Вирішено задачу автоматизації видобування чисельних асоціативних правил. Метою роботи було створення методу видобування чисельних асоціативних правил з урахуванням значущості ознак. Запропоновано метод пошуку асоціативних правил, у якому використовується апріорна інформація про значущість ознак, що дозволяє скоротити простір пошуку та час видобування правил, зменшити кількість правил, підвищити інтерпретабельність синтезованої бази правил

Ключові слова: асоціативне правило, база правил, нечітка логіка, транзакція, фаззифікація, функція належності

Решена задача автоматизации извлечения численных ассоциативных правил. Целью работы являлось создание метода извлечения численных ассоциативных правил с учетом значимости признаков. Предложен метод поиска ассоциативных правил, в котором используется априорная информация о значимости признаков, что позволяет сократить пространство поиска и время извлечения правил, уменьшить количество правил, повысить интерпретабельность синтезированной базы правил

Ключевые слова: ассоциативное правило, база правил, нечеткая логика, транзакция, фаззификация, функция принадлежности

УДК 004.93

ИЗВЛЕЧЕНИЕ ЧИСЛЕННЫХ АССОЦИАТИВНЫХ ПРАВИЛ С УЧЕТОМ ЗНАЧИМОСТИ ПРИЗНАКОВ

Т. А. Зайко
Аспирант*
E-mail: tzyakun@mail.ru
А. А. Олейник
Кандидат технических наук, доцент*
E-mail: olejnikaa@gmail.com
С. А. Субботин
Кандидат технических наук, профессор*
E-mail: subbotin@zntu.edu.ua
*Кафедра программных средств
Запорожский национальный технический университет
ул. Жуковского, 64, г. Запорожье, Украина, 69063

1. Введение

Исследование сложных объектов и процессов связано с необходимостью извлечения новых знаний путем обработки больших массивов данных [1]. Для извлечения новых знаний из больших массивов информации при решении задач диагностирования и распознавания образов широко применяются методы и средства интеллектуального анализа данных, эффективным инструментом которого являются ассоциативные правила [1, 2]. Такие правила представляются в виде импликаций $X \to Y$, в которых X и Y являются непересекающимися множествами элементов.

Существующие методы поиска ассоциативных правил [1 – 3], как правило, извлекают бинарные правила, в которых множества X и Y содержат информацию лишь о том, произошел ли какой-то набор событий или нет.

Однако большинство реальных задач диагностирования, распознавания образов и др. связаны с необходимостью обработки не только качественной, но и количественной информации. В таких случаях целесообразным является выделение численных ассоциативных правил [1, 3 – 7], содержащих информацию не только о наличии некоторого набора событий, но и об их численных характеристиках. Предложенные методы извлечения таких правил [3 – 7] связаны с проблемами выбора интервалов дискретизации диапазонов значений переменных, определения количества интервалов разбиений признаков, поскольку неудачное разбиение в некоторых случаях может привести к существенному увеличению пространства поиска и требований к вычислительным ресурсам ЭВМ, а также к недостаточной точности прогнозирования или классификации по синтезированной базе ассоциативных правил [3 – 9]. Кроме того, такие методы предполагают, что каждый признак (элемент транзакции базы данных) имеет одинаковую значимость, что, как правило, на практике не соответствует действительности и приводит к построению баз ассоциативных правил с неприемлемыми аппроксимационными свойствами.

Поэтому актуальной задачей является разработка метода синтеза численных ассоциативных правил, свободного от указанных недостатков.

Целью настоящей работы является создание метода извлечения численных ассоциативных правил с учетом значимости признаков.

2. Постановка задачи синтеза численных ассоциативных правил

Пусть задана база транзакций D:

$$D = \left\{ T_1, T_2, ..., T_{N_D} \right\},$$

в которой каждый элемент T_j, $j = 1,2,...,N_D$ содержит информацию о некоторых взаимосвязанных событиях, где $N_D = |D|$ – количество элементов (транзакций) в наборе данных D.

Элементы T_j могут представляться в виде:

$$T_j = \left(tid_j, item_j\right),$$

где tid_j – идентификатор j-й транзакции T_j; $item_j = \left\{t_{1j}, t_{2j}, ..., t_{N_{item_j} j}\right\} \subseteq I$ – список элементов, входящих в транзакцию T_j; t_{ij} – i-й элемент списка $item_j$, $i = 1,2,...,N_{item_j}$; $N_{item_j} = \left|item_j\right|$ – количество элементов множества $item_j$; $I = \left\{\tau_1, \tau_2, ..., \tau_{N_I}\right\}$ – множество возможных переменных (признаков), которые могут входить в список элементов $item_j$ каждой транзакции T_j, $j = 1,2,...,N_D$ набора данных D; τ_a – a-й элемент множества I, $a = 1,2,...,N_I$; $N_I = |I|$ – количество элементов множества I.

В случае, если база транзакций D содержит кроме бинарных, еще и вещественные переменные, элементы t_{ij} транзакции T_j представляются кортежем:

$$t_{ij} = \left\langle \tau_{ij}; v\left(\tau_{ij}\right)\right\rangle,$$

где τ_{ij} – признак из множества I, соответствующий элементу t_{ij}; $v\left(\tau_{ij}\right)$ – значение признака τ_{ij} в транзакции T_j, $v\left(\tau_{ij}\right) \in \Delta_{ij} = \left[\tau_{ijmin}; \tau_{ijmax}\right]$; τ_{ijmin} и τ_{ijmax} – минимальное и максимальное значения из диапазона возможных значений Δ_{ij} признака τ_{ij}.

Тогда на основе заданной транзакционной базы данных D необходимо построить набор численных ассоциативных правил в виде импликаций $\left\langle X, v(X)\right\rangle \rightarrow \left\langle Y, v(Y)\right\rangle$, в которых наборы X и Y не пересекаются [1, 3]:

$$\left\langle X, v(X)\right\rangle \rightarrow \left\langle Y, v(Y)\right\rangle : X \subset I,\ Y \subset I,\ X \cap Y = \varnothing,$$

где $v(X)$ и $v(Y)$ – множества значений признаков, принадлежащих множествам X и Y, соответственно.

Таким образом, в результате синтеза ассоциативных правил основе имеющегося набор данных D выполняется поиск закономерностей между событиями $\tau_a \in I$, $a = 1,2,...,N_I$.

3. Метод синтеза численных ассоциативных правил

Для возможности извлечения ассоциативных правил из транзакционных баз данных D, содержащих численные атрибуты, такие атрибуты преобразовываются к формату, доступному для применения известных методов поиска ассоциативных правил [1 – 5]. При этом требуется выполнять разбиение численных признаков на непересекающиеся интервалы, каждый из которых рассматривается затем как новый атрибут. Однако в таких случаях возникают проблемы выбора числа интервалов и разбиения на интервалы, кроме того существенно возрастает размерность решаемой задачи и требования к вычислительным ресурсам ЭВМ.

Поэтому в разработанном методе синтеза численных ассоциативных правил предлагается использовать подход на основе теории нечетких множеств [10 – 12], позволяющий разбивать исходные признаки на нечеткие интервалы и работать с каждым признаком, а не с отдельными интервалами его разбиения. Кроме того, в предложенном методе при поиске ассоциативных правил используются рассчитанные оценки индивидуальной информативности признаков, что позволяет учитывать их значимость в исходной базе данных.

Предлагаемый метод может быть представлен следующими этапами:

– фаззификация транзакционной базы данных D;

– определение индивидуальной значимости признаков;

– вычисление пороговых значений поддержки;

– построение базы численных ассоциативных правил.

На начальном этапе выполняется фаззификация базы транзакций D, т.е. приведение всех ее численных значений к нечеткому виду: $D \rightarrow FuzzyD$. Такое преобразование позволит выделить нечеткие термы каждого признака для возможности выполнения дальнейшего извлечения ассоциативных правил. Для фаззификации определяются функции принадлежности μ_a для каждого численного a-го признака $\tau_a \in I$. Функции принадлежности могут быть заданы экспертом, исходя из его знаний и опыта относительно исследуемого объекта или процесса [10 – 15]. Однако использование субъективной информации и некоторых допущений при преобразовании ее в степени принадлежности нечетких множеств в некоторых случаях может привести к неприемлемым результатам такого преобразования, вследствие чего синтезированная база ассоциативных правил не будет содержать интересные правила, а новые знания, выделенные на основе построенной таким образом базы ассоциативных правил, будут необъективно отражать исследуемые объекты или процессы.

Как правило, признаки, описывающие исследуемые объекты или процессы, имеют различную информативность [11, 13, 14], поэтому с целью извлечения интересных ассоциативных правил, адекватно описывающих исследуемые зависимости, целесообразно учитывать индивидуальную значимость признаков. Поскольку выходной параметр в транзакционных базах данных, как правило, не задан, предлагается оценивать индивидуальную значимость признаков с помощью параметров, характеризующих границы областей группирования экземпляров (транзакций) в пространстве признаков.

При этом признаки предварительно нормируются с целью приведения значений всех признаков к одному диапазону, что устранит влияние величины граничных значений признака на его индивидуальную значимость.

В результате кластеризации выделяется $N_{кл.}$ кластеров. Для определения значимости каждого элемента $\tau_a \in I$ будем оценивать его влияние для отнесения транзакции к каждому из кластеров. Очевидно, чем меньше ширина диапазона изменения значений a-го признака во множестве транзакций кластера K_b $(b = 1,2,...,N_{кл.})$, тем более его значимость в данном кластере.

Ширину диапазона будем оценивать как среднеквадратическое отклонение [16]:

$$\sigma_{ab} = \sqrt{\sum_{g=1}^{N_{тр.b}} \left(\overline{\tau_{ab}} - \tau_{abg}\right)^2},$$

где $\overline{\tau_{ab}}$ – среднее значение a-го признака в b-м кластере; τ_{abg} – g-е значение a-го признака в b-м кластере; $N_{тр.ab}$ – количество транзакций в b-м кластере.

Признаку с минимальным значением величины σ_{ab} будем присваивать максимальное значение ранга $Rg_{ab} = |I|$ в b-м кластере, следующему по возрастанию значения σ_{ab} признаку присвоим ранг $Rg_{ab} = |I| - 1$ и т.д. В случае, если признаки имеют одинаковое значение σ_{ab}, им присваиваются одинаковые значения Rg_{ab}. Редко встречающиеся признаки со средним значением в группе $\overline{\tau_{ab}}$, ниже минимально допустимого ($\overline{\tau_{ab}} < \tau_{min}$), считаются неинформативными в данном кластере, вследствие чего им присваивается нулевое значение ранга: $Rg_{ab} = 0$.

Затем для каждого a-го признака τ_a складываются значения рангов по всем кластерам:

$$Rg_a = \sum_{b=1}^{N_{кл.}} Rg_{ab}.$$

Значимость (вес) w_a признака τ_a может определяется следующим образом:

– как отношение ранга Rg_a к сумме рангов всех признаков:

$$w_a = \frac{Rg_a}{\sum_{A=1}^{|I|} Rg_A};$$

– как отношение ранга Rg_a к максимальному значению рангов:

$$w_a = \frac{Rg_a}{\max\limits_{A=1,2,...,|I|} Rg_A}.$$

Кроме предложенного выше подхода можно использовать подход, учитывающий границы интервалов разбиения признаков в кластерах.

В данном методе предлагается сортировать массив значений каждого признака τ_a по возрастанию. Левая l_{ak} и правая r_{ak} границы k-го интервала Δ_{ak} a-го признака τ_a выбираются таким образом, чтобы экземпляры (транзакции) со значением признака $\tau_a \in \Delta_{ak} = [l_{ak}; r_{ak})$ относились к одному кластеру K_b, а экземпляры из соседних интервалов – к другим кластерам $K_c \neq K_b$.

В качестве меры информативности a-го признака в транзакционной базе данных D целесообразно использовать количество интервалов $N_{инт.a}$, на которые разбивается диапазон его значений $\Delta_a = [\tau_{amin}; \tau_{amax}]$: чем меньше количество таких интервалов, тем больше информативность признака.

Поэтому значимость признака τ_a будем вычислять по одной из формул:

– отношение минимального количества интервалов среди всех признаков к величине $N_{инт.a}$ a-го признака:

$$w_a = \frac{\min\limits_{A=1,2,...,|I|} N_{инт.A}}{N_{инт.a}};$$

– нормированное значение величины N_{8-Ba}:

$$w_a = 1 - \frac{N_{инт.a} - \min\limits_{A=1,2,...,|I|} N_{инт.A}}{\max\limits_{A=1,2,...,|I|} N_{инт.A} - \min\limits_{A=1,2,...,|I|} N_{инт.A}} =$$

$$= \frac{\max\limits_{A=1,2,...,|I|} N_{инт.A} - N_{инт.a}}{\max\limits_{A=1,2,...,|I|} N_{инт.A} - \min\limits_{A=1,2,...,|I|} N_{инт.A}}.$$

Предложенный подход позволяет вычислять информативность каждого признака в транзакционной базе данных D, а также выделять интервалы разбиения признаков без необходимости задания количества интервалов разбиений, что уменьшает степень участия пользователя и влияние его субъективных оценок на результаты процесса извлечения ассоциативных правил, что в свою очередь снижает вероятность извлечения ассоциативных правил, некорректно описывающих исследуемые объекты или процессы.

Важным этапом является определение пороговых значений поддержки наборов элементов, которое в предложенном методе происходит с использованием информации об индивидуальной значимости признаков, определенной ранее. Кроме того, предусматривается возможность извлечения наборов, не являющихся часто встречающимися, однако являющихся интересными и позволяющими выявлять новые знания об исследуемых объектах или процессах.

При поиске ассоциативных правил важной характеристикой, используемой в процессе их извлечения, является поддержка наборов элементов, а также ее пороговое значения, задаваемое, как правило, пользователем в качестве параметра метода.

В разработанном методе извлечения численных ассоциативных правил поддержку транзакции T_j будем рассчитывать как пересечение функций принадлежности признаков, входящих в транзакцию T_j:

$$supp(T_j) = \bigcap_{\tau_a \in T_j} \mu_a(T_j),$$

где $\mu_a(T_j)$ – значение функция принадлежности a-го признака, вычисленное для его значения в транзакции T_j.

Тогда поддержка набора X определяется как сумма поддержек всех транзакций, содержащих это множество:

$$supp(X) = \sum_{X \subseteq T_j} supp(T_j) = \sum_{X \subseteq T_j} \bigcap_{\tau_a \in T_j} \mu_a(T_j).$$

Взвешенную поддержку набора X, учитывающую оценки индивидуальной информативности признаков, входящих в данных набор, определим следующим образом:

$$wsupp(X) = supp(X) \sum_{\tau_a \in X} w_a,$$

где величина $\sum_{\tau_a \in X} w_a$ определяет оценку информативности набора признаков X.

Взвешенная поддержка ассоциативного правила $X \to Y$ может быть определена по формуле:

$$wsupp(X \to Y) = supp(X \cup Y) \sum_{\tau_a \in X \cup Y} w_a.$$

Будем считать набор X часто встречающимся взвешенным набором, если будет выполняться условие:

$$\mathrm{wsupp}(X) \geq \mathrm{wminsupport},$$

где wminsupport – пороговое (минимально допустимое) значение взвешенной поддержки.

Важно отметить, что в некоторых случаях кроме часто встречающихся наборов X важными для извлечения новых знаний об исследуемых объектах или процессах являются нечастые наборы элементов, позволяющие выявлять косвенные (непрямые) ассоциации.

Если два набора элементов X и Y существенно зависят от наличия третьего набора Z, тогда будем считать, что пара X и Y косвенно связана по набору Z: $X \xrightarrow{Z} Y$. Наличие такой связи будем определять, исходя из истинности таких условий:

1) значение взвешенной поддержки набора $X \cup Y$ меньше минимально допустимой:

$$\mathrm{wsupp}(X \cup Y) < \beta_{\mathrm{wsupp}(X \cup Y)},$$

где $\beta_{\mathrm{wsupp}(X \cup Y)}$ – пороговое значение взвешенной нечеткой поддержки между наборами X и Y – величина, указывающая на то, что наборы X и Y встречаются не часто. Величину $\beta_{\mathrm{wsupp}(X \cup Y)}$ можно установить следующим образом: $\beta_{\mathrm{wsupp}(X \cup Y)} = \mathrm{wminsupport}$;

2) существует непустой набор Z ($\exists Z \neq \varnothing$), для которого выполняются условия:

$$\begin{cases} \mathrm{wsupp}(X \cup Z) \geq \beta_{\mathrm{wsupp}(Z)}; \\ \mathrm{wsupp}(Y \cup Z) \geq \beta_{\mathrm{wsupp}(Z)}; \end{cases} \text{ и } \begin{cases} w(X,Z) \geq w_{\min}; \\ w(Y,Z) \geq w_{\min}, \end{cases}$$

где $\beta_{\mathrm{wsupp}(Z)}$ – пороговое значение взвешенной нечеткой поддержки между некоторым набором и набором Z, являющимся ключевым для появления пары наборов X и Y, – величина, указывающая на то, что наборы X и Y встречаются часто при наличии множества Z. Величину $\beta_{\mathrm{wsupp}(Z)}$ целесообразно установить следующим образом: $\beta_{\mathrm{wsupp}(Z)} \geq \beta_{\mathrm{wsupp}(X \cup Y)}$; $w(X,Z)$ и $w(Y,Z)$ – значения критерия оценивания взаимосвязи между множествами X и Z, а также Y и Z, соответственно; $w_{\min}$ – минимально допустимое значение критерия оценивания взаимосвязи между множествами элементов базы транзакций.

В качестве меры $w(X,Z)$ целесообразно использовать следующую:

$$w(X,Z) = \frac{p(X \cap Z)}{\sqrt{p(X)p(Z)}},$$

где $p(X)$, $p(Z)$, $p(X \cap Z)$ – вероятность появления наборов X, Z и $X \cap Z$ в базе данных D.

Таким образом, использование предложенных выше критериев и их пороговых значений позволит извлекать не только часто встречающиеся наборы, но и наборы, редко возникающие в исходной базе данных, однако являющиеся интересными и позволяющие выявлять новые знания об исследуемых объектах или процессах.

При построении базы ассоциативных правил в процессе их извлечения используются значения индивидуальной информативности признаков, рассчитанные ранее, что позволяет учитывать значимость каждого атрибута при поиске правил. При генерации новых наборов-кандидатов в процессе синтеза ассоциативных правил учитывается свойство антимонотонности поддержки [1, 3], применение которого позволяет существенно сократить пространство поиска. Для извлечения ассоциативных правил каждое j-е численное значение τ_{aj} a-го признака τ_a в транзакции T_j преобразовывается к нечеткому значению $f\tau_{aj}$:

$$f\tau_{aj} = \sum_{k=1}^{N_{S=Ba}} \frac{\mu_{ak}(\tau_a \in T_j)}{|\Delta_{ak}|},$$

где $\mu_{ak}(\tau_a \in T_j)$ – функция принадлежности a-го признака k-му терму, вычисленная для значения признака τ_a в транзакции T_j; $|\Delta_{ak}|$ – ширина k-го диапазона разбиения a-го признака.

После этого вычисляется мощность каждого k-го диапазона разбиения a-го признака:

$$C\Delta_{ak} = \sum_{j=1}^{N_D} \mu_{ak}(\tau_a \in T_j),$$

и находится максимальное значение такой величины для каждого a-го признака:

$$\max C\Delta_a = \max_{k=1,2,...,N_{S=Ba}} C\Delta_{ak},\ a = 1,2,...,|I|,$$

а также соответствующий величине $\max C\Delta_a$ интервал разбиения $\max \Delta_a$, который в дальнейшем процессе извлечения ассоциативных правил будет использоваться для представления нечетких характеристик элемента τ_a.

Для каждого интервала $\max \Delta_a$, $a = 1,2,...,|I|$ вычисляется взвешенная поддержка $\mathrm{wsupp}(\max \Delta_a)$ по формулам, приведенным выше (до этого определяется значимость w_a каждого из признаков τ_a. Все интервалы $\max \Delta_a$, значения взвешенной поддержки которых не менее минимально допустимого порогового значения wminsupport, заносятся в массив FI_1, содержащий одноэлементные часто встречающиеся наборы:

$$FI_1 = \{\max \Delta_a \mid \mathrm{wsupp}(\max \Delta_a) \geq \mathrm{wminsupport}\}.$$

Интервалы с малыми значениями взвешенных поддержек $\mathrm{wsupp}(\max \Delta_a)$ заносятся в массив RI_1 редко встречающихся одноэлементных наборов:

$$RI_1 = \{\max \Delta_a \mid \mathrm{wsupp}(\max \Delta_a) < \mathrm{wminsupport}\}.$$

В случае, если множество FI_1 является пустым, метод прекращает свою работу, поскольку сгенерировать часто встречающиеся и достоверные ассоциативные правила не представляется возможным.

Затем на основе текущего множества FI_d d-элементных наборов генерируется множество C_{d+1} (d + 1)-элементных кандидатов в часто встречающиеся наборы. При этом аналогично методу Apriori [1 – 3] для уменьшения количества кандидатов на (d + 1)-й итерации используется свойство антимонотонности поддержки, заключающееся в том, что поддержка любого множества элементов X не превышает значения минимальной поддержки любого его подмножества $Y \subset X$ [1, 3]. Поэтому

на этапе генерации множества кандидатов C_{d+1} отсекаются (не создаются и не заносятся в C_{d+1}) те наборы, которые не могут стать часто встречающимися, что определяется на основе информации о наборах с низкими значениями поддержки wsupp, рассчитанными на предыдущих этапах и хранящимися во множестве RI. Таким образом, при создании нового множества C_{d+1} кандидатов используется идея о том, что у набора, который потенциально является часто встречающимся, все подмножества также должны быть часто встречающимися (значения всех поддержек подмножеств должно быть не ниже порогового значения).

Следовательно, кандидат X, содержащий подмножество $Y \subset X$, отброшенное на предыдущих этапах как нечасто встречающееся ($Y \in RI$), не включается в следующее множество C_{d+1} кандидатов в часто встречающиеся наборы.

После формирования множества C_{d+1} для каждого набора $X = \{\tau_1, \tau_2, ..., \tau_{d+1}\} \in C_{d+1}$ ($|X| = d+1$) вычисляется его нечеткая характеристика для j-й транзакции T_j:

$$\mu_X(T_j) = \bigcap_{a:\tau_a \in X} \mu_a(\tau_a \in T_j),$$

далее определяется взвешенная поддержка набора X:

$$wsupp(X) = \sum_{X \in T_j, T_j \in D} \mu_X(T_j) \sum_{\tau_a \in X} w_a .$$

Если значение $wsupp(X)$ не менее минимально допустимого порога wminsupport, множество X заносится в массив FI_{d+1} часто встречающихся наборов элементов, в противном случае – в массив редко встречающихся наборов RI_{d+1}.

В случае, если $FI_{d+1} \neq \varnothing$, выполняются действия, аналогичные описанным выше.

В противном случае считается, что дальнейшее генерирование часто встречающихся наборов является невозможным. Поэтому далее выполняется извлечение ассоциативных правил с приемлемым уровнем достоверности.

Ассоциативные правила будем генерировать исходя из того, что:

$$wconf(X \to Y) = \frac{wsupp(X \to Y)}{wsupp(X)} \geq wminconfidence,$$

$$X \cap Y = \varnothing.$$

Массив всех часто встречающихся наборов, найденных ранее, может быть сформирован как совокупность массивов FI_C:

$$FI = \bigcup_{C=1}^{d} FI_C .$$

Для каждого набора $A \in FI$ и каждого его подмножества $X \in A$ выполняются проверки:

$$\frac{wsupp(A)}{wsupp(X)} \geq wminconfidence$$

и $$\frac{wsupp(A)}{wsupp(A \setminus X)} \geq wminconfidence .$$

Пусть $Y = A \setminus X$. Тогда, если выполняется первое условие, то генерируется ассоциативное правило $X \to Y$. Если выполняется второе условие, то генерируется правило $Y \to X$. При невыполнении обоих условий, генерации правила для $A \in FI$ и $X \in A$ не происходит.

После этого выполняется поиск интересных, но редко встречающихся правил вида $X \xrightarrow{Z} Y$. Для этого формируется множество RI:

$$RI = \bigcup_{C=1}^{d} RI_C ,$$

и для каждого его элемента $A \in RI$ выполняются следующие действия: $X = A_{|A|}$ – последний элемент множества A; $Y = A_{|A|-1}$ – предпоследний элемент множества A; $Z = A \setminus (X \cup Y)$. Тогда будем извлекать ассоциативные правила вида $X \xrightarrow{Z} Y$ при выполнении следующих условий:

$$\begin{cases} wsupp(X \cup Y) < \beta_{wsupp(X \cup Y)}; \\ (wsupp(X \cup Z)) \cap (wsupp(Y \cup Z)) \geq \beta_{wsupp(Z)}; \\ w(X,Z) \cap w(Y,Z) \geq w_{min}. \end{cases}$$

После извлечения импликаций вида $X \to Y$ и $X \xrightarrow{Z} Y$ на их основе синтезируется база ассоциативных правил, описывающая исследуемые объекты и процессы.

Разработанный метод обеспечивает интеграцию описанных выше принципов, позволяет по заданным транзакционным базам данных строить наборы численных ассоциативных правил и на их основе извлекать новые знания об исследуемых объектах или процессах.

Предложенный метод предполагает фаззификацию заданной базы транзакций и автоматическое разбиение диапазонов значений признаков на интервалы, учитывает индивидуальную значимость признаков, использует критерии для оценивания косвенных ассоциаций, что понижает степень участия пользователя в процессе поиска ассоциативных правил, уменьшает вероятность извлечения правил, некорректно описывающих исследуемые объекты и процессы, а также позволяет извлекать не только часто встречающиеся наборы, но редко возникающие интересные ассоциативные правила.

4. Эксперименты и результаты

С целью проведения экспериментов по исследованию свойств и характеристик предложенного метода извлечения численных ассоциативных правил он был программно реализован на языке программирования C#.

Экспериментальное исследование разработанного метода выполнялось на основе данных, представленной в виде транзакционной базы данных, содержащей информацию о состоянии здоровья детей, рожденных от родителей, пострадавших от аварии на Чернобыльской АЭС [17]. В результате обследования пациентов получен набор данных, содержащих диагностические критерии формирования различных заболеваний, а также установленные диагнозы. С целью выявления взаимосвязи между заболеваниями, а также влияния значений различных показателей на тот или иной диагноз выполнялось извлечение ассоциативных данных.

Поскольку большинство результатов лабораторных исследований носят численный характер, целесообразным является извлечение численных ассоциативных правил.

База данных содержала $N_D = |D| = 344$ записей (транзакций), каждая из которых представляла информацию о конкретном пациенте и могла характеризоваться несколькими из $N_I = |I| = 69$ признаков. Каждая запись содержала в среднем 14 признаков.

Результаты проведения экспериментов позволили выявить взаимосвязи различных заболеваний вида «Если установлен набор диагнозов $D_1, D_2, ..., D_{kD}$ и значения численных показателей находятся в определенных пределах $x_i \in A_{iyj}$, то у пациента с вероятностью P_j может быть установлен диагноз y_j». Это позволит выполнять диагностирование некоторых болезней на ранних стадиях, а также предоставлять своевременные рекомендации для проведения комплекса профилактических мероприятий по недопущению возникновения болезней, с большой степенью вероятности сопровождающихся или возникающих вследствие заболеваний, диагноз по которым уже установлен. Кроме того, выявлены факторы, являющиеся пусковым механизмом для перехода от латентной формы заболевания к открытой. В частности, выявлено, что наиболее информативными факторами, позволяющими диагностировать нейро-артритические аномалии на ранних стадиях, являются: уменьшение концентрации 4-пиридоксиновой кислоты, эмоциональная лабильность, диспептический синдром, ацетонемическая рвота, уратурия в период новорожденности.

Выявленные факторы и зависимости позволят своевременно предпринимать необходимые действия для предотвращения нежелательных переходов от латентной формы к открытой форме заболевания. Таким образом, результаты экспериментов показали, что разработанный метод позволяет извлекать из баз транзакций численные ассоциативные правила, используя при этом априорную информацию о значимости признаков, что сокращает пространство поиска и время извлечения правил, уменьшает количество извлеченных правил, и, соответственно, повышает уровни обобщения и интерпретабельности синтезированной базы ассоциативных правил.

5. Выводы

В работе решена актуальная задача автоматизации извлечения численных ассоциативных правил.

Научная новизна работы заключается в том, что предложен метод извлечения численных ассоциативных правил, основными этапами которого являются: фаззификация транзакционной базы данных, определение индивидуальной значимости признаков, вычисление пороговых значений поддержки и построение базы численных ассоциативных правил. Предложенный метод предполагает фаззификацию заданной базы транзакций и автоматическое разбиение диапазонов значений признаков на интервалы, учитывает индивидуальную значимость признаков, использует критерии для оценивания косвенных ассоциаций, что понижает степень участия пользователя в процессе поиска ассоциативных правил, уменьшает вероятность извлечения правил, некорректно описывающих исследуемые объекты и процессы, а также позволяет извлекать не только часто встречающиеся наборы, но и редко возникающие интересные ассоциативные правила.

Использование априорной информации о значимости признаков в разработанном методе позволяет сократить пространство поиска и время извлечения правил, уменьшить количество извлеченных правил, и, соответственно, повысить уровни обобщения и интерпретабельности синтезированной базы ассоциативных правил.

Практическая ценность полученных результатов заключается в том, что на основе предложенного метода разработано программное обеспечение, позволяющее выполнять извлечение численных ассоциативных правил, а также решена практическая задача медицинского диагностирования.

Работа выполнена в рамках госбюджетной научно-исследовательской темы Запорожского национального технического университета «Интеллектуальные информационные технологии автоматизации проектирования, моделирования, управления и диагностирования производственных процессов и систем» (номер государственной регистрации 0112U005350).

Литература

1. Zhang, C. Association rule mining: models and algorithms [Text] / C. Zhang, S. Zhang. – Berlin : Springer-Verlag. – 2002. – 238 p.
2. Gkoulalas-Divanis, A. Association Rule Hiding for Data Mining [Text] / A. Gkoulalas-Divanis,V. S. Verykios. – New York : Springer-Verlag. – 2010. – 150 p.
3. Zhao, Y. Post-mining of association rules: techniques for effective knowledge extraction [Text] / Y. Zhao, C. Zhang, L. Cao. – New York : Information Science Reference. – 2009. – 372 p.
4. Dubois, D. A Systematic Approach to the Assessment of Fuzzy Association Rules [Text] / D. Dubois, E. Hullermeier, H. Prade // Data Mining and Knowledge Discovery. – 2006. – Vol. 13. – P. 167-192.
5. Khan, M. S. Weighted Association Rule Mining from Binary and Fuzzy Data [Text] / M. S. Khan, M. Muyeba, F. Coenen // Lecture Notes in Computer Science. – 2008. – Vol. 5077. – P. 200-212.
6. Lian, W. An efficient algorithm for finding dense regions for mining quantitative association rules [Text] / W. Lian, D. W. Cheung, S. M. Yiu // Computers & Mathematics With Applications. – 2005. – Vol. 50, № 3. – P. 471-490.
7. Sohn, S. Y. Searching customer patterns of mobile service using clustering and quantitative association rule [Text] / S. Y. Sohn, Y. Kim // Expert Systems With Applications. – 2008. – Vol. 34, № 2. – P. 1070-1077.
8. Adamo, J.-M. Data mining for association rules and sequential patterns: sequential and parallel algorithms [Text] / J.-M. Adamo. – New York : Springer-Verlag. – 2001. – 259 p.

9. Koh, Y. S. Rare Association Rule Mining and Knowledge Discovery [Text] / Y. S. Koh, N. Rountree. – New York : Information Science Reference. – 2009. – 320 p.
10. Zadeh, L. Fuzzy sets [Text] / L. Zadeh // Information and Control. – 1965. – № 8. – P. 338–353.
11. Субботін, С. О. Неітеративні, еволюційні та мультиагентні методи синтезу нечіткологічних і нейромережних моделей: монографія [Текст] / С. О. Субботін, А. О. Олійник, О. О. Олійник ; під заг. ред. С. О. Субботіна. – Запоріжжя : ЗНТУ, 2009. – 375 с.
12. Encyclopedia of artificial intelligence [Text] / Eds.: J. R. Dopico, J. D. de la Calle, A. P. Sierra. – New York : Information Science Reference, 2009. – Vol. 1–3. – 1677 p.
13. Интеллектуальные информационные технологии проектирования автоматизированных систем диагностирования и распознавания образов : монография [Текст] / [С. А. Субботин, Ан. А. Олейник, Е. А. Гофман, С. А. Зайцев, Ал. А. Олейник под ред. С. А. Субботина]. – Харьков : ООО "Компания Смит", 2012. – 317 с.
14. Прогрессивные технологии моделирования, оптимизации и интеллектуальной автоматизации этапов жизненного цикла авиадвигателей : монография [Текст] / [А. В. Богуслаев, Ал. А. Олейник, Ан. А. Олейник, Д. В. Павленко, С. А. Субботин под ред. Д. В. Павленко, С. А. Субботина]. – Запорожье : ОАО «Мотор Сич», 2009. – 468 с.
15. Гибридные нейро-фаззи модели и мультиагентные технологии в сложных системах : монография [Текст] / [В. А. Филатов, Е. В. Бодянский, В. Е. Кучеренко и др. под общ. ред. Е. В. Бодянского]. – Дніпропетровськ : Системні технології, 2008. – 403 с.
16. Айвазян, С. А. Прикладная статистика: Исследование зависимостей [Текст] / С. А. Айвазян, И. С. Енюков, Л. Д. Мешалкин. – М.: Финансы и статистика, 1985. – 487 с.
17. Диагностирование нейро-артритических аномалий на основе ассоциативных правил [Текст] / Т. А. Зайко, А. А. Олейник, Н. В. Жихарева, С. А. Субботин // Бионика интеллекта. – 2012. – № 2 (79). – С. 53–57.

В роботі досліджуються питання визначення показників якості для атомарних сервісів в сервіс-орієнтованих системах. Визначено кількісні оцінки показників якості для атомарних сервісів. Запропоновано методи моніторингу та управління сервісами на підставі статистичних даних показників якості сервісів. У статті описані методи вибору екземплярів атомарних сервісів з однаковим інтерфейсом із пулу сервісів

Ключові слова: web-сервіс, SOA, якість, час відгуку, доступність, надійність, якість послуг

В работе исследуются вопросы показателей качества для атомарных сервисов в сервис-ориентированных системах. Определены количественные оценки показателей качества для атомарных сервисов. Предложены методы мониторинга и управления сервисами на основании статистических данных по показателям качества сервисов. В статье описаны методы выбора экземпляров атомарных сервисов, предоставляющих с одинаковым интерфейсом из пула сервисов

Ключевые слова: web-сервис, SOA, время отклика, доступность, надежность, качество услуг

УДК 004.052

МЕТОД ОЦЕНИВАНИЯ ПОКАЗАТЕЛЕЙ КАЧЕСТВА WEB-СЕРВИСОВ

О. В. Рогов*
E-mail: olehrgf@gmail.com
Т. В. Дуравкина
Кандидат технических наук, старший преподаватель*
E-mail: stv_@list.ru
А. Г. Морозова
Кандидат технических наук, старший преподаватель*
E-mail: a.morozova@karazin.ua
*Кафедра теоретической и прикладной информатики
Харьковский национальный университет им. В. Н. Каразина
пл. Свободы, 4, г. Харьков, Украина, 61022

1. Введение

В настоящее время успех бизнеса сильно зависит от того, насколько он автоматизирован и как быстро компания может предложить новую услугу или продукт на рынок.

Практически перед любым IT подразделением компании всегда стоит задача бесперебойного предоставления IT сервисов бизнесу. Реализация традиционных решений для интеграции прикладных программ - непростая задача, требующая существенных капиталовложений. Кроме того, часто при внедрении