УДК 519.7

DOI: 10.15827/0236-235X.118.187-195

Дата подачи статьи: 17.03.17

2017. Т. 30. № 2. С. 187–195

СЖАТОЕ ПРЕДСТАВЛЕНИЕ СТРОГИХ АССОЦИАТИВНЫХ ПРАВИЛ В АНАЛИЗЕ ДАННЫХ

В.В. Быкова, д.ф.-м.н., профессор, bykvalen@mail.ru;

А.В. Катаева, аспирант, kataeva_av@mail.ru

(Сибирский федеральный университет,

Свободный просп., 79, г. Красноярск, 660041, Россия)

Современные методы и средства поиска ассоциативных правил в больших массивах данных приводят к значительному множеству правил, многие из которых являются избыточными. Избыточные ассоциативные правила не представляют ценности, но могут вводить в заблуждение. Для решения этой проблемы предложен алгоритм MClose, являющийся модификацией алгоритма Close.

Известно, что с помощью алгоритма Close можно построить минимаксный базис для строгих ассоциативных правил (правил с достоверностью 1). Минимаксный базис интересен для экспертов тем, что каждое входящее в него правило имеет минимальную посылку и максимальное следствие. Однако минимаксный базис может содержать избыточные ассоциативные правила. Алгоритм MClose позволяет в процессе построения минимаксного базиса устранять избыточные правила. Результирующий базис назван сжатым строгим базисом. Удаленные ассоциативные правила всегда можно получить из сжатого строгого базиса с сохранением их поддержки и достоверности без обращений к анализируемому массиву данных. Алгоритм MClose основан на соответствиях Галуа и выводимостях, подобных аксиомам Амстронга, которые используются в теории реляционных БД для функциональных зависимостей.

Как показали вычислительные эксперименты, алгоритм MClose по времени работы сопоставим с алгоритмом Close. Однако он примерно в два раза уменьшает число ассоциативных правил минимаксного базиса. В работе дано описание программы, в которой представлены алгоритмы MClose и Close.

Ключевые слова: анализ данных, соответствия Галуа, замкнутые множества, ассоциативные правила, избыточность, сжатый строгий базис.

Интеллектуальный анализ данных направлен на исследование больших объемов информации с целью выявления зависимостей между данными. Ассоциативные правила (*association rule*) – один из хорошо изученных классов зависимостей, которые отражают, какие признаки или события встречаются совместно и насколько часто это происходит. В настоящее время ассоциативные правила нашли широкое применение в медицине, информационной безопасности, анализе компьютерных сетей и маркетинге при решении задач диагностики и прогнозирования [1–3].

При поиске ассоциативных правил анализируемое множество данных обычно описывается бинарным контекстом – матрицей, строки которой соответствуют объектам рассматриваемой предметной области, а столбцы – признакам этих объектов. Единичное значение элемента матрицы трактуется как наличие у объекта соответствующего признака, а нулевое – как его отсутствие. Бинарное представление данных значительно расширяет математический аппарат для их исследования. Современные методы поиска ассоциативных правил базируются преимущественно на анализе формальных понятий и теории вероятностей [4, 5]. Анализ формальных понятий, как прикладная ветвь алгебраической теории решеток, является удобным математическим аппаратом описания методов поиска ассоциативных правил [6–8].

Главная проблема при поиске ассоциативных правил – огромное число правил, возникающих при исследовании больших бинарных контекстов.

Это существенно усложняет экспертный анализ выявленных ассоциативных правил. Для решения этой проблемы используются различные меры значимости, такие как поддержка (*support*) и достоверность (*confidence*) [9]. С их помощью найденные ассоциативные правила фильтруются, и для анализа предъявляются только те, для которых значения мер значимости превышают заданные пороговые значения. Подобная фильтрация, конечно, уменьшает число правил, но не решает проблему размерности полностью. Не помогает и расширение мер значимости [10]. Часто после фильтрации все равно остается значительное число ассоциативных правил, при этом многие из них избыточные. Ассоциативное правило считается избыточным, если его удаление из множества правил не приводит к потере информации о связях между данными в рассматриваемой предметной области.

Существуют различные формальные определения избыточных ассоциативных правил и методы их устранения [11]. Наиболее развиты методы устранения избыточности для строгих ассоциативных правил (*strong association rules*). Такие правила имеют достоверность 1 и считаются самыми важными. Для этих правил имеются плотная параллель с функциональными зависимостями из теории реляционных БД, а также четкое понятие избыточности [12]. Строгие ассоциативные правила представляют интерес для тех приложений, где требуется высокая степень уверенности в обнаруженных зависимостях между данными, например, в медицине и информационной безопасности [1–3].

Множество ассоциативных правил, не содержащее избыточных (в некотором смысле) правил, принято называть базисом. В настоящее время известен ряд алгоритмов, позволяющих строить различные базисы для строгих ассоциативных правил. Наиболее значимыми являются канонический и минимаксный базисы. Канонический базис (или базис Дюкена–Гига) состоит из минимального числа ассоциативных правил, рекуррентно описываемых в терминах псевдосодержаний [13]. Канонический базис глубоко исследован в анализе формальных понятий, однако все предложенные на сегодняшний день алгоритмы его построения в большей степени представляют теоретический, чем практический интерес [14]. Минимаксный базис состоит из строгих ассоциативных правил, имеющих минимальную посылку и максимальное следствие [15]. Для построения минимаксного базиса существует ряд хорошо апробированных практикой алгоритмов. К ним относится, например, алгоритм Close [11, 16]. Исследования показали, что в построенных канонических и минимаксных базисах остается некоторая избыточность, которая может быть устранена на основе выводимостей, подобных аксиомам Амстронга, известным в теории реляционных БД для функциональных зависимостей.

В данной статье для строгих ассоциативных правил предложен алгоритм MClose – модификация известного алгоритма Close. Алгоритм MClose позволяет в процессе построения минимаксного базиса устранять избыточность с сохранением поддержки и достоверности без дополнительного обращения к исходному бинарному контексту. Результирующий базис назван сжатым строгим базисом. Введено понятие избыточного ассоциативного правила. Приведены выводимости, доказывающие корректность предлагаемого алгоритма, а также результаты вычислительных экспериментов. Поскольку алгоритм Close основан на анализе формальных понятий, необходимые определения и обозначения этой теории возьмем из работ [6–8].

Основные определения и обозначения анализа формальных понятий

Пусть для некоторой предметной области определены два непустых конечных множества G и M , объектов и признаков соответственно. Предполагаем, что все объекты в G и признаки в M различны. Пусть задано отношение $I \subseteq G \times M$ инцидентности между множествами G и M . Существование в I пары (g, m) , $g \in G$ и $m \in M$, означает, что объект g имеет признак m и, наоборот, признак m характерен для объекта g . Тройку $K = (G, M, I)$ принято называть контекстом предметной области.

Выберем в $K = (G, M, I)$ два любых элемента $g \in G$ и $m \in M$. Определим для них два отображения φ и ψ :

$\varphi(g) = \{m \in M \mid (g, m) \in I\}$ – множество признаков, присущих объекту g ;

$\psi(m) = \{g \in G \mid (g, m) \in I\}$ – множество объектов, обладающих признаком m .

Отображения φ и ψ обобщаются на произвольные множества $A \subseteq G$ и $B \subseteq M$ следующим образом:

$$\varphi(A) = \bigcap_{g \in A} \varphi(g) = \{m \in M \mid \forall g \in A (g, m) \in I\},$$

$$\psi(B) = \bigcap_{m \in B} \psi(m) = \{g \in G \mid \forall m \in B (g, m) \in I\}.$$

Следовательно, $\varphi(A)$ – множество признаков, общих для всех объектов из A , а $\psi(B)$ – множество объектов, которые обладают всеми признаками из B . Полагаем, что $\varphi(\emptyset) = M$ и $\psi(\emptyset) = G$: пустому множеству объектов присущи все признаки из M и каждый объект рассматриваемого контекста $K = (G, M, I)$ обладает пустым множеством признаков.

Обычно в анализе формальных понятий для отображений φ и ψ применяется единое обозначение $(\cdot)'$, а приведенные выше формулы для $\varphi(A)$, $\psi(B)$ записываются так:

$$A' = \bigcap_{g \in A} g' = \{m \in M \mid \forall g \in A (g, m) \in I\}, \quad (1)$$

$$B' = \bigcap_{m \in B} m' = \{g \in G \mid \forall m \in B (g, m) \in I\}. \quad (2)$$

Если $g \in G$ и $m \in M$, то обозначения g' и m' служат сокращенной формой записи множеств $\varphi(g) = \{g\}'$ и $\psi(m) = \{m\}'$ соответственно.

Отображения «'» удовлетворяют свойствам, вытекающим из их определения и вполне реалистичного и постулируемого в анализе данных положения: расширение (сокращение) множества признаков уменьшает (увеличивает) число объектов, обладающих этими признаками. Формально эти свойства можно выразить в виде следующих утверждений.

Утверждение 1. Для всякого контекста $K = (G, M, I)$ и любых $B_1, B_2 \subseteq M$ верны свойства:

– *антимонотонность*:

если $B_1 \subseteq B_2$, то $(B_2)' \subseteq (B_1)'$;

– *экстенсивность*:

$B_1 \subseteq (B_1)''$, где $(B_1)'' = ((B_1)')' \subseteq M$.

Множество $(B_1)'' = \varphi(\psi(B_1))$ можно трактовать как набор признаков, которые всегда появляются в объектах контекста $K = (G, M, I)$ вместе с признаками из B_1 , причем это множество является наибольшим по включению в пределах $K = (G, M, I)$. Очевидно, что $(\emptyset)'' = \varphi(\psi(\emptyset)) = G'$, где G' – множество признаков, свойственных всем объектам из G . Для подмножеств множества G справедливо утверждение 2, аналогичное утверждению 1.

Утверждение 2. Для всякого контекста $K = (G, M, I)$ и любых $A_1, A_2 \subseteq G$ верны свойства:

– *антимонотонность*:

если $A_1 \subseteq A_2$, то $(A_2)' \subseteq (A_1)'$;

– *экстенсивность*:

$A_1 \subseteq (A_1)''$, где $(A_1)'' = ((A_1)')' \subseteq G$.

Множество $(A_1)'' = \psi(\varphi(A_1))$ можно интерпретировать как наибольшее по включению множество

объектов, обладающих всеми признаками, которые имеют объекты из A_1 .

Заметим, что, согласно утверждениям 1 и 2, отображения φ и ψ составляют пару соответствий Галуа между множествами 2^G и 2^M , частично упорядоченными по включению [6, 7]. Здесь традиционно 2^G и 2^M – совокупность всех подмножеств рассматриваемых множеств G и M соответственно. Известно, что для соответствий Галуа [7] φ и ψ справедливо равенство

$$\varphi(\psi(\varphi(A))) = \varphi(A), \varphi(\psi(\psi(B))) = \psi(B),$$

или в единых обозначениях

$$((A')') = (A')' = A', ((B')') = (B')' = B'.$$

Двойное применение отображения «'» определяет оператор замыкания на 2^M в алгебраическом смысле [8]. Ему присущи

– *рефлексивность*:

для любого $B \subseteq M$ всегда $B \subseteq B''$;

– *монотонность*:

если $B_1 \subseteq B_2 \subseteq M$, то $(B_1)'' \subseteq (B_2)'' \subseteq M$;

– *идемпотентность*:

для любого $B \subseteq M$ всегда $(B'')'' = B''$.

Справедливость этих свойств вытекает из утверждений 1 и 2. Множество признаков $B \subseteq M$, для которого $B = B''$, называется замкнутым относительно оператора «''» в контексте $K = (G, M, I)$. Принято говорить, что множество B'' является замыканием для $B \subseteq M$ в контексте $K = (G, M, I)$. Заметим, что, исходя из (1) и (2), при $B' \neq \emptyset$ замыкание для $B \subseteq M$ можно вычислить по формуле

$$B'' = \bigcap_{g \in G} \{g' \mid B \subseteq g'\}. \quad (3)$$

Если $B' = \emptyset$, то всегда $B'' = \varphi(\psi(B)) = \varphi(\emptyset) = M$.

Важно отметить, что применение формулы (3) позволяет найти для заданного множества признаков замыкание за один просмотр контекста $K = (G, M, I)$.

Ассоциативные правила и основные меры их значимости

Определим ассоциативное правило, используя приведенные выше понятия и обозначения. Ассоциативным правилом на множестве признаков контекста $K = (G, M, I)$ называется упорядоченная пара $r = (X, Y)$, $X, Y \subseteq M$.

Принято ассоциативное правило $r = (X, Y)$ записывать в виде $X \Rightarrow Y$. В ассоциативном правиле $X \Rightarrow Y$ множества X и Y называют посылкой (или причиной) и заключением (или следствием) соответственно [9]. В анализе ассоциативных правил часто полагают, что посылка и заключение – непустые непересекающиеся множества. С формальных позиций эти ограничения несущественны.

Применительно к заданному контексту $K = (G, M, I)$ всякое ассоциативное правило $X \Rightarrow Y$ количественно характеризуется с помощью двух числовых функций: $\delta(X \Rightarrow Y)$ – поддержка, $\gamma(X \Rightarrow Y)$ – достоверность [10]. Поддержка и достоверность ассо-

циативного правила определяются через понятие поддержки множества признаков.

Поддержка $\delta(X)$ множества признаков $X \subseteq M$ в контексте $K = (G, M, I)$ – отношение числа объектов, которым присущи признаки X , к общему числу объектов, представленных в этом контексте:

$$\delta(X) = |X'| / |G|. \quad (4)$$

Величину $\delta(X)$ можно интерпретировать как частоту встречаемости в контексте $K = (G, M, I)$ объектов, имеющих признаки X . Из формулы (4) следует, что для любого $X \subseteq M$ значение $\delta(X)$ неизменно находится в естественных границах $0 \leq \delta(X) \leq 1$. Чем ближе значение $\delta(X)$ к 1, тем большее число объектов рассматриваемого контекста обладает всеми признаками из X . В силу антимонотонности отображения «'» поддержка множества признаков также удовлетворяет свойству антимонотонности: для всякого контекста $K = (G, M, I)$ и любых $X, Y \subseteq M$ при $X \subseteq Y$ верно неравенство

$$\delta(Y) \leq \delta(X). \quad (5)$$

Согласно (5), поддержка множества признаков не может превышать поддержку любого из его подмножеств. Так, для произвольного $X \subseteq M$ всегда $\delta(M) \leq \delta(X) \leq \delta(\emptyset) = 1$.

Множество признаков $X \subseteq M$ называется частым в контексте $K = (G, M, I)$, если его поддержка больше или равна заданному пороговому значению $\delta_0 \in [0, 1]$. Если $\delta(X) \geq \delta_0$ и $X = X''$, то X называется частым замкнутым множеством признаков в $K = (G, M, I)$. Частые множества и частые замкнутые множества признаков традиционно служат основой для поиска ассоциативных правил в заданном контексте.

Поддержкой $\delta(X \Rightarrow Y)$ ассоциативного правила $X \Rightarrow Y$ относительно контекста $K = (G, M, I)$ называется величина

$$\delta(X \Rightarrow Y) = \delta(X \cup Y) = |(X \cup Y)'| / |G|, \quad (6)$$

указывающая, какая доля объектов этого контекста имеет признаки $X \cup Y$. Достоверность $\gamma(X \Rightarrow Y)$ ассоциативного правила $X \Rightarrow Y$ относительно контекста $K = (G, M, I)$ определяется как отношение числа объектов, обладающих всеми признаками из $X \cup Y$, к числу объектов, которым свойственны только признаки X : $\gamma(X \Rightarrow Y) = |(X \cup Y)'| / |X'|$.

Достоверность ассоциативного правила через функцию поддержки выражается формулой

$$\gamma(X \Rightarrow Y) = \delta(X \Rightarrow Y) / \delta(X) = \delta(X \cup Y) / \delta(X). \quad (7)$$

Заметим, что достоверность определяется формулой (7) только для тех ассоциативных правил $X \Rightarrow Y$, для которых $\delta(X) \neq 0$. Если $\delta(X) = 0$ (в контексте нет ни одного объекта, который обладал бы признаками X), то также $\delta(X \cup Y) = 0$. В этом особом случае полагают $\gamma(X \Rightarrow Y) = 1$. Исходя из (4)–(7), достоверность ассоциативного правила $X \Rightarrow Y$ при произвольных $X, Y \subseteq M$ всегда находится в границах $0 \leq \gamma(X \Rightarrow Y) \leq 1$. Чем ближе значение $\gamma(X \Rightarrow Y)$ к 1, тем с большей уверенностью

можно сказать, что признаки Y появляются в объектах рассматриваемого контекста вместе с признаками X .

Ассоциативное правило $X \Rightarrow Y$ называется минимаксным, если не существует другого ассоциативного правила $X^* \Rightarrow Y^*$, такого, что $X^* \subseteq X$, $Y \subseteq Y^*$ и $\delta(X^* \Rightarrow Y^*) = \delta(X \Rightarrow Y)$, $\gamma(X^* \Rightarrow Y^*) = \gamma(X \Rightarrow Y)$.

В анализе ассоциативных правил часто достоверность ассоциативного правила называют величиной его значимости. Самыми значимыми считаются строгие ассоциативные правила – правила с достоверностью 1. Известен критерий наличия в контексте строгого ассоциативного правила: достоверность ассоциативного правила $X \Rightarrow Y$ относительно контекста $K = (G, M, I)$ равна 1 тогда и только тогда, когда $Y \subseteq X''$. Заметим, что данный критерий тривиальным образом выполняется для $X \Rightarrow X''$.

Из определения замыкания и формулы (3) следует, что для всякого контекста $K = (G, M, I)$ и любого множества $X \subseteq M$ поддержка X'' относительно $K = (G, M, I)$ совпадает с поддержкой X :

$$\delta(X'') = \delta(X). \quad (8)$$

Таким образом, если $X \subseteq M$ является частым множеством в $K = (G, M, I)$, то X'' также является частым в этом контексте. Следует отметить, что в худшем случае число частых замкнутых множеств признаков контекста $K = (G, M, I)$ совпадает с числом частых множеств признаков и экспоненциально зависит от $|M|$. Однако на практике обычно число частых замкнутых множеств значительно меньше числа частых множеств. Примеры контекстов с полиномиальным относительно $|M|$ числом частых замкнутых множеств можно найти в [16]. Использование формулы (8) позволяет при поиске ассоциативных правил вместо частых множеств признаков применять частые замкнутые множества и тем самым сокращать пространство поиска ассоциативных правил. В силу (3) и (8) расширение исходного контекста путем добавления в него новых объектов не изменяет замыкания $X'' \subseteq M$, но может расширить состав объектов, которым присущи все признаки из X'' . Это означает, что при такой трансформации исходного контекста возможно лишь увеличение поддержек для ранее найденных замкнутых множеств признаков. Поэтому, если эти множества были частыми, они по-прежнему остаются частыми.

Задача поиска ассоциативных правил и известные алгоритмы ее решения

Пусть заданы контекст $K = (G, M, I)$ и δ_0, γ_0 – вещественные числа из $[0, 1]$. Будем говорить, что $X \Rightarrow Y$ является (δ_0, γ_0) -ассоциативным правилом в $K = (G, M, I)$, если выполняются два условия:

$$\gamma_0 \leq \gamma(X \Rightarrow Y) \leq 1, \quad (9)$$

$$\delta_0 \leq \delta(X \Rightarrow Y) \leq 1. \quad (10)$$

Требуется найти для заданного контекста $K = (G, M, I)$ множество AR всех (δ_0, γ_0) -ассоциативных правил. Заметим, что искомым набор правил AR параметризован относительно пороговых значений δ_0 и γ_0 . Например, при $\delta_0 = 0$ условие (10) отражает естественные границы поддержки. Данная ситуация свидетельствует о том, что нет ограничений на частоту появления признаков $X \cup Y$ в $K = (G, M, I)$. При $\gamma_0 = 1$ условие (9) приводит к равенству $\gamma(X \Rightarrow Y) = 1$. В этом случае будем иметь строгие ассоциативные правила.

Решение поставленной задачи предполагает выполнение следующих двух этапов: вначале построение различных подмножеств множества M , вычисление для каждого из них поддержки и проверка условия (10); далее генерация ассоциативных правил с учетом найденных частых наборов и проверка условия (9). Наличие на обоих этапах комбинаторных переборов приводит к значительным вычислительным затратам для нахождения решения данной задачи. Очевидно, что, чем больше пороговые значения, тем быстрее находится соответствующий набор ассоциативных правил и тем меньшее число правил будет содержать этот набор. К сожалению, в худшем случае время поиска всех (δ_0, γ_0) -ассоциативных правил в заданном контексте $K = (G, M, I)$ экспоненциально зависит от $|M|$.

К настоящему времени разработано большое число алгоритмов поиска ассоциативных правил. Их обзор представлен в работах [15, 16]. Основополагающими среди них являются алгоритмы Apriori и Close. Алгоритм Apriori использует свойство антимонотонности функции поддержки [3, 9]. Он генерирует (δ_0, γ_0) -ассоциативные правила для любых δ_0 и γ_0 исходя из найденных частых множеств признаков. Алгоритм Close представляет класс алгоритмов, генерирующих только $(\delta_0, 1)$ -ассоциативные правила, то есть строгие ассоциативные правила. Этот алгоритм порождает строгие ассоциативные правила из частых замкнутых наборов признаков [15]. Переход от частых множеств к частым замкнутым множествам позволяет сузить пространство поиска. Другим достоинством алгоритма Close и его многочисленных версий является способность формировать минимаксные ассоциативные правила.

Пример 1. Рассмотрим контекст $K = (G, M, I)$, представленный в таблицах 1 и 2, где $G = \{g_1, g_2, g_3, g_4, g_5\}$ – множество объектов, $M = \{a, b, c, d, e\}$ – множество признаков, I – матрица инцидентности. Заметим, что именно этот контекст традиционно используется в публикациях по алгоритмам поиска ассоциативных правил для демонстрации и тестирования этих алгоритмов [9]. В таблицах 1, 2 и далее при написании множеств для краткости опущены фигурные скобки, запятые между элемен-

тами этих множеств и элементы расположены в лексикографическом порядке. Например, вместо $\{a, b, c\}$ записано abc , а вместо $\{a, b\} \Rightarrow \{c\} - ab \Rightarrow c$.

Таблица 1

Исходный контекст

Table 1

Input context

Объект	Признаки, присущие объектам
g_1	acd
g_2	bce
g_3	$abce$
g_4	be
g_5	$abce$

Таблица 2

Матрица инцидентности исходного контекста

Table 2

The incidence matrix of the input context

	a	b	c	d	e
g_1	1	0	1	1	0
g_2	0	1	1	0	1
g_3	1	1	1	0	1
g_4	0	1	0	0	1
g_5	1	1	1	0	1

Для рассматриваемого контекста при $\delta_0 = 1/5$ и $\gamma_0 = 1$ алгоритм Apriori приводит к семнадцати строгим ассоциативным правилам:

$AR = \{a \Rightarrow c, d \Rightarrow a, e \Rightarrow b, b \Rightarrow e, d \Rightarrow c, ab \Rightarrow c, ae \Rightarrow b, ab \Rightarrow e, cd \Rightarrow a, ad \Rightarrow c, d \Rightarrow ac, ae \Rightarrow c, ce \Rightarrow b, bc \Rightarrow e, ace \Rightarrow b, abe \Rightarrow c, abc \Rightarrow e\}$. (11)

Полученное множество AR содержит много избыточных правил в том смысле, что, если их исключить, то они выводимы из оставшихся правил. Это следующие ассоциативные правила: $\{d \Rightarrow c, ab \Rightarrow c, ae \Rightarrow b, ab \Rightarrow e, cd \Rightarrow a, ad \Rightarrow c, d \Rightarrow ac, ae \Rightarrow c, ce \Rightarrow b, bc \Rightarrow e, ace \Rightarrow b, abe \Rightarrow c, abc \Rightarrow e\}$.

Некоторые из них удаляются алгоритмом Close в процессе построения множества AR . В результате алгоритм Close извлекает лишь восемь минимаксных строгих ассоциативных правил:

$AR = \{a \Rightarrow c, b \Rightarrow e, e \Rightarrow b, d \Rightarrow ac, bc \Rightarrow e, ce \Rightarrow b, ab \Rightarrow ce, ae \Rightarrow bc\}$, (12)

что в два раза меньше чем в (11). Алгоритм Close находит минимаксные строгие ассоциативные правила по следующим частым замкнутым наборам признаков: $c, ac, be, bce, acd, abce$. Все эти правила имеют допустимые поддержки: $\delta(a \Rightarrow c) = 3/5$, $\delta(b \Rightarrow e) = 4/5$, $\delta(e \Rightarrow b) = 4/5$, $\delta(d \Rightarrow ac) = 1/5$, $\delta(bc \Rightarrow e) = 3/5$, $\delta(ce \Rightarrow b) = 3/5$, $\delta(ab \Rightarrow ce) = 2/5$, $\delta(ae \Rightarrow bc) = 2/5$.

Классический алгоритм Close

Суть классического алгоритма Close заключается в пошаговом извлечении генераторов и частных

замкнутых наборов признаков [11, 15]. Множество $\rho \subseteq M$ называется генератором замкнутого набора признаков $X \subseteq M$, $X = X''$, тогда и только тогда, когда $\rho'' = X$ и не существует другого множества $\tau \subseteq M$, такого, что $\tau \subset \rho$ и $\tau'' = X$. Другими словами, генератор замкнутого набора признаков X – наименьшее по мощности множество признаков, имеющее замыкание X . Число признаков, входящих в генератор ρ , называется мощностью этого генератора. Если $|\rho| = k$, то ρ является k -генератором.

На вход алгоритма Close подаются исходный контекст $K = (G, M, I)$ и пороговое значение δ_0 . Изначально множество AR считается пустым и $k = 1$. На первом шаге в качестве k -генераторов рассматриваются все одноэлементные подмножества множества M . Замыкание ρ_k'' для генератора ρ_k вычисляется по формуле (3). Поддержка для ρ_k'' находится по формуле (4). Если $\delta(\rho_k'') \geq \delta_0$, то по частому замкнутому множеству ρ_k'' строится минимаксное строгое ассоциативное правило

$$\rho_k \Rightarrow \rho_k'' \setminus \rho_k \quad (13)$$

и сохраняется в AR . Согласно (6) и (8), для него $\delta(\rho_k \Rightarrow \rho_k'' \setminus \rho_k) = \delta(\rho_k'') \geq \delta_0$, $\gamma(\rho_k \Rightarrow \rho_k'' \setminus \rho_k) = 1$. То, что ассоциативное правило (13) является минимаксным, следует из определения генератора. Очевидно, что различные генераторы могут порождать одинаковые замыкания. Так, для контекста из примера 1 имеем: $b'' = be$, $e'' = be$. Поэтому (12) содержит минимаксные строгие ассоциативные правила $b \Rightarrow e$ и $e \Rightarrow b$.

После генерации ассоциативного правила по ρ_k'' создаются кандидаты в $(k+1)$ -генераторы для следующей итерации. Каждый такой кандидат формируется путем объединения двух k -генераторов, обладающих одинаковыми первыми $k-1$ признаками. Далее проверяется, вложен ли найденный кандидат в ρ_k'' . Если вложен, то он исключается из рассмотрения. После нахождения всех $(k+1)$ -генераторов осуществляется переход к следующей итерации. Алгоритм Close завершает работу, когда исчерпаны все генераторы.

Множество ассоциативных правил, полученных в результате работы алгоритма Close, образует минимаксный базис строгих ассоциативных правил контекста $K = (G, M, I)$. Корректность алгоритма Close доказана в работе [15]. Алгоритм Close может быть модифицирован так, чтобы в результирующее множество AR не попадали заведомо избыточные строгие ассоциативные правила.

Алгоритм MClose для построения сжатого строгого базиса

Верна следующая теорема.

Для любого контекста $K = (G, M, I)$ и произвольных $X, Y, Z, W \subseteq M$ справедливы следующие свойства строгих ассоциативных правил:

D_1 . Рефлексивность: $X \Rightarrow X$.

D_2 . Пополнение посылки:

если $X \Rightarrow Y$, то $X \cup Z \Rightarrow Y$.

D_3 . Аддитивность:

если $X \Rightarrow Y$ и $X \Rightarrow Z$, то $X \Rightarrow Y \cup Z$.

D_4 . Проективность:

если $X \Rightarrow Y$ и $Z \subseteq Y$, то $X \Rightarrow Z$.

D_5 . Транзитивность:

если $X \Rightarrow Y$ и $Y \Rightarrow W$, то $X \Rightarrow W$.

D_6 . Псевдотранзитивность:

если $X \Rightarrow Y$ и $Y \cup Z \Rightarrow W$, то $X \cup Z \Rightarrow W$.

Свойства (или выводимости) D_1 – D_6 позволяют из некоторого множества строгих ассоциативных правил вывести многие другие правила без дополнительного сканирования контекста. С одной стороны, именно выводимости D_1 – D_6 обуславливают экспоненциальное число возможных строгих ассоциативных правил для рассматриваемого контекста и неоднозначность их представления. С другой стороны, они предоставляют возможность построить для множества строгих ассоциативных правил сжатый строгий базис (*concise strong basis*), компактно описывающий связи между данными изучаемой предметной области.

Примечательно, что выводимости, подобные D_1 – D_6 , справедливы для функциональных зависимостей, имеющих место в теории реляционных БД, где их принято называть аксиомами Амстронга. Такое совпадение не случайно, поскольку строгие ассоциативные правила можно рассматривать как особый случай функциональных зависимостей [5]. В работе [12] выводимости D_1 – D_6 доказаны применительно к функциональным зависимостям. Однако они могут быть доказаны непосредственно для ассоциативных правил на основе соответствий Галуа. Важно, что выводимости D_1 , D_3 , D_4 , D_5 гарантируют сохранение поддержки. Это означает, что результатом применения их $(\delta_0, 1)$ -ассоциативным правилом всегда является $(\delta_0, 1)$ -ассоциативное правило. Поэтому эти выводимости применяются в алгоритме MClose для построения сжатого строгого базиса. Введем определения избыточного ассоциативного правила и сжатого строгого базиса для множества $(\delta_0, 1)$ -ассоциативных правил.

Пусть AR – множество строгих ассоциативных правил, каждое из которых справедливо для контекста $K = (G, M, I)$. Будем говорить, что строгое ассоциативное правило $X \Rightarrow Y$ логически следует из множества AR , если оно может быть выведено из AR с помощью выводимостей D_1 , D_3 , D_4 , D_5 . Тот факт, что строгое ассоциативное правило $X \Rightarrow Y$ логически следует из AR , будем обозначать так: $AR \vdash X \Rightarrow Y$.

Строгое ассоциативное правило $X \Rightarrow Y$ называется избыточным в AR , если $AR \setminus \{X \Rightarrow Y\} \vdash X \Rightarrow Y$. Множество строгих ассоциативных правил неизбыточное, если оно не содержит избыточных стро-

гих ассоциативных правил. Множество CSB называется сжатым строгим базисом множества AR , если оно неизбыточное и состоит только из минимаксных строгих ассоциативных правил. Данное определение сжатого строгого базиса указывает способ его нахождения – генерация минимаксных строгих ассоциативных правил (например, с помощью алгоритма Close) и устранение среди них избыточных.

Распознавание избыточного строгого ассоциативного правила в AR основано на проверке логического следования

$$AR \setminus \{X \Rightarrow Y\} \not\vdash X \Rightarrow Y. \quad (14)$$

Алгоритм такой проверки использует понятие замыкания множества признаков относительно множества AR и является полиномиальным относительно $|M|$ и $|AR|$. Замыканием множества $X \subseteq M$ относительно AR (обозначается X^+) называется множество всех признаков $m \in M$, таких, что верно логическое следование $AR \vdash X \Rightarrow m$. Заметим, что неизменно $X^+ \subseteq M$. Из выводимостей D_1 , D_3 , D_4 вытекает справедливость следующего утверждения.

Утверждение 3. Логическое следование $AR \vdash X \Rightarrow Y$ имеет место тогда и только тогда, когда $Y \subseteq X^+$.

Отсюда всегда $AR \vdash X \Rightarrow X^+$, $AR \not\vdash X \Rightarrow X^+ \setminus X$. В силу утверждения 3, чтобы убедиться в справедливости (14), достаточно вычислить X^+ относительно $AR \setminus \{X \Rightarrow Y\}$ и проверить включение $Y \subseteq X^+$. Если $Y \subseteq X^+$, то строгое ассоциативное правило $X \Rightarrow Y$ избыточно в AR , иначе – не избыточно.

Алгоритм вычисления X^+ целиком базируется на выводимостях D_1 , D_3 , D_4 , D_5 и сводится к выполнению следующих действий. Сначала полагается $X^+ = X$. Далее осуществляются просмотр правил из AR и пополнение замыкания по следующему принципу: если для правила $Y \Rightarrow Z \in AR$ верно включение $Y \subseteq X^+$, то множество Z добавляется к X^+ . Этот процесс повторяется до тех пор, пока изменяется X^+ . Поскольку множества M и AR конечные, то процесс вычисления X^+ конечен.

Заметим, что процесс исключения избыточных строгих ассоциативных правил не требует доступа к контексту $K = (G, M, I)$ и поэтому время его выполнения незначительно по сравнению со временем получения частых замкнутых множеств признаков. Чтобы исключить добавление заведомо избыточного строгого ассоциативного правила в AR , необходимо всякий раз после построения ρ_k'' выполнять следующие действия. Если посылка ρ_k найденного правила не равна ρ_k'' , то необходимо найти замыкание ρ_k^+ относительно вычисленного множества AR . Если $\rho_k^+ = \rho_k''$, то минимаксное ассоциативное правило $\rho_k \Rightarrow \rho_k'' / \rho_k$ является избыточным (по утверждению 3), иначе оно включается в AR . После завершения генерации минимаксных строгих ассоциативных правил обязательно требу-

ется дополнительный просмотр результирующего множества AR с целью обнаружения оставшихся избыточных правил. Такие правила вполне возможны: они избыточные по отношению к ранее выявленным правилам, однако после пополнения AR новыми правилами могут оказаться избыточными. Таким образом, по построению результирующее множество AR состоит из минимаксных строгих ассоциативных правил и является избыточным. Заметим, что оперативное удаление избыточных правил сдерживает рост мощности AR и снижает время выполнения алгоритма. Опишем алгоритм MClose.

Алгоритм 1. MClose

```

1: begin
2:    $AR \leftarrow \emptyset$ 
3:    $k \leftarrow 1$ 
4:   while  $\rho_k \neq \emptyset$ 
5:     Gen-Closure ( $\rho_k$ )
6:     if  $\delta(\rho_k) \geq \delta_0$ 
7:       if  $\rho_k \neq \rho_k''$ 
8:          $\rho_k^+ \leftarrow SX(\rho_k)$ 
9:       end if
10:      if  $\rho_k^+ \neq \rho_k''$ 
11:         $AR \leftarrow (\rho_k \Rightarrow \rho_k'' \setminus \rho_k)$ 
12:      end if
13:    end if
14:    Gen-Generator ( $k + 1$ )
15:     $k \leftarrow k + 1$ 
16:  end while
17:  Non-Redundancy ( $AR$ )
18: end

```

В описании алгоритма MClose процедуры Gen-Closure и Gen-Generator аналогичны одноименным процедурам классического алгоритма Close (см. [11, 15]). Процедура SX осуществляет построение замыкания ρ_k^+ . Процедура Non-Redundancy, выполняемая на шаге 17 алгоритма MClose, устраняет в AR избыточные строгие ассоциативные правила после завершения его генерации.

Пример 2. Для контекста, представленного в таблицах 1 и 2, при $\delta_0 = 1/5$ и $\gamma_0 = 1$ минимаксный базис состоит из восьми правил (12), в котором избыточными являются

$$bc \Rightarrow e, ce \Rightarrow b, ab \Rightarrow ce, ae \Rightarrow bc. \quad (15)$$

Сжатый строгий базис, построенный алгоритмом MClose, содержит только четыре минимаксных строгих ассоциативных правила:

$$CSB = \{a \Rightarrow c, b \Rightarrow e, e \Rightarrow b, d \Rightarrow ac\}.$$

Заметим, что на основе CSB можно получить всякое правило из (15) с помощью алгоритма вычисления X^+ . Так, чтобы найти максимальное следствие для посылки bc , достаточно вычислить bc^+ относительно CSB . В результате имеем $bc^+ = bce$. Отсюда $CSB \models bc \Rightarrow e$. Поскольку $bc^+ = bc'' = bce$ и $e \in bc''$, то ассоциативное правило $bc \Rightarrow e$ является строгим.

Программная реализация алгоритма MClose и результаты экспериментов

Алгоритм MClose программно реализован на языке программирования C++ в среде разработки Embarcadero RAD Studio XE8. Исходными данными для него служат контекст $K = (G, M, I)$ и пороговое значение поддержки. Для ввода исходных данных имеется интерфейс (см. http://www.swsys.ru/uploaded/image/2017_2/2017-2-dop/7.jpg). Для удобства пользователя пороговое значение вводится в виде целого положительного числа $\delta_0 \cdot |G|$. Результатом работы алгоритма MClose являются все выявленные частые замкнутые множества признаков и сжатый строгий базис (см. http://www.swsys.ru/uploaded/image/2017_2/2017-2-dop/8.jpg). Значения поддержек масштабируются и выводятся в виде целых положительных чисел.

В программе существует возможность изменения результирующего множества ассоциативных правил путем добавления и удаления отдельных правил. Это сделано для того, чтобы эксперт мог исключить из рассмотрения ассоциативные правила, которые, на его взгляд, не отвечают реальной действительности, и добавить существенные правила. Допускается сокращение признакового пространства (вкладка «Анализ признаков») с помощью методов, описанных в работе [17]. Предусмотрена функция вычисления следствия для всякой заданной посылки относительно сжатого строгого базиса. Для оценки эффективности алгоритма MClose были также программно реализованы алгоритмы Apriori и Close.

Алгоритмы Apriori, Close и MClose сравнивались по числу сгенерированных строгих ассоциативных правил и времени работы. Эксперименты осуществлялись на компьютере с процессором Intel® Core™ i5 CPU & 2.30 GHz и ОЗУ размером 4 Гб. Эксперименты выполнялись на контекстах различной размерности и плотности, сгенерированных случайным образом. Результаты экспериментов представлены в таблице 3.

Таблица 3

Результаты экспериментов

Table 3

Results of experiments

Характеристика контекста $K = (G, M, I)$			Алгоритм					
			Apriori		Close		MClose	
$ G $	$ M $	σ	1	2	1	2	1	2
20	10	0,38	1797	17562	45	250	22	297
30	10	0,39	2029	18347	46	374	19	412
30	10	0,55	15438/	187202	69	390	20	484
50	10	0,53	27769	375178	46	78	13	124
500	10	0,53	27769	376154	42	124	13	168
10000	10	0,53	27769	378400	42	671	13	872

Примечание: в колонках под цифрой 1 – число извлеченных строгих ассоциативных правил, под цифрой 2 – время, мс.

В таблице 3 для всякого анализируемого контекста $K = (G, M, I)$ указаны $|G|$ – число объектов, $|M|$ – число признаков, $\sigma = n / (|G| \cdot |M|)$ – плотность контекста, где n задает число единичных элементов матрицы инцидентности I . Контекст из 10 000 объектов был сформирован многократным копированием контекста, состоящего из 500 объектов.

Из таблицы 3 видно, что алгоритмы Close и MClose эффективнее алгоритма Apriori как по числу извлеченных строгих ассоциативных правил, так и по времени работы. Алгоритм MClose по времени работы сопоставим с алгоритмом Close. Однако алгоритм MClose позволяет более чем в два раза уменьшить мощность минимаксного базиса, формируемого алгоритмом Close.

Заключение

Современные методы и средства поиска ассоциативных правил в больших массивах данных приводят к значительному числу правил, большинство из которых являются избыточными. Усилия многих исследователей направлены на разработку методов устранения избыточности в представлении ассоциативных правил. Для решения этой проблемы применительно к строгим ассоциативным правилам предложен алгоритм MClose, являющийся модификацией известного алгоритма Close и основанный на свойствах частых замкнутых множеств. Предложенный алгоритм формирует для заданного контекста сжатый строгий базис – не избыточное множество минимаксных строгих ассоциативных правил с сохранением поддержки. Вычислительные эксперименты показали, что алгоритм MClose существенно сокращает число извлеченных строгих ассоциативных правил без потери информации о связях между данными анализируемой предметной области. Перспективы исследования, направленные на устранение избыточности для ассоциативных правил с любыми поддержками и достоверностями.

Литература

1. Батура Т.В. Модели и методы анализа компьютерных социальных сетей // Программные продукты и системы. 2013. № 3. С. 130–137.
2. Платонов В.В., Семенов П.О. Методы сокращения размерности в системах обнаружения сетевых атак // Проблемы информационной безопасности. Компьютерные системы. 2012. № 3. С. 40–45.
3. Ilayaraja M., Meyyappan T. Mining medical data to identify frequent diseases using Apriori algorithm. Pattern Recognition, Informatics and Mobile Engineering (PRIME), IEEE, 2013, pp. 194–199.
4. Городецкий В.И., Самойлов В.В. Ассоциативный и причинный анализ и ассоциативные байесовские сети // Тр. СПИИРАН. 2009. Вып. 9. С. 13–65.
5. Кузнецов С.О. Автоматическое обучение на основе анализа формальных понятий // Автоматика и телемеханика. 2001. № 10. С. 3–27.
6. Биркгоф Г., Барти Т. Современная прикладная алгебра. СПб: Лань, 2005. 400 с.
7. Гуров С.И. Булевы алгебры, упорядоченные множества, решетки: определения, свойства, примеры. М.: ЛИБРОКОМ, 2013. 352 с.
8. Ganter B., Wille R. Formal Concept Analyses: mathematical foundations, Springer Science and Business Media, 2012, 314 p.
9. Zhang C., Zhang S. Association rules mining. Springer, 2002, 240 p.
10. Geng L., Hamilton H.J. Interestingness measures for data mining: a survey. ACM Computing Surveys, 2006, vol. 38, no. 3, article 9.
11. Pasquier N., Bastide Y., Taouil R. & Lakhal L. Generating a condensed representation for association rules. Jour. of Intelligent Information Systems, 2005, vol. 24, no. 1, pp. 29–60.
12. Мейер Д. Теория реляционных баз данных. М.: Мир, 1987. 608 с.
13. Duquenne V., Obiedkov S.A. Attribute-incremental construction of the canonical implication basis. Annals of Mathematics and Artificial Intelligence, 2007, vol. 49, no. 1–4, pp. 77–99.
14. Rudolph S. Some notes on pseudo-closed sets. Proc. ICFA 2007, LNCS, Springer, 2007, vol. 4390, pp. 151–165.
15. Zaki M.J., Hsiao Ch.-J. Efficient Algorithms for Mining Closed Itemsets and Their Lattice Structure. IEEE Transaction on Knowledge and Data Engineering, 2005, vol. 17, no. 4, pp. 462–478.
16. Uno T., Asai T., Uchida Y., Arimura H. An Efficient Algorithm for Enumerating Closed Patterns in Transaction Databases. Proc. DS'04, LNAI 3245, 2004, pp. 16–31.
17. Быкова В.В., Катаева А.В. Методы и средства анализа информативности признаков при обработке медицинских данных // Программные продукты и системы. 2016. № 2. С. 172–178.

A CONTRACTED REPRESENTATION OF STRONG ASSOCIATIVE RULES IN DATA ANALYSIS

V.V. Bykova¹, Dr. Sc. (Physics and Mathematics), Professor, bykvalen@mail.ru

A.V. Kataeva¹, Postgraduate Student, kataeva_av@mail.ru

¹ Siberian Federal University, Svobodny Ave. 79, Krasnoyarsk, 660041, Russian Federation

Abstract. Modern methods and means of searching for association rules in big data lead to a significant number of rules, many of which are redundant. Redundant association rules are generally of no value, but they can misinform. To solve this problem, the paper proposes an algorithm MClose, which is a modification of the algorithm Close.

It is known that Close algorithm might help to construct mini-max basis for strict association rules (association rules with the confidence of 1). Mini-max basis consists of only min-max association rules. Association rules with minimal antecedent and maximal consequent are called min-max association rules. Such rules are interesting for experts. However, mini-max basis may contain redundant association rules. The algorithm MClose immediately eliminates redundant association rules when creating mini-max basis. The resulting basis is called concise strong basis (CSB). Redundant association rules might always be obtained from the CSB without sacrificing their support and confidence, without references to the data set. Algorithm MClose is based on Galois connection. MClose algorithm is also based on derivability, which are similar on Armstrong axioms for functional dependencies.

Experiments have shown that running time of algorithm MClose is comparable with the algorithm Close. However, it reduces the number of association rules mini-max basis about twice. We provide a description of the program which presents MClose and Close algorithms.

Keywords: data mining, Galois connection, closed sets, association rules, non-redundancy, concise strong basis.

References

1. Batura T.V. Social networks analysis models and methods. *Programmnye produkty i sistemy* [Software & Systems]. 2013, no. 3, pp. 130–137 (in Russ.).
2. Platonov V.V., Semenov P.O. Dimension reduction in network attacks detection systems. *Problemy informatsionnoy bezopasnosti. Kompyuternye sistemy* [Information Security Problems. Computer Systems]. 2012, no. 3, pp. 40–45 (in Russ.).
3. Ilayaraja M., Meyyappan T. Mining medical data to identify frequent diseases using Apriori algorithm. *Pattern Recognition, Informatics and Mobile Engineering (PRIME), IEEE*, 2013, pp. 194–199.
4. Gorodetsky V.I., Samoylov V.V. Association and casual rule mining using associative bayesian networks. *Tr. SPIIRAN* [SPIIRAS Proc.]. 2009, no. 9, pp. 13–65 (in Russ.).
5. Kuznetsov S.O. Automatic learning using analysis of formal concepts. *Avtomatika i telemekhanika* [Automation and Telemechanics]. 2001, no. 10, pp. 3–27 (in Russ.).
6. Birkhof G., Barti T. *Sovremennaya prikladnaya algebra* [Modern Applied Algebra]. St. Petersburg, Lan Publ., 2005, 400 p.
7. Gurov S.I. *Bulevy algebrы, uporyadochennyye mnozhestva, reshetki: opredeleniya, svoystva, primery* [Boolean algebra, ordered sets, grids: definitions, features, examples]. Moscow, Librokom Publ., 2013, 352 p.
8. Ganter B., Wille R. *Formal Concept Analyses: mathematical foundations*. Springer Science and Business Media Publ., 2012, 314 p.
9. Zhang C., Zhang S. *Association Rules Mining*. Springer Publ., 2002, 240 p.
10. Geng L., Hamilton H.J. Interestingness measures for data mining: a survey. *ACM Computing Surveys*. 2006, vol. 38, no. 3, article 9.
11. Pasquier N., Bastide Y., Taouil R., Lakhal L. Generating a condensed representation for association rules. *Jour. of Intelligent Information Systems*. 2005, vol. 24, no. 1, pp. 29–60.
12. Meyer D. *Teoriya relyatsionnykh baz dannykh* [The Relational Database Theory]. Moscow, Mir Publ., 1987, 608 p.
13. Duquenne V., Obiedkov S.A. Attribute-incremental construction of the canonical implication basis. *Annals of Mathematics and Artificial Intelligence*. 2007, no. 49 (1–4), pp. 77–99.
14. Rudolph S. Some notes on pseudo-closed sets. *Proc. ICFCA 2007*. LNCS, Springer Publ., 2007, vol. 4390, pp. 151–165.
15. Zaki M.J., Hsiao Ch.-J. Efficient Algorithms for Mining Closed Itemsets and Their Lattice Structure. *IEEE Trans. on Knowledge and Data Engineering*. 2005, vol. 17, no. 4, pp. 462–478.
16. Uno T., Asai T., Uchida Y., Arimura H. An Efficient Algorithm for Enumerating Closed Patterns in Transaction Databases. *Proc. DS'04*. LNAI Publ., 2004, no. 3245, pp. 16–31.
17. Bykova V.V., Katayeva A.V. Methods and tools for analysing informative features when processing medical data. *Programmnye produkty i sistemy* [Software & Systems]. 2016, no. 2 (114), pp. 172–178 (in Russ.).

Примеры библиографического описания статьи

1. Быкова В.В., Катаева А.В. Сжатое представление строгих ассоциативных правил в анализе данных // Программные продукты и системы. 2017. Т. 30. № 2. С. 187–195; DOI: 10.15827/0236-235X.118.187-195.
2. Bykova V.V., Kataeva A.V. A contracted representation of strong associative rules in data analysis. *Programmnye produkty i sistemy* [Software & Systems]. 2017, vol. 30, no. 2, pp. 187–195 (in Russ.); DOI: 10.15827/0236-235X.118.187-195.