20.
18. *Ibrahim S., Chandran K.R.* Compact Weighted Class Association Rule Mining using Information Gain // International Journal of Data Mining & Knowledge Management Process (IJDKP). 2011. vol.1, no.6. pp. 1–13.
19. *Cohen W.* Fast effective rule induction // Proceedings of the ICML'95. Tahoe City, CA. 1995. pp. 115–123.
20. *Huang Z., Zhou Z., He T., Wang X.* ACAC: Associative Classification based on All-Confidence // Proceedings of IEEE International Conference on Granular Computing (GrC). 2011. pp. 289-293.
21. *Dong G., Li J.* Efficient Mining of Emerging Patterns: Discovering Trends and Differences // Proceedings of the KDD'99. 1999. pp. 43–52.
22. *Dong G., Zhang X., Wong L., Li J.* CAEP: Classification by Aggregating Emerging Patterns // Proceedings of the DS'99, .1999. pp. 30–42.
23. *Fan H., Ramamohanarao K.* Fast Discovery and the Generalization of Strong Jumping Emerging Patterns for Building Compact and Accurate Classifiers // IEEE Trans. Knowl. Data Eng. 2006. vol. 18(6). pp. 721-737.
24. *Li J., Dong G, Ramamohanarao K.* Making use of the most expressive jumping emerging patterns for classification // Proceedings of the Fourth Pacific-Asia Conference on Knowledge Discovery and Data Mining. Kyoto, Japan. 2000. pp. 220-230.

References

1. Wikipedia.org: the free encyclopedia. Available at: http://en.wikipedia.org/w/index.php?title=Big_data&oldid=556537897 (Accessed: 20.06.2014 г.).
2. Douglas L. 3D Data Management: Controlling Data Volume, Velocity and Variety. Available at: http://blogs. gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf (Accessed: 20.06.2014 г.).
3. Liu B., Hsu W., Ma Y. Integrating classification and association rule mining. Proceedings of the KDD'98, New York, NY, Aug. 1998. pp. 80–86.
4. Gorodetsky V., Samoylov V. [Associative and causal analysis and associative Bayesian networks]. *Trudy SPIIRAN – SPIIRAS Proceedings*. 2009. vol.9. pp. 13-65. (In Russ.).
5. Brin S., Motwani R., Silverstein C. Beyond market baskets: generalizing association rules to correlations. Proceedings of the ACM SIGMOD Intern. Conf. on Management of Data. 1997. pp. 255–264.
6. Agrawal R., Sricant R. Fast Algorithm for Mining Association rules. Proceedings of the 20th Intern. Conference on Very Large Databases, Santiago, Chile. 1994. pp. 68-77.
7. Han J., Pei J., Yin Y. Mining frequent patterns without candidate generation. Proceedings of the ACM SIGMOD Intern. Conf. on Management of Data. 2000. pp. 1–12.
8. Piatetsky–Shapiro G. Discover, analysis, and presentation of strong rules. Knowledge discovery from Databases. G. Piatetsky–Shapiro and W.Frawley (Eds.). AAAI Press/MIT Press. 1991. pp. 229-248.
9. Aliferis C.F., Statnikov A., et al. Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification Part I: Algorithms and Empirical Evaluation. Journal of Machine Learning Research. 2010. no. 11. pp. 171-234.
10. Aliferis C.F., Statnikov A., et al. Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification Part II: Analysis and Extensions. Journal of Machine Learning Research. 2010. no. 11. pp. 235 – 299.
11. Schapire R.E. The Boosting Approach to Machine Learning. An Overview Nonlinear Estimation and Classification. Springer. Lecture Notes in Statistics. Vol. 171. Denison D.D., Hansen M.H, Holmes C.C., Mallick B., Yu B. (Eds). 2003. pp. 149–172.
12. Blake C.L., Murphy P.M. UCI Repository of machine learning database / University of California, Department of Information and Computer Science. Irvine, CA. 1998. Available at: http://www.cs.uci.edu/ mleam/mlrepository.html (Accessed:20.06.2014).
13. Li W., Han J., Pei J. CMAR: Accurate and efficient classification based on multiple class-association rules. Proceedings of the ICDM'01, San Jose, CA, Nov. 2001. pp. 369–376.
14. Wedyan S. Review and Comparison of Associative Classification Data Mining Approaches. International Journal of Computer, Information, Systems and Control Engineering. 2014. vol. 8 no.1. pp. 34–45.
15. Yin X., Han J. CPAR: Classification Based on Predictive Association Rule. Proceedings of the SDM'03. 2003. pp. 369–376.
16. Quinlan J.R. C4.5: Programs for Machine Learning. Morgan Kaufmann, 1993.
17. Quinlan J.R., Cameron-Jones R.M. FOIL: A midterm report. Proceedings of the European Conference on Machine Learning. Vienna, Austria. 1993. pp. 3–20.
18. Ibrahim S., Chandran K.R. Compact Weighted Class Association Rule Mining using Information Gain. International Journal of Data Mining & Knowledge Management Process (IJDKP). 2011. vol.1, no.6. pp. 1–13.
19. Cohen W. Fast effective rule induction. Proceedings of the ICML'95. Tahoe City, CA. 1995. pp. 115–123.

20. Huang Z., Zhou Z., He T., Wang X. ACAC: Associative Classification based on All-Confidence. Proceedings of IEEE International Conference on Granular Computing (GrC). 2011. pp. 289-293.
21. Dong G., Li J. Efficient Mining of Emerging Patterns: Discovering Trends and Differences. Proceedings of the KDD'99. 1999. pp. 43–52.
22. Dong G., Zhang X., Wong L., Li J. CAEP: Classification by Aggregating Emerging Patterns. Proceedings of the DS'99, .1999. pp. 30–42.
23. Fan H., Ramamohanarao K. Fast Discovery and the Generalization of Strong Jumping Emerging Patterns for Building Compact and Accurate Classifiers. IEEE Trans. Knowl. Data Eng. 2006. vol. 18(6). pp. 721-737.
24. Li J., Dong G, Ramamohanarao K. Making use of the most expressive jumping emerging patterns for classification. Proceedings of the Fourth Pacific-Asia Conference on Knowledge Discovery and Data Mining. Kyoto, Japan. 2000. pp. 220-230.

Тушканова Ольга Николаевна — аспирант, Федеральное государственное бюджетное учреждение науки Санкт-Петербургский институт информатики и автоматизации Российской академии наук. Область научных интересов: машинное обучение, интеллектуальный анализ данных, извлечение знаний, многоагентные системы, рекомендующие системы, облачные технологии, онтологии. Число научных публикаций — 12. tushkanova.on@gmail.com; 199178, Санкт-Петербург, 14 линия, д. 39; р.т.: +79817343119.

Tushkanova Olga Nikolaevna — Ph.D. student, St.Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences. Research interests: data mining, multi-agent systems, recommender systems, cloud computing, ontologies, knowledge extraction technologies. The number of publications — 12. tushkanova.on@gmail.com; 39, 14-th Line, St. Petersburg, 199178, Russia; office phone: +79817343119.

Городецкий Владимир Иванович — д-р техн. наук, профессор, заведующий лабораторией интеллектуальных систем, Федеральное государственное бюджетное учреждение науки Санкт-Петербургский институт информатики и автоматизации Российской академии наук. Область научных интересов: искусственный интеллект, технология многоагентных систем, распределенное обучение, извлечение знаний из баз данных, анализ и объединение данных различных источников, P2P сети принятия решений и P2P методы извлечения знаний из данных, обработка больших данных, планирование и составление расписаний, алгоритмы улучшения изображений, рекомендующие системы. Число научных публикаций — 200. gor@mail.iias.spb.su; 199178, Санкт-Петербург, 14 линия, д. 39; р.т.: +7-812-328-3311.

Gorodetsky Vladimir Ivanovich — Ph.D., professor, head of laboratory of intelligent systems, St.Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences. Research interests: intelligent data analysis, information fusion, P2P data mining and machine learning, multi-agent systems technology and software tools, agent-based applications, recommender systems, mobile image enhancement. The number of publications — 200. gor@mail.iias.spb.su; 39, 14-th Line, St. Petersburg, 199178, Russia; office phone: +7-812-328-3311.

РЕФЕРАТ

Городецкий В.И., Тушканова О.Н. **Ассоциативная классификация: аналитический обзор. Часть 1.**

В настоящее время в области работы с большими данными, в частности, в интересах решения задач классификации, развиваются новые методы, модели и алгоритмы интеллектуальной обработки. Одним из таких относительно новых направлений является *ассоциативная классификация*. В данной работе описываются, анализируются и сравниваются начальные результаты, модели и методы, разработанные в области ассоциативной классификации, применительно к работе с данными большого объема и устанавливается их связь с классическими результатами в области индуктивного обучения и методами поиска часто встречающихся паттернов для генерации ассоциативных правил. В работе дается постановка задачи ассоциативной классификации, вводится необходимая терминология и формальные обозначения, используемые в ассоциативной классификации. Приводится описание и сравнительный анализ начальных алгоритмов в области ассоциативной классификации. Дается оценка вклада первых работ, посвящённых ассоциативной классификации, в развитие этого направления, а также и формулируются цели дальнейшего исследования.

SUMMARY

Gorodetsky V., Tushkanova O. **Associative Classification: Analytical Overview. Part 1.**

Currently new methods, models and algorithms for intelligent processing of big data (in particular for solving classification problems) are rapidly developing. One of these areas - associative classification - is relatively new. The paper topic is early methods of associative classification intended for processing of big data. It formulates corresponding problem statement and introduces basic concepts and formal notation used in associative classification. An extended overview and comparative analysis of the basic approaches, models and algorithms developed for associative classification form the main paper contents. The paper assesses the contribution of the first papers devoted to associative classification to the development of this area and formulates goals of the further research.