**УДК 004.93**

***Т.А. ЗАЙКО***, аспирантка, ЗНТУ, Запорожье,
***А.А. ОЛЕЙНИК,*** канд. техн. наук, доц., ЗНТУ, Запорожье,
***С.А. СУББОТИН***, канд. техн. наук, проф., ЗНТУ, Запорожье

# АССОЦИАТИВНЫЕ ПРАВИЛА В ИНТЕЛЛЕКТУАЛЬНОМ АНАЛИЗЕ ДАННЫХ

Рассмотрена задача построения моделей на основе ассоциативных правил. Проанализирован процесс поиска ассоциативных правил. Исследованы различные виды ассоциативных правил (негативные, численные, обобщенные, временные и нечеткие ассоциативные правила) при использовании их для решения задач интеллектуального анализа данных. Библиогр.: 15 назв.

**Ключевые слова**: ассоциативное правило, различные виды ассоциативных правил, интеллектуальный анализ данных, нечеткие ассоциативные правила.

**Постановка проблемы и анализ литературы.** В настоящее время в связи со снижением удельной стоимости хранения данных возрастает объем хранимой информации на предприятиях [1, 2], в результате чего возникают задачи, связанные с необходимостью обработки больших массивов данных с целью поиска новых закономерностей, установления и выявления новых знаний.

Задачи прикладного характера, связанные с необходимостью обработки больших массивов данных, возникают на промышленных предприятиях, а также в организациях, занимающихся розничной торговлей, финансовым анализом, логистикой и коммуникациями [2 – 4].

Для анализа данных в настоящее время широко применяют методы и средства искусственного интеллекта [1, 4, 5], в частности нейронные сети, нечеткие модели, деревья решений, байесовские сети, методы регрессионного анализа и др. [1, 4 – 7].

Однако такие методы, как правило, используются для обработки структурированных данных, представленных в виде массивов, содержащих значения признаков и выходных параметров экземпляров выборки [1, 4 – 6].

В настоящее время наблюдается переизбыток так называемых неструктурированных данных [1, 2], в которых каждая единица хранения не может быть представлена конечным числом признаков (атрибутов). Такие данные могут содержать, например, информацию о товарах, купленных одним покупателем у предприятия розничной торговли; результаты ответов респондента при проведении анкетирования; набор

установленных диагнозов и результатов лабораторных исследований у пациентов лечебных учреждений; набор различного рода данных о клиентах предприятий и др.

Такие данные представляются, как правило, в виде последовательностей связанных событий [1 – 4]. При этом нет чёткого понимания, что является входными данными, а что выходными. Кроме того, размер каждой транзакции (множества событий, произошедших одновременно) не является фиксированным.

В связи с этим возникают задачи:

– сокращения объемов неструктурированных данных путем удаления избыточных транзакций, исключение которых из дальнейшего рассмотрения не повлияет на качество синтезируемых правил и моделей;

– выявления интересных правил, позволяющих извлекать новые знания на основе имеющихся неструктурированных данных;

– построения моделей на основе больших массивов неструктурированных данных для решения практических задач прогнозирования, классификации и кластеризации данных.

Для обработки больших массивов неструктурированных данных и решения указанных задач целесообразно использовать методы поиска ассоциативных правил [2, 4, 5, 8 – 10], позволяющие выявлять новые закономерности вида "если условие, то действие" в имеющихся данных и синтезировать на их основе интерпретабельные базы правил, понятные экспертам в прикладных областях.

В настоящее время предложено достаточно большое количество видов ассоциативных правил, каждый из которых целесообразно применять для решения определенного класса задач. Поэтому актуальным является обзор и классификация ассоциативных правил для дальнейшего их применения с целью решения практических задач интеллектуального анализа данных.

**Цель статьи** – анализ ассоциативных правил и методов их построения для решения задач интеллектуального анализа данных.

**Синтез ассоциативных правил.** Пусть задан набор данных *D*:

$$D = \{T_1, T_2, \ldots, T_{N_D}\}, \quad (1)$$

представляющий собой транзакционную базу данных [2], в которой каждый элемент $T_j$, $j = 1, 2, \ldots, N_D$ содержит информацию о некоторых взаимосвязанных событиях, где $N_D = |D|$ – количество элементов (транзакций) в наборе данных *D*.

Элементы $T_j$ могут представляться в виде (2):

$$T_j = \left(tid_j,\ item_j\right), \tag{2}$$

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

Таким образом, каждая транзакция $T_j$ набора данных $D$ представляет собой список элементов $item_j$, являющийся подмножеством множества $I$.

Ассоциативным правилом (АП) называется импликация $X \to Y$, в которой наборы $X$ и $Y$ не пересекаются (3) [2, 8 – 10]:

$$X \to Y:\ X \subset I,\ Y \subset I,\ X \bigcap Y = \varnothing. \tag{3}$$

Т.е. ассоциативное правило описывает закономерности вида: "Из события $X$ следует событие $Y$" или "если условие, то действие" [2, 9].

Задача поиска ассоциативных правил $AR$ заключается в том, чтобы на основе имеющегося набор данных $D$ (транзакционной базы данных) найти закономерности между событиями $\tau_a \in I$, $a = 1, 2, ..., N_I$.

Задача построения АП связана с необходимостью вычисления поддержки достоверности правил $a$. Набор $X \subset I$ из базы $D$ имеет поддержку $\mathrm{supp}(X)$, определяемую как отношение количества транзакций $T$ в наборе данных $D$, содержащих множество элементов $X$, к общему количеству транзакций в базе данных $D$.

Поддержкой $\mathrm{supp}(X \to Y)$ правила $X \to Y$ является поддержка множества $X \bigcup Y$: $\mathrm{supp}(X \to Y) = \mathrm{supp}(X \bigcup Y)$.

Достоверностью $\mathrm{conf}(X \to Y)$ правила $X \to Y$ называют отношение его поддержки $\mathrm{supp}(X \to Y)$ к поддержке $\mathrm{supp}(X)$ множества $X$.

Процесс синтеза ассоциативных правил может быть разбит на два этапа [2, 9]:

– генерирование всех наборов $X$ с уровнем поддержки, не ниже заданного экспертом порогового значения minsupport($X$), в результате чего формируются часто встречаемые наборы $X \subset I$;

– генерирование всех правил $X \to Y$ с уровнем достоверности, не ниже заданного экспертом порогового значения minconfidence( $X \to Y$ ).

**Анализ видов ассоциативных правил**. При поиске взаимосвязей между различными элементами в транзакционных базах данных $D = \{T_1, T_2, ..., T_{N_T}\}$ часто необходимо выявлять не только, так называемые, позитивные ассоциативные правила (positive association rules) $X \to Y$, но и другие виды правил. К таким правилам относятся [8 – 15]: негативные, численные, обобщенные, временные и нечеткие АП.

Негативные АП (negative association rules) характеризуют отрицательную взаимосвязь между различными событиями типа: "Если произошло событие $X$, то событие $Y$ не наступит" ( $X \to \overline{Y}$ ) или "Если не произошло событие $X$, то наступит событие $Y$" ( $\overline{X} \to Y$ ) [11].

Необходимость извлечения негативных ассоциативных правил $X \to \overline{Y}$ или $\overline{X} \to Y$ наряду с позитивными правилами $X \to Y$ обуславливается следующей причиной. Построение полного набора ассоциативных правил ( $X \to Y$, $X \to \overline{Y}$, $\overline{X} \to Y$ ) между различными объектами $\tau_a \in I$, $a = 1, 2, ..., N_I$ базы данных $D = \{T_1, T_2, ..., T_{N_T}\}$ позволит более детально описать исследуемые зависимости, что в свою очередь приведет к более точным результатам прогнозирования по синтезированной базе правил [8, 11].

Для поиска интересных негативных правил (таких, которые представляют интерес в конкретной прикладной области в соответствии с заданным набором данных $D$) необходимо учитывать уровень их интереса, определяемый в соответствии с критерием Пятецкого-Шапиро [2, 9] следующим образом

$$\text{supp}(X \to \overline{Y}) - \text{supp}(X)\text{supp}(\overline{Y}) \geq \varepsilon_I . \qquad (4)$$

При выполнении неравенства (4) правила $X \to \overline{Y}$ считаются интересными. Аналогичным образом можно определить и неравенства для поиска интересных негативных правил типа $\overline{X} \to Y$ [11].

Таким образом, при извлечении негативных правил $X \to \overline{Y}$ из набора данных $D$ поиск происходит таких транзакций $T_j$, в результате чего извлекается набор негативных правил $X \to \overline{Y}$, удовлетворяющих списку условий (5):

$$\begin{cases} \left(\text{supp}(X) \ge \text{minsupport}\right) \cap \left(\text{supp}(Y) \ge \text{minsupport}\right) \cap \\ \cap \left(\text{supp}(X \to \overline{Y}) \ge \text{minsupport}\right); \\ \text{conf}(X \to Y) \ge \min \text{confidence}\ ; \\ \left|\text{supp}(X \to \overline{Y}) - \text{supp}(X)\text{supp}(\overline{Y})\right| \ge \varepsilon_I . \end{cases} \quad (5)$$

Приведенные условия позволяют выявлять достоверные негативные правила $X \to \overline{Y}$ с приемлемым уровнем поддержки и являющиеся интересными в исследуемой предметной области.

Важно отметить, что при идентификации негативных ассоциативных правил необходимо обрабатывать нечастые последовательности в заданной базе данных *D*. Однако большинство методов синтеза АП основаны на извлечении и анализе часто встречаемых наборов, что затрудняет их применение на практике для поиска негативных АП [2, 8–11], и обуславливает необходимость разработки новых методов синтеза АП, позволяющих извлекать как позитивные, так и негативные АП.

Часто признаки $\tau_a \in I$, $a = 1, 2, \ldots, N_I$ могут принимать не только бинарные, но и численные значения из некоторого диапазона значений $\tau_a \in [\tau_{a\min}; \tau_{a\max}]$ или множества значений $\text{T}(\tau_a) = \{\tau_{a1}, \tau_{a2}, \ldots, \tau_{aN_{\tau a}}\}$. Поэтому актуальной является задача выделения правил вида Если $X \in [X_{\min}; X_{\max}]$, то $Y \in [Y_{\min}; Y_{\max}]$. Численным ассоциативным правилом (quantitative association rule) называется импликация вида (6) [2, 12]:

$$\langle X, v(X) \rangle \to \langle Y, v(Y) \rangle , \quad (6)$$

где $v(X) \in \text{T}(X)$ и $v(Y) \in \text{T}(Y)$ – значения переменных $X$ и $Y$, соответственно, принадлежащие множествам возможных значений $\text{T}(X)$ и $\text{T}(Y)$.

Поддержка $\text{supp}(X \to Y)$ численного АП $X \to Y$ вида (6) определяется по формуле (7) [12]:

$$\text{supp}(X \to Y) = \frac{1}{|D|} \sum_{j=1}^{|D|} \prod_{i=1}^{N_X + N_Y} v_j(\tau_a), \ \tau_a \in (X \cup Y), \quad (7)$$

где $N_X$ и $N_Y$ – количество элементов $\tau_a \in I$ в множествах $X = \{\tau_1, \tau_2, \ldots, \tau_{N_X}\}$ и $Y = \{\tau_{N_X+1}, \tau_{N_X+2}, \ldots, \tau_{N_X+N_Y}\}$, соответственно; $v_j(\tau_a)$ – значение *a*-го признака $\tau_a$ в *j*-й транзакции $T_j$ базы данных $D$.

Достоверность $\mathrm{conf}(X \to Y)$ численного АП $X \to Y$ вида (6) определяется аналогично позитивным АП. При этом поддержка $\mathrm{supp}(X)$ множества $X$ вычисляется в соответствии с формулой (8) [12]:

$$\mathrm{supp}(X) = \frac{1}{|D|} \sum_{j=1}^{|D|} \prod_{i=1}^{N_X} v_j(\tau_a), \ \tau_a \in X \ . \qquad (8)$$

Процесс поиска численных АП вида (6) по заданным наборам данных $D$ связан с необходимостью разбиения на интервалы (дискретизации) диапазонов возможных значений элементов $\tau_a \in I$, входящих в транзакции $T_j$, $j = 1, 2, ..., N_T$. В результате такого разбиения каждая $j$-я транзакция $T_j = (tid_j, \ item_j)$ представляется списком элементов $item_j = \{t_{1j}, \ t_{2j}, \ ..., \ t_{N_{item_j} j}\} \subseteq I$, в котором каждый $i$-й элемент $t_{ij}$ представляется в виде (9):

$$t_{ij} = (\text{элемент } \tau_a \in I \text{ ; диапазон значений элемента } \tau_a), \qquad (9)$$

при этом множество $I = \{\tau_1, \tau_2, ..., \tau_{N_I}\}$ возможных значений, которые могут входить в список элементов $item_j$ каждой транзакции $T_j$, содержит элементы $\tau_a$ (10):

$$\tau_a \in [\tau_{a\min}; \ \tau_{a\max}] = \bigcup_{c=1}^{N_{\text{разб}\tau_a}} [\tau_{a\min c}; \ \tau_{a\max c}], \ a = 1, 2, ..., N_I, \qquad (10)$$

где $\tau_{a\min}$ и $\tau_{a\max}$ – минимальное и максимальное значение, которые может принимать $a$-й элемент $\tau_a$ множества $I$; $\tau_{a\min c}$ и $\tau_{a\max c}$ – минимальное и максимальное значение $c$-го интервала разбиения значений $a$-го элемента $\tau_a$ множества $I$; $N_{\text{разб}\tau_a}$ – количество интервалов разбиения $a$-го элемента $\tau_a$.

После дискретизации значений численных переменных выполняется поиск ассоциативных правил $X \to Y$. При этом используются методы извлечения ассоциативных правил, удовлетворяющих приведенным выше условиям к позитивным правилам вида $X \to Y$, но при таком поиске каждый диапазон дискретизации каждой переменной считается отдельным элементом, который может быть использован при построении ассоциативного правила [2].

Однако необходимость дискретизации диапазонов значений переменных для извлечения численных АП существенно увеличивает пространство поиска и требования к вычислительным ресурсам ЭВМ. Кроме того, в некоторых случаях дискретизация приводит к неудачным разбиениям диапазона значений переменных, в результате чего не обеспечивается приемлемая точность прогнозирования или классификации по синтезированной базе АП. Поэтому актуальной задачей является разработка новых методов поиска численных АП, свободных от указанных недостатков.

Иногда элементы $\tau_a \in I$, образующие транзакции $T_j$ базы данных $D$, могут быть объединены в группы, а группы элементов могут формировать группы более высокого уровня и т.д., образуя, таким образом, некоторую иерархическую древоподобную структуру. В таких случая целесообразно извлекать правила, связывающие не только наборы конкретных элементов $\tau_a \in I$ из базы данных $D$, но и элементы с группами, а также группы с группами. Это особенно полезно, например, в задачах анализа взаимосвязей потребительского спроса на различные группы товаров, взаимосвязей различных групп болезней в медицинском диагностировании и т.п. [8 – 10, 13].

Пусть задана транзакционная база данных $D$ (1), содержащая элементы $\tau_a \in I$, которые могут быть отнесены к определенным группам $Gr = \{gr_1, gr_2, ..., gr_{N_{Gr}}\}$. То есть может быть построено дерево, описывающее связи типа "часть – целое" исследуемой предметной области. Для узла-потомка $\tau_{child}$ будем считать узел $\tau_{parent}$, который располагается на более высоком уровне дерева, и от которого имеется путь к узлу $\tau_{child}$.

Обобщенным АП (generalized association rule) называется импликация $X \to Y$, в которой ни один из элементов множества $Y = \{\tau_{Y1}, \tau_{Y2}, ..., \tau_{N_Y}\}$ не является предком какого-либо элемента множества $X = \{\tau_{X1}, \tau_{X2}, ..., \tau_{N_X}\}$ [2, 8 – 10, 13]. При этом остальные условия для АП сохраняются.

Задача поиска обобщенных АП заключается в том, чтобы на основе заданной базы данных $D$ и построенной иерархии элементов идентифицировать все правила с уровнями поддержки, достоверности и интереса, не ниже заданных пороговых значений minsupport, minconfidence и $\varepsilon_I$. При этом поддержка $\mathrm{supp}(X \to Y)$ и достоверность

$\mathrm{conf}(X \to Y)$ обобщенных АП $X \to Y$ вычисляются аналогично позитивным АП.

Использование дополнительной информации об иерархических связях и о возможности группировки элементов $\tau_a \in I$, а также введение дополнительного множества $Gr = \{gr_1, gr_2, ..., gr_{N_{Gr}}\}$ при поиске АП позволяет извлекать правила между различными иерархическими уровнями, что, в свою очередь, обеспечивает возможность выявления скрытых связей между различными наборами элементов из $D$ (1). Это достигается в том числе за счет того, что поддержка группы $\mathrm{supp}(gr\}$ может быть больше поддержки элементов $\tau_a \in I$, ее образующих, и, соответственно, быть более минимальной поддержки $\mathrm{minsupport}(X)$, что позволит синтезировать правила типа $gr \to Y$, $X \to gr$, $gr_1 \to gr_2$, в то время как некоторые правила типа $X \to Y$ не будут извлечены в силу невыполнения условий соответствия, условий удовлетворения пороговым значениям поддержки и достоверности.

С целью извлечения обобщенных АП применяют методы поиска позитивных АП. При этом каждая транзакция $T_j$ базы данных $D$ (1) расширяется путем дополнения ее всеми предками каждого из элементов, в нее входящих. Однако применение такого подхода связано с такими проблемами [8, 13]:

– элементы *gr*, расположенные на верхних иерархических уровнях дерева, характеризуются существенно более высокими значениями поддержки, что, как правило, приводит к их появлению в большинстве синтезированных правил и, следовательно, к построению избыточных правил и к усложнению построенной базы правил;

– существенное увеличение пространства поиска, связанное с добавлением групповых элементов *gr* в транзакции $T_j$, что усложняет задачу извлечения АП и увеличивает количество извлеченных правил.

Необходимость устранения приведенных недостатков обуславливает потребность разработки новых и модификации существующих методов извлечения АП.

При работе с базами данных, содержащих информацию о событиях, связанных во времени, целесообразно синтезировать временные АП (temporal association rules) [14]. Существенным отличием баз данных типа $D$ (1) от тех, с которыми приходится иметь дело при извлечении временных АП, является наличие информации о времени транзакций $T_j$. Поэтому большинство понятий и определений, приведенных для

позитивных АП являются верными и для временных АП. Исключение составляет поддержка правил.

К временным АП относятся [14]:

а) АП, описывающие зависимости, связанные с некоторыми интервалами времени. Такие правила могут быть представлены в виде: "В интервал времени $time(X)$ истинным является выражение: Если $X$, то $Y$". Временным интервалом $time(X)$ набора элементов $X$ считается интервал времени, на протяжении которого выполняется этот набор (11):

$$time(X) = [time(T_f\}; \ time(T_e)], \tag{11}$$

где $time(T_f)$ и $time(T_e)$ – времена первой $T_f$ и последней $T_e$ транзакций, содержащих набор элементов $X$, $time(T_f) < time(T_e)$.

При таком подходе поддержкой $\text{supp}(X)$ набора элементов $X$ является отношение количества $N_{T_j \in D|X \subseteq T}$ транзакций $T_j$ базы данных $D$, содержащих набор $X$, к количеству $N_{T_j \in D|time(t) \subseteq time(X)}$ транзакций, происходящих во временном интервале $time(X)$. Поддержка $\text{supp}(X)$ в таком случае вычисляется по формуле (12):

$$\text{supp}(X) = \frac{N_{T_j \in D|X \subseteq T}}{N_{T_j \in D|time(t) \subseteq time(X)}}; \tag{12}$$

б) циклические АП – описывают регулярные циклические действия во времени. Например, к таким АП могут быть отнесены правила, которые являются истинными в определенный промежуток времени каждого дня. Такие правила могут быть представлены в виде: "В интервал времени $time(X)$ с периодичностью $Period$ истинным является выражение: Если $X$, то $Y$".

Пусть $\Delta t$ – некоторый интервал времени. Тогда транзакция $T_j$ выполняется в $j$-й интервал времени $time(T_j)$: $[j \cdot \Delta t; \ (j+1) \cdot \Delta t]$. Обозначим $ST(time(T_j))$ – множество транзакций, выполняемых в $j$-й интервал времени $time(T_j)$: $ST(time(T_j)) = \{T \in D \mid time(T) \subseteq time(T_j)\}$. Циклическая поддержка $\text{supp}(X \to Y, \ time(T_j))$ АП $X \to Y$ в $j$-й интервал времени $time(T_j)$ отношение количества транзакций в

множестве $ST(time(T_j))$ к общему количеству транзакций в базе данных $D$. Аналогичным образом определяется и достоверность $\text{conf}(X \to Y, time(T_j))$ правила $X \to Y$ в $j$-й интервал времени (13):

$$\text{conf}(X \to Y, time(T_j)) = \frac{\text{supp}(X \to Y, \ time(T_j))}{\text{supp}(X, \ time(T_j))}. \qquad (13)$$

Определим цикл как пару $c = (Period, O)$, в которой $Period$ – период или длина цикла (например, $L$ = 24 часа, $\Delta t$ = час), $O$ – смещение, т.е. конкретное время срабатывания правила, $0 \le O \le Period$. Таким образом, АП $X \to Y$ является циклическим с циклом $c = (Period, O)$, если оно срабатывает в каждый $O$-й интервал времени с периодичностью $Period$.

Для извлечения временных АП, аналогично поиску других видов правил, задаются пороговые значения минимальной поддержки, достоверности и интересности правила: minsupport, minconfidence и $\varepsilon_I$, соответственно.

Однако непосредственное применение методов поиска позитивных АП для извлечения временных АП связано с проблемой вычисления интересности правил с учетом времени их выполнения [2, 8 – 10, 14]. Кроме того, существуют проблемы извлечения временных АП с разными периодами срабатывания правил, поиска оптимальных интервалов дискретизации времени $\Delta t$ и др. Поэтому целесообразной является разработка новых методов извлечения временных АП, устраняющих недостатки существующих методов.

Пусть транзакционная база данных $D$ характеризуется набором транзакций $T_j$, состоящих из элементов $\tau_a \in I = \{\tau_1, \tau_2, ..., \tau_{N_I}\}$, а также неотрицательных весовых коэффициентов (весов) $W = \{w_1, w_2, ..., w_{N_I}\}$, где $w_a$ – вес элемента $\tau_a$, $a = 1, 2, ..., N_I$.

Понятие нечетких АП (fuzzy association rules) связано с численными правилами [15]. При этом основные понятия и определения относительно бинарных (позитивных) АП расширяются на нечеткие АП для численных транзакций. Как отмечено выше, числовые значения элементов $\tau_a$ при поиске численных АП должны быть дискретизированы – разбиты на непересекающиеся интервалы, каждый из которых рассматривается в дальнейшем как отдельный атрибут или терм. С целью устранения проблемы неэффективного разбиения на интервалы при извлечении АП в

случае наличия в транзакционной базе данных $D$ числовых переменных используется теория нечетких множеств, в сочетании с теорией АП позволяющая извлекать нечеткие АП и строить на их основе нечеткие базы правил.

Нечеткая поддержка $\text{supp}(X, T_j)$ набора $X$ в транзакции $T_j$ определяется по формуле (14) [15]

$$\text{supp}(X, T_j) = \bigcap_{i=1}^{|X|} \mu_{X_i}(T_j), \tag{14}$$

где $\mu_{X_i}(T_j)$ – функция принадлежности переменной $X_i$ множеству $T_j$.

Нечеткая поддержка $\text{supp}(X)$ набора $X$ по всей транзакционной базе данных $D$ может быть вычислена аналогично по формуле (15)

$$\text{supp}(X) = \sum_{T_j : X \subseteq T_j} \text{supp}(X, T_j) = \sum_{T_j : X \subseteq T_j} \prod_{i=1}^{|X|} \mu_{X_i}(T_j). \tag{15}$$

При извлечении нечетких АП используются понятия взвешенной нечеткой поддержки набора элементов $X$ и правила $X \to Y$ [15].

Взвешенная нечеткая поддержка $\text{wsupp}(X)$ набора элементов $X$ определяется как произведение поддержка $\text{supp}(X)$ на сумму весовых коэффициентов $w_a$ элементов $\tau_a$, присутствующих в наборе $X$ (16)

$$\text{wsupp}(X) = \left( \sum_{\tau_a \in X} w_a \right) \text{supp}(X). \tag{16}$$

Взвешенная нечеткая поддержка правила $X \to Y$ вычисляется по формуле (17)

$$\text{wsupp}(X \to Y) = \left( \sum_{\tau_a \in X \cup Y} w_a \right) \text{supp}(X \cup Y). \tag{17}$$

Взвешенная нечеткая достоверность правила $X \to Y$ вычисляется по формуле (18)

$$\text{wconf}\left(X \to Y\right) = \frac{\text{wsupp}\left(X \to Y\right)}{\text{wsupp}\left(X\right)} = \frac{\left(\sum\limits_{\tau_a \in X \cup Y} w_a\right) \text{supp}\left(X \cup Y\right)}{\left(\sum\limits_{\tau_a \in X} w_a\right) \text{supp}\left(X\right)}. \qquad (18)$$

Для извлечения нечетких АП задаются уровни минимальной взвешенной поддержки $\text{wminsupport}$, минимальной взвешенной достоверности $\text{wminconfidence}$ и минимальной интересности правила $\varepsilon_I$. Если взвешенная нечетка поддержка $\text{wsupp}(X)$ набора элементов $X$ не меньше минимально допустимого значения $\left(\text{wsupp}(X) \geq \text{wminsupport}\right)$, то набор $X$ считается часто встречаемым. Аналогично, если значение взвешенной нечеткой достоверности $\text{wconf}\left(X \to Y\right)$ правила $X \to Y$ не меньше минимально допустимого значения $\left(\text{wconf}\left(X \to Y\right) \geq \text{wminconfidence}\right)$, то правило $X \to Y$ считается достоверным. Кроме того, уровень интересности правила должен быть не менее заданного порогового значения $\varepsilon_I$.

Методы извлечения нечетких АП на начальном этапе выполняют преобразование каждой числовой переменной к нечеткому множеству с соответствующими лингвистическими термами, используя для этого функции принадлежности. Далее выполняется расчет скалярной мощности каждого лингвистического терма по всей базе данных *D*, а также вычисление поддержки наборов элементов, используя итеративный подход для поиска больших наборов данных. Затем выполняется поиск нечетких АП из этих больших наборов данных.

Каждый элемент используется только в лингвистическом терме с максимальной мощностью на более поздних итерациях, в результате чего количество обрабатываемых нечетких областей такое же, как количество элементов $\tau_a$ в множестве $I = \{\tau_1, \tau_2, ..., \tau_{N_I}\}$. Таким образом, основной акцент осуществляется на наиболее важные лингвистические термы, что снижает временную сложность таких методов [15].

Однако необходимость выделения нечётких термов при извлечении АП требует решения задачи кластерного анализа (в результате чего термы могут быть определены как проекции границ кластеров на оси признаков), что усложняет процесс поиска нечетких АП, либо участия пользователя – эксперта в прикладной области, что уменьшает уровень автоматизации программных средств извлечения нечетких АП на основе соответствующих методов. Необходимость устранения указанных

недостатков обуславливает потребность разработки новых методов поиска нечетких АП.

Таким образом, с целью устранения выявленных недостатков существующих методов извлечения АП целесообразно разработать новые и модифицировать существующие методы для поиска негативных, численных, обобщенных, временных и нечетких ассоциативных правил.

**Выводы**. В работе решена актуальная задача исследования видов ассоциативных правил. Показано, что для обработки больших массивов неструктурированных данных целесообразно использовать ассоциативные правила, которые позволяют синтезировать базы правил, удобные для дальнейшего восприятия и анализа экспертами в прикладных областях.

Проанализирован процесс извлечения ассоциативных правил. Отмечено, что проанализированные методы не позволяют решать задачи отбора информативных признаков, кластерного анализа, построения моделей на основе больших массивов неструктурированных данных и др., возникающие при решении реальных практических задач прогнозирования, классификации и кластеризации данных.

Проанализированы основные виды АП (негативные, численные, обобщенные, временные и нечеткие АП), необходимость построения которых возникает при решении реальных задач. Показано, что существующие методы извлечения бинарных АП неэффективно функционируют при извлечении особых видов АП, что обуславливает необходимость разработки новых и модификации существующих методов для поиска негативных, численных, обобщенных, временных и нечетких АП.

С целью устранения выявленных недостатков предлагается:

– разработать эффективные методы извлечения негативных, численных, обобщенных, временных и нечетких АП;

– создать метод отбора информативных признаков с использованием АП;

– разработать метод кластерного анализа на основе АП;

– создать метод построения нейро-нечетких сетей с использованием АП;

– выполнить программную реализацию предложенных методов. На основе разработанного программного обеспечения исследовать предложенные методы, выполнить их сравнение с существующими аналогами, решить практические задачи распознавания образов.

Работа выполнена в рамках госбюджетной НИР кафедры радиотехники и телекоммуникаций Запорожского национального

технического университета "Методы, модели и устройства принятия решений в системах распознавания образов" (№ гос. регистрации 0111U000059), а также в рамках НИР ДБ 04922 "Интеллектуальные информационные технологии автоматизации проектирования, моделирования, управления и диагностирования производственных процессов и систем".

**Список літератури:** **1.** *Shin Y.C.* Intelligent systems: modeling, optimization, and control / *C.Y. Shin, C. Xu*. – Boca Raton: CRC Press, 2009. – 456 p. **2**. *Zhang C*. Association rule mining: models and algorithms / *C. Zhang, S. Zhang*. – Berlin: Springer-Verlag. – 2002. – 238 p. **3**. *Ding S.X*. Model-based fault diagnosis techniques: design schemes, algorithms, and tools / *S.X. Ding*. – Berlin: Springer, 2008. – 473 p. **4**. Encyclopedia of artificial intelligence / Eds.: *J.R. Dopico, J. D. de la Calle, A.P. Sierra*. – New York: Information Science Reference, 2009. – Vol. 1–3. – 1677 p. **5**. Интеллектуальные информационные технологии проектирования автоматизированных систем диагностирования и распознавания образов: монография / [*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. **6**. Гибридные нейро-фаззи модели и мультиагентные технологии в сложных системах: монография / [*В.А. Филатов, Е.В. Бодянский, В.Е. Кучеренко* и др.; под общ. ред. Е.В. Бодянского). – Дніпропетровськ: Системні технології, 2008. – 403 с. **7**. *Айвазян С.А.* Прикладная статистика: Исследование зависимостей / *С.А. Айвазян, И.С. Енюков, Л.Д. Мешалкин*. – М.: Финансы и статистика, 1985. – 487 с. **8**. *Zhao Y.* Post-mining of association rules: techniques for effective knowledge extraction / *Y. Zhao, C. Zhang, L. Cao*. – New York: Information Science Reference. – 2009. – 372 p. **9**. *Adamo J.-M.* Data mining for association rules and sequential patterns: sequential and parallel algorithms / *J.-M. Adamo*. – New York: Springer-Verlag. – 2001. – 259 p. **10**. *Koh Y.S*. Rare Association Rule Mining and Knowledge Discovery / *Y.S. Koh, N. Rountree*. – New York: Information Science Reference. – 2009. – 320 p. **11**. *Piao X*. Research on Mining Positive and Negative Association Rules Based on Dual Confidence / *X. Piao, Z. Wang, G. Liu* // Internet Computing in Science and Engineering: The Fifth International Conference ICICSE, Harbin, China, 1–2 November 2010: Proceedings of the Conference. – Washington: IEEE Press, 2010. – P. 102-105. **12**. *Ke Y.* An information-theoretic approach to quantitative association rule mining / *Y. Ke, J. Cheng, N. Wilfred* // Knowledge and Information Systems. – 2008. – Vol. 16. – № 2. – P. 213-244. **13**. *Wu C.* Generalized association rule mining using an efficient data structure / *C. Wu, Y. Huang* // Expert Systems With Applications. – 2011. – Vol. 38. – № 6. – P. 7277-7290. **14**. *Nam H.* Identification of temporal association rules from time-series microarray data sets / *H. Nam, K. Y. Lee, D. Lee* // BMC Bioinformatics. – 2009. – Vol. 10. – № 3. – P. 6-9. **15**. *Pach F.P.* Compact fuzzy association rule-based classifier / *F.P. Pach, A. Gyenesei, J. Abonyi* // Expert Systems With Applications. – 2008. – Vol. 34. – № 4. – P. 2406-2416.

*Поступила в редакцию 03.09.2012*

*Работу представил декан математического факультета Запорожского національного университета, д-р техн. наук, проф. Гоменюк С.И.*.

УДК 004.93

**Асоціативні правила в інтелектуальному аналізі даних / Зайко Т.А., Олійник А.О., Субботін С.О.** // Вісник НТУ "ХПІ". Серія: Інформатика та моделювання. – Харків: НТУ "ХПІ". – 2013. – № 39 (1012). – С. 82– 96.

Розглянуто завдання побудови моделей на основі асоціативних правил. Проаналізовано процес пошуку асоціативних правил. Досліджено різні види асоціативних правил (негативні, чисельні, узагальнені, часові та нечіткі асоціативні правила) при використанні їх для розв'язання завдань інтелектуального аналізу даних. Бібліогр.: 15 назв.

**Ключові слова:** асоціативне правило, різні види асоціативних правил, інтелектуальний аналіз даних, нечіткі асоціативні правила.

UDC 004.93

**Association rules in data mining / Zayko T.A., Oliinyk A.A., Subbotin S.A.** // Herald of the National University "KhPI". Subject issue: Information science and Modeling. – Kharkov: NTU "KhPI". – 2013. – № 39 (1012). – С. 82– 96.

The problem of synthesis of models based on association rules is concidered. The process of mining association rules is analyzed. Various types of association rules (negative, quantitative, generalized, temporal and fuzzy association rules) for solving data mining problems are investigated. Refs.: 15 titles.

**Keywords:** association rule, various types of association rules, data mining, fuzzy association rules.