АВТОМАТИКА И ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА. – 2014. – No.2. – С.34-47.

ФАКТОРНЫЙ АНАЛИЗ ТРАНЗАКЦИОННЫХ БАЗ ДАННЫХ

А.А. ОЛЕЙНИК, *кандидат технических наук, доцент*
Т.А. ЗАЙКО, *аспирантка*
С.А. СУББОТИН, *кандидат технических наук, профессор*
Запорожский национальный технический университет,
ул. Жуковского 64, Запорожье 69063, Украина
E-mail: subbotin@zntu.edu.ua

Рассмотрено решение задачи факторного анализа. Разработан метод факторного анализа, позволяющий обрабатывать данные, представленные в виде баз транзакций. Предложенный метод предусматривает извлечение правил из заданных баз транзакций, в результате чего выполняется обобщение данных, и, соответственно, исключение из дальнейшего рассмотрения избыточных признаков, что позволяет сократить пространство поиска и время выполнения факторного анализа. Выполнен анализ вычислительной сложности разработанного метода. Проведены эксперименты по решению практических и тестовых задач.

Ключевые слова: ассоциативное правило, база транзакций, терм, транзакция, факторный анализ, эволюционный поиск

1. ВВЕДЕНИЕ

Сложные технические и медицинские объекты, процессы и системы характеризуются, как правило, большим количеством переменных, некоторые из которых взаимосвязаны [1]. Для поиска таких скрытых зависимостей применяют методы факторного анализа [2], позволяющего выявлять взаимосвязи между различными признаками, характеризующими исследуемые объекты или процессы, объясняя таким образом их внутреннюю природу.

Однако применение известных методов факторного анализа [3] возможно при выполнении ряда условий [2–4]: признаки, описывающие исследуемые объекты или процессы, должны быть численными; количество экземпляров выборки должно превышать не менее чем в два раза число признаков; однородность обучающей выборки; симметричность распределения признаков обучающей выборки. Выборки данных, характеризующие реальные технические и медицинские объекты, не всегда удовлетворяют приведенным выше условиям. Кроме того, ряд задач диагностирования и классификации связан с необходимостью обработки выборок данных, представленных в виде баз транзакций, в которых каждая транзакция описывает некоторый набор характеристик конкретного объекта или процесса, при этом количество признаков в разных транзакциях может быть разным [5].

Поэтому актуальной является разработка метода факторного анализа, свободного от указанных недостатков и позволяющего обрабатывать данные, представленные в виде баз транзакций.

Целью настоящей работы является создание метода факторного анализа на основе ассоциативных правил для поиска скрытых зависимостей в транзакционных базах данных.

2. ПОСТАНОВКА ЗАДАЧИ ФАКТОРНОГО АНАЛИЗА В ТРАНЗАКЦИОННЫХ БАЗАХ ДАННЫХ

Пусть задана база транзакций $D = \{T_1, T_2, ..., T_{N_D}\}$, в которой каждый элемент T_j, $j = 1, 2, ..., N_D$ содержит информацию о некоторых взаимосвязанных событиях, где $N_D = |D|$ – количество транзакций в наборе данных *D*.

Элементы T_j могут представляться таким образом:

$$T_j = (tid_j, item_j),$$

где tid_j – идентификатор *j*-й транзакции T_j;

$item_j = \{t_{1j}, t_{2j}, ..., t_{N_{item_j} j}\} \subseteq I$ – список элементов, входящих в транзакцию T_j;

t_{ij} – *i*-й элемент списка $item_j$, $i = 1, 2, ..., N_{item_j}$;

$N_{item_j} = |item_j|$ – количество элементов множества $item_j$;

$I = \{\tau_1, \tau_2, ..., \tau_{N_I}\}$ – 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 вещественные переменные, элементы t_{ij} транзакции T_j представляются кортежем:

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

где τ_{ij} – признак из множества *I*, соответствующий элементу t_{ij};

$v(\tau_{ij})$ – значение признака τ_{ij} в транзакции T_j, $v(\tau_{ij}) \in \Delta_{ij} = [\tau_{ij\min}; \tau_{ij\max}]$;

$\tau_{ij\min}$ и $\tau_{ij\max}$ – минимальное и максимальное значения из диапазона возможных значений Δ_{ij} признака τ_{ij}.

Тогда задача факторного анализа заключается в выявлении набора $\Psi = \{\Psi_1, \Psi_2, ..., \Psi_{N_\Psi}\}$ ($N_\Psi = |\Psi| \le N_I$), состоящего из факторов Ψ_d, каждый из которых характеризует группу тесно связанных признаков $\tau \in I$. Таким образом, в результате факторного анализа обеспечивается сжатие информации путем объединения множества признаков через некоторый набор факторов.

3. МЕТОД ФАКТОРНОГО АНАЛИЗА НА ОСНОВЕ АССОЦИАТИВНЫХ ПРАВИЛ

Как отмечалось выше, известные методы факторного анализа [2–4, 6, 7] требуют достаточно большого количества экземпляров в обучающих выборках, наличия исключительно количественных признаков, однородности обучающей выборки, а также не предназначены для обработки данных, представленных в виде баз транзакций. С целью устранения указанных недостатков и возможности выполнения факторного анализа в транзакционных базах данных в разработанном методе выполняется извлечение ассоциативных правил, что позволяет исключить из дальнейше-

го рассмотрения избыточные признаки, сократив тем самым пространство поиска и уменьшив время факторного анализа.

Разработанный метод факторного анализа в транзакционных базах данных на основе ассоциативных правил состоит из следующих этапов: инициализация, извлечение ассоциативных правил и построение базы правил, выделение термов признаков, определение эквивалентности термов и признаков, поиск групп качественно близких признаков.

В предложенном методе факторного анализа на начальном этапе задается транзакционная база данных D, которая может содержать как численные, так и бинарные или качественные признаки.

Затем из заданной базы транзакций D извлекаются ассоциативные правила, используя известные методы поиска таких правил [8–14] $D \to \mathrm{RB}$, в результате чего выполняется обобщение данных, и, соответственно, исключение из дальнейшего рассмотрения избыточных признаков, а также некоторых термов неизбыточных признаков. Это позволяет сократить пространство поиска и время выполнения факторного анализа.

В процессе поиска численных ассоциативных правил [5, 8, 10] выполняется разбиение диапазона значений численных признаков τ_a на некоторые интервалы. В результате такого перехода от непрерывных переменных к дискретным интервалам происходит потеря информации, что приводит к возникновению ошибок дискретизации e_d. При дискретизации с округлением значение признака $\tau_a \in I$ заменяется на среднее значение соответствующего интервала $\Delta\tau_{ak} \in \left[\Delta\tau_{ak\min};\ \Delta\tau_{ak\max}\right]$, поэтому максимальное значение ошибки дискретизации будет равно половине интервала разбиения признаков: $e_{d\max} = 0{,}5\,\Delta\tau_a$.

Оценим также среднее значение ошибки дискретизации $\overline{e_d}$ и ее дисперсию $D(e_d)$. Будем считать, что значения признака $\tau_a \in I$ распределяются равномерно по всему диапазону его возможных значений $\Delta\tau_a \in \left[\Delta\tau_{a\min};\ \Delta\tau_{a\max}\right]$. Предполагая, что длина интервалов $\Delta\tau_{ak}$ дискретизации признака τ_a является достаточно малой по сравнению с диапазоном его значений $\Delta\tau_a$ ($\Delta\tau_{ak} << \Delta\tau_a$, $\forall k$), плотность $w(e_d)$ распределения ошибки дискретизации e_d в пределах интервала $\Delta\tau_{ak}$ может быть оценена как $w(e_d) = (\Delta\tau_{ak})^{-1}$.

Следовательно, среднее значение ошибки дискретизации $\overline{e_d}$ и ее дисперсия $D(e_d)$. могут быть вычислены по формулам (1) и (2), соответственно:

$$\overline{e_d} = \int_{-\Delta\tau_{ak}/2}^{\Delta\tau_{ak}/2} e_d\, w(e_d)\, de_d = \left(\Delta\tau_{ak}\right)^{-1} \int_{-\Delta\tau_{ak}/2}^{\Delta\tau_{ak}/2} e_d\, de_d = \\ = \frac{{e_d}^2}{2\Delta\tau_{ak}}\Bigg|_{-\Delta\tau_{ak}/2}^{\Delta\tau_{ak}/2} = \frac{1}{2\Delta\tau_{ak}}\left(\frac{{\Delta\tau_{ak}}^2}{4} - \frac{{\Delta\tau_{ak}}^2}{4}\right) = 0. \tag{1}$$

$$D\left(e_d\right) = \int_{-\Delta\tau_{ak}/2}^{\Delta\tau_{ak}/2} \left(e_d - \overline{e_d}\right)^2 w(e_d)\, de_d = \left(\Delta\tau_{ak}\right)^{-1} \int_{-\Delta\tau_{ak}/2}^{\Delta\tau_{ak}/2} {e_d}^2\, de_d = \frac{{e_d}^3}{3\Delta\tau_{ak}}\Bigg|_{-\Delta\tau_{ak}/2}^{\Delta\tau_{ak}/2} = \\ = \frac{1}{3\Delta\tau_{ak}}\left(\frac{{\Delta\tau_{ak}}^3}{8} - \left(-\frac{{\Delta\tau_{ak}}^3}{8}\right)\right) = \frac{{\Delta\tau_{ak}}^2}{12}. \tag{2}$$

Полученные оценки являются вполне приемлемыми и показывают, что накопление погрешности при дискретизации не происходит ($\overline{e_d} = 0$) при несущественной дисперсии ошибки дискретизации.

Далее выполняется упрощение синтезированной базы ассоциативных правил RB [1, 9], объединяя по возможности некоторые правила.

После этого на основе построенной базы правил RB выделяются термы признаков $\tau_a \in I$. Для этого анализируется каждое ассоциативное правило из базы RB ($\mathrm{AR}_l \in \mathrm{RB}$), в результате чего формируются массивы термов каждого из признаков $\tau_a \in I$: $\Delta\tau_a = \left\{\Delta\tau_{a1}, \Delta\tau_{a2}, ..., \Delta\tau_{aN_{\Delta\tau_a}}\right\}$,

где $\Delta\tau_{ak} \in \left[\Delta\tau_{ak\min}; \Delta\tau_{ak\max}\right]$ – k-й терм (интервал) a-го признака;

$\Delta\tau_{ak\min}$ и $\Delta\tau_{ak\max}$ – минимальное и максимальное значения в k-ом терме a-го признака, соответственно;

$N_{\Delta\tau_a}$ – количество термов a-го признака.

Важно отметить, что не обязательно границы соседних интервалов пересекаются (условие $\Delta\tau_{ak\max} = \Delta\tau_{a(k+1)\min}$ не должно выполняться для всех k), поскольку ранее были синтезированы ассоциативные правила и, соответственно, исключены избыточные (неинформативные) термы некоторых признаков.

Затем определяется эквивалентность термов признаков. Будем считать, что термы тем эквивалентнее, чем выше вероятность (частота) того, что экземпляры (ассоциативные правила), попавшие в один терм $\Delta\tau_{ak}$ первого признака $\tau_a \in I$, попадут в другой терм $\Delta\tau_{bm}$ второго признака $\tau_b \in I$. Поэтому для определения эквивалентности термов признаков будем рассчитывать частоту попадания ассоциативных правил в термы различных признаков (3):

$$\alpha_M = \frac{\sum_{l=1}^{N_{\mathrm{RB}}} \beta_{Ml}}{N_{\mathrm{RB}}}, \tag{3}$$

где $M = \langle a, b, k, m \rangle$ – кортеж, определяющий взаимосвязь k-го терма $\Delta\tau_{ak}$ a-го признака $\tau_a \in I$ и m-го терма $\Delta\tau_{bm}$ b-го признака $\tau_b \in I$; N_{RB} – количество правил в базе RB; β_{Ml} – величина, определяющая наличие связи между термами признаков картежа M в l-ом правиле $\mathrm{AR}_l \in \mathrm{RB}$ синтезированной базы правил RB (4):

$$\beta_{Ml} = \begin{cases} 1, \left(\Delta\tau_{ak} \wedge \Delta\tau_{bm}\right) \in AR_l; \\ 0, \left(\Delta\tau_{ak} \wedge \Delta\tau_{bm}\right) \notin AR_l. \end{cases} \tag{4}$$

После определения эквивалентности термов α_M определяется эквивалентность признаков γ_{ab} (5):

$$\gamma_{ab} = \frac{\sum_{m=1}^{N_{\Delta\tau_b}} \sum_{k=1}^{N_{\Delta\tau_a}} \alpha_{abkm}}{N_{\Delta\tau_b} N_{\Delta\tau_a}}. \tag{5}$$

Для формирования групп $\Psi = \{\Psi_1, \Psi_2, ..., \Psi_{N_\Psi}\}$ близких признаков создается массив признаков I' как копия массива $I = \{\tau_1, \tau_2, ..., \tau_{N_I}\}$: $I' = I$. Затем выбираются два признака $\tau_A \in I'$ и $\tau_B \in I'$ с наибольшей оценкой γ_{AB} эквивалентности (6):

$$\tau_A, \tau_B : \gamma_{AB} = \max_{a,b=1,2,...,N_{I'}} \gamma_{ab}, \quad (6)$$

где $N_{I'}$ – количество признаков во множестве I'.

В случае, если признаки $\tau_A \in I'$ и $\tau_B \in I'$ абсолютно эквивалентны (при наличии *k*-го терма $\Delta\tau_{Ak}$ *A*-го признака $\tau_A \in I'$ в *l*-ом правиле $\mathrm{AR}_l \in \mathrm{RB}$ базы правил RB также будет находиться *k*-й терм $\Delta\tau_{Bk}$ *B*-го признака $\tau_B \in I'$, и наоборот), в *d*-ю группу эквивалентности Ψ_d включается только один признак, например, τ_A: $\Psi_d = \Psi_d \cup \tau_A$. В случае неабсолютной эквивалентности признаков τ_A и τ_B в *d*-ю группу эквивалентности Ψ_d включаются оба признака: $\Psi_d = \Psi_d \cup \{\tau_A, \tau_B\}$.

Признаки τ_A и τ_B исключаются из дальнейшего рассмотрения как такие, которые уже входят в некоторую группу эквивалентности: $I' = I' \setminus \{\tau_A, \tau_B\}$.

Затем находится следующий по значению оценки эквивалентности γ признак $\tau_C \in I'$, после чего он включается в текущую группу близких признаков Ψ_d и исключается из множества I': $\Psi_d = \Psi_d \cup \tau_C$, $I' = I' \setminus \{\tau_C\}$.

Формирование группы Ψ_d эквивалентных признаков продолжается до тех пор, пока выполняется условие наличия признаков τ_a во множестве I' с минимально приемлемым значением оценки эквивалентности с каждым из признаков в наборе Ψ_d (7):

$$\exists \tau_a \in I' : \gamma_{ab} \geq \gamma_{\min}, \quad b = 1, 2, ..., N_{\Psi_d}, \quad (7)$$

где $\gamma_{\min}$ – минимально приемлемое значение оценки γ эквивалентности признаков; N_{Ψ_d} – количество элементов во множестве Ψ_d.

После формирования группы Ψ_d аналогичным образом происходит генерация группы Ψ_{d+1}. Формирование групп Ψ происходит до тех пор, пока выполняется выше приведенное условие. Оставшиеся во множестве I' признаки считаются такими, которые не могут быть объединены в группы эквивалентности Ψ. Таким образом сформированное множество групп Ψ является решением задачи факторного анализа, при этом каждая группа Ψ_d характеризует набор тесно связанных признаков $\tau \in I$.

При необходимости извлечения искусственных признаков (латентных переменных) на основе факторных групп, сформированных из исходных признаков, выполняется этап конструирования факторов, количественно характеризующих соответствующие признаки. Для этого должны быть заданы критерий оценивания качества преобразования и множество Ω возможных преобразований $\omega_1, \omega_2, ...\omega_{N_\Omega}$ (например, аддитивных, мультипликативных, полиномиальных и др.), с помощью которых выполняется конструирование искусственных признаков $\upsilon = \{\upsilon_1, \upsilon_2, ..., \upsilon_{N_\upsilon}\}$ на основе признаков из соответствующих факторных групп $\Psi = \{\Psi_1, \Psi_2, ..., \Psi_{N_\Psi}\}$, где $N_\upsilon = N_\Psi$ – количество искусственно создаваемых переменных. Затем с помощью известных методов конструирования признаков (например, с помощью метода главных компонент) происходит синтез новых латентных переменных, количественно описывающих исходные признаки.

4. ЭВОЛЮЦИОННЫЙ МЕТОД ФОРМИРОВАНИЯ ФАКТОРНЫХ ГРУПП

Однако иногда результаты факторного анализа могут оказаться неприемлемыми: найдено слишком малое количество групп эквивалентности Ψ_d ($N_\Psi \le N_{\Psi\min}$), или размер этих групп менее минимально допустимого ($N_{\Psi_d} < N_{\Psi_d\min}$, $\Psi_d \in \Psi$). Кроме внутренней природы исследуемых объектов, процессов или систем такое неприемлемое решение в некоторых случаях может объясняться жадным характером поиска групп эквивалентности на последнем этапе разработанного метода факторного анализа на основе ассоциативных правил. Использование жадной стратегии поиска обусловлено незначительным временем ее выполнения. Однако в случае ее применения при выборе одного признака и включении его в группу Ψ_d не представляется возможным в дальнейшем его исключить из данной группы и ввести в другую.

Поэтому в некоторых случаях целесообразно этап формирования групп $\Psi = \{\Psi_1, \Psi_2, ..., \Psi_{N_\Psi}\}$ близких признаков выполнять с помощью оптимизационных процедур. Поскольку поиск факторных групп необходимо выполнять в дискретном множестве признаков, для решения данной задачи целесообразным является применение методов эволюционной оптимизации [15–18], которые являются методами глобального поиска, не выдвигают требований к целевой функции и входным переменным, а также являются методами дискретной оптимизации.

С целью применения эволюционного поиска для факторного анализа определим способ представления решения (множества факторных групп $\Psi = \{\Psi_1, \Psi_2, ..., \Psi_{N_\Psi}\}$) в виде хромосом $H: \Psi \to H$. Поскольку размер множества Ψ (количество факторных групп N_Ψ) заранее не известен, размер хромосом N_{H_j} будет переменным: $N_{H_j^{(t)}} \ne N_{H_j^{(t+1)}}$, $N_{H_j^{(t)}} \ne N_{H_k^{(t)}}$, где $N_{H_j^{(t)}}$ – размер *j*-й хромосомы H_j на *t*-й итерации эволюционного поиска. Гены h_{ji} хромосомы H_j будут соответствовать *i*-му элементу множества $\Psi \to H_j$. Таким образом, каждая *j*-ая хромосома *t*-й популяции $H_j^{(t)}$ будет соответствовать *j*-му решению на *t*-й итерации эволюционного поиска, представляющему собой *j*-ое множество факторных групп (8):

$$H_j^{(t)} \to \Psi_j^{(t)} = \left\{ \Psi_{1j}^{(t)}, \Psi_{2j}^{(t)}, ..., \Psi_{N_{\Psi_j^{(t)}} j}^{(t)} \right\}. \qquad (8)$$

Гены h_{ji} хромосом H_j представляют собой векторы целых чисел, соответствующих номерам признаков $\tau_a \in I$ из множества *I*.

Таким образом, предлагается представлять хромосомы $H_j^{(t)}$ в виде векторов переменной длины, каждый элемент (ген) которых также является массивом элементов переменной длины. Причем гены в векторных хромосомах при таком способе кодирования имеют свойства негомологичности, т.е. числа в каждом векторе могут принимать значения из заданного множества и не должны повторяться [15–18].

Пример хромосомы при решении задачи факторного анализа приведен на рис.1.

Как видно из рис.1, некоторые признаки (τ_3, τ_7, τ_8 и др.) не вошли в хромосому H_j, что свидетельствует о том, что данные признаки не являются членами ни одной из четырех факторных групп $\Psi_j = \{\Psi_{j1}, \Psi_{j2}, \Psi_{j3}, \Psi_{j4}\}$, определенных в хромосоме $H_j \to \Psi_j$ в виде генов h_{j1}, h_{j2}, h_{j3}, h_{j4}, соответственно. Множество не вошедших в хромосому H_j признаков (τ_3, τ_7, τ_8 и др.) является аналогией множеству I' при использовании жадной стратегии. Гены h_{j1}, h_{j2}, h_{j3}, h_{j4} хромосомы H_j имеют различную длину, соответствующую количеству признаков в соответствующей факторной группе.

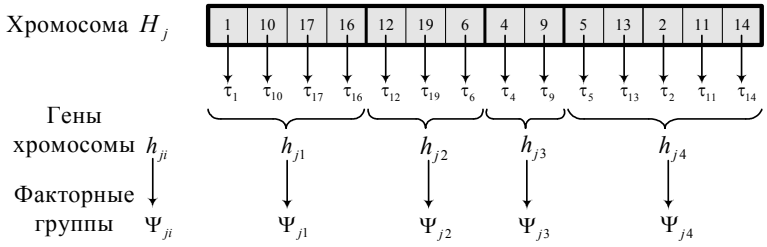

Рис.1. Пример хромосомы при решении задачи факторного анализа.

Для оценивания хромосом H_j предлагается использовать целевую функцию γ_{H_j} (9), учитывающую эквивалентность γ признаков τ_a в каждом из генов h_{ji} (факторной группе):

$$\gamma_{H_j} = \left(1 + \sum_{i=1}^{N_{H_j}} \gamma_{h_{ji}}\right)^{-1} N_{H_j} \to \min, \tag{9}$$

где $\gamma_{h_{ji}}$ – среднее значение эквивалентности признаков в *i*-м гене h_{ji} *j*-й хромосомы H_j (10):

$$\gamma_{h_{ji}} = \left(\frac{1}{2} N_{h_{ji}} \left(N_{h_{ji}} - 1\right)\right)^{-1} \sum_{\substack{a=1 \\ a,b:\, \tau_a, \tau_b \in h_{ji}}}^{N_{h_{ji}}} \sum_{b=a+1}^{N_{h_{ji}}} \gamma_{ab}, \tag{10}$$

где $N_{h_{ji}}$ – размер *i*-го гена *j*-й хромосомы.

Инициализация эволюционного поиска в предлагаемом методе происходит путем генерации заданного количества хромосом N_H, каждая из которых H_j соответствует некоторому множеству Ψ_j факторных групп $\Psi_1, \Psi_2, ..., \Psi_{N_\Psi}$.

Генерация *j*-й хромосомы H_j начинается со случайного формирования массива целых неповторяющихся чисел из интервала $[1;\ N_I]$, каждое из которых соответствует определенному признаку $\tau_a \in I$. Затем созданный массив разбивается на некоторое случайно сгенерированное количество $N_{H_j} << N_I$ генов h_{ji}, представляющих собой определенные факторные группы переменной длины. Таким образом, на начальном этапе эволюционного поиска считается, что каждый признак $\tau_a \in I$ может входить в некоторую факторную группу.

В качестве оператора отбора предлагается использовать пропорциональный отбор [15–18], при котором к скрещиванию и мутации допускаются хромосомы H_j с уровнем приспособленности (значением целевой функции γ_{H_j}), не ниже средней по популяции.

Для скрещивания хромосом предлагается использовать модифицированный с учетом решаемой задачи оператор одноточечного скрещивания. Затем происходит обмен генами между хромосомами H_1 и H_2, в результате чего формируется два потомка

$$H_{\text{desc1}} = \left\{h_{11}, h_{12}, ..., h_{1\text{tr}}, h_{2(\text{tr}+1)}, h_{2N_{H_2}}\right\} \text{ и } H_{\text{desc2}} = \left\{h_{21}, h_{22}, ..., h_{2\text{tr}}, h_{1(\text{tr}+1)}, h_{1N_{H_1}}\right\}.$$

После формирования хромосом-потомков H_{desc1} и H_{desc2}, как правило, в них возникают недопустимые гены (в различных генах одной хромосомы содержатся одинаковые элементы, соответствующие одним и тем же признакам $\tau_a \in I$). Поскольку в скрещивании принимало участие

две хромосомы, в хромосомах-потомках могут содержаться гены h_1 и h_2 не более чем с двумя одинаковыми элементами $\tau_a \in I$. Для приведения хромосом $H_{\text{desc}1}$ и $H_{\text{desc}2}$ к допустимому виду из каждой из них исключаются повторяющиеся признаки $\tau_a \in I$. При этом, если в двух генах h_{j1} и h_{j2} одной хромосомы содержатся одинаковые элементы, соответствующие одному признаку $\tau_a \in I$, происходит исключение соответствующего элемента из гена h_{ji}, в котором он имеет меньшее влияние на общую эквивалентность $\gamma_{h_{ji}}$. Для определения влияния признака $\tau_a \in I$ в каждой из хромосом h_{j1} и h_{j2} вычисляются величины

$$\Delta\gamma_{h_{j1}\setminus\tau_a} = \gamma_{h_{j1}} - \gamma_{h_{j1}\setminus\tau_a} \quad \text{и} \quad \Delta\gamma_{h_{j2}\setminus\tau_a} = \gamma_{h_{j2}} - \gamma_{h_{j2}\setminus\tau_a},$$

где $\gamma_{h_{j1}\setminus\tau_a}$ и $\gamma_{h_{j2}\setminus\tau_a}$ – средние значения эквивалентности без признака τ_a в генах h_{j1} и h_{j2}, соответственно.

Признак $\tau_a \in I$ исключается из гена h_{ji} с меньшим значением величины $\Delta\gamma_{h_{ji}\setminus\tau_a}$. Таким образом, предложенный оператор скрещивания позволяет генерировать новое множество допустимых решений путем передачи информации от хромосом-родителей.

После скрещивания выполняется мутация случайно выбранных генов некоторых хромосом, позволяющая более детально исследовать пространство поиска и разнообразить его. Предлагается выполнять мутации вставки, обмена и удаления некоторых признаков из хромосом. Мутация вставки выполняется с целью добавления признаков в факторные группы. Для этого случайно выбранный признак $\tau_a \in I$, не содержащийся в хромосоме H_j ($\tau_a \notin H_j$), добавляется в один из ее генов. В случае, если выполняется условие $\gamma_{h_{ji}\cup\tau_a} \geq \gamma_{h_{ji}}$, добавленный признак τ_a остается в хромосоме H_j.

Мутация обмена предусматривает выбор и замену двух случайно отобранных признаков $\tau_a \in h_1$ и $\tau_b \in h_2$ из различных генов h_1 и h_2 одной хромосомы H_j, выбранной для мутации. Операция мутации обмена считается успешной при выполнении условия (11):

$$\Delta\gamma_{h_1\cup\tau_b\setminus\tau_a} + \Delta\gamma_{h_2\cup\tau_a\setminus\tau_b} > 0, \tag{11}$$

где $\Delta\gamma_{h_1\cup\tau_b\setminus\tau_a} = \gamma_{h_1} - \gamma_{h_1\cup\tau_b\setminus\tau_a}$ и $\Delta\gamma_{h_2\cup\tau_a\setminus\tau_b} = \gamma_{h_2} - \gamma_{h_2\cup\tau_a\setminus\tau_b}$ – приросты средней эквивалентности факторных групп, соответствующих генам h_1 и h_2 при обмене признаками τ_a и τ_b.

Мутация удалением направлена на исключение из факторной группы $\Psi_{ji} \leftarrow h_{ji}$ признака $\tau_a \in h_{ji}$, слабо коррелирующего с другими признаками $\tau_b \in h_{ji}$ из этой же группы Ψ_{ji}. Для этого случайным образом в отобранной для мутации хромосоме H_j выбирается мутирующий ген $h_{ji} \in H_j$, из которого исключается признак $\tau_a \in h_{ji}$, наименее влияющий на среднюю эквивалентность гена $\gamma_{h_{ji}}$ (12):

$$\tau_a : \Delta\gamma_{h_{ji}\setminus\tau_a} = \min_{\substack{b=1,2,\ldots,N_{h_{ji}} \\ \tau_b \in h_{ji}}} \Delta\gamma_{h_{ji}\setminus\tau_b}. \tag{12}$$

Таким образом, операция мутации в некоторых случаях изменяет размер хромосомы и позволяет сформировать факторные группы с более приемлемыми оценками эквивалентности. После скрещивания и мутации происходит формирование нового множества решений. На новую итерацию переходят хромосомы с наилучшими значениями целевой функции, а также хромосомы, полученные в результате скрещивания и мутации.

Эволюционная оптимизация продолжается до тех пор, пока не будет достигнуто максимально допустимое количество итераций N_{it} или не найдено решение $H_j \to \Psi_j$ с приемлемым значением целевой функции, не превышающим минимально допустимое значение $\gamma_{H_j} \le \gamma_{H\min}$. Использование эволюционного подхода для поиска групп качественно близких признаков позволяет более детально по сравнению с жадной стратегией исследовать пространство поиска, а также формировать группы близких признаков, характеризующиеся более приемлемыми оценками эквивалентности.

Предложенный метод факторного анализа на основе ассоциативных правил предусматривает извлечение правил из заданных баз транзакций, в результате чего выполняется обобщение данных, и, соответственно, исключение из дальнейшего рассмотрения избыточных признаков, что позволяет сократить пространство поиска и время выполнения факторного анализа. В разработанном методе определение эквивалентности признаков для формирования факторных групп выполняется исходя из частоты их совместного попадания в ассоциативные правила синтезированной базы правил, что позволяет оценивать тесноту связи между различными признаками (качественными, количественными), не выдвигать требований к входным данным и выполнять факторный анализ в транзакционных базах данных.

5. АНАЛИЗ ВЫЧИСЛИТЕЛЬНОЙ СЛОЖНОСТИ

Проанализируем вычислительную сложность предложенного метода. Пусть O – количество элементарных операций, которые могут быть использованы при выполнении конкретных действий с целью решения определенной задачи.

Тогда вычислительная сложность O метода может быть определена как сумма величин $O_1 + O_2 + ... + O_{N_{st}}$, характеризующих вычислительную сложность каждого из N_{st} соответствующего метода. Вычислительная сложность этапа извлечения ассоциативных правил O_{AR} определяется как $O_{\mathrm{AR}} = O_{\mathrm{AR}}\left(|I| \cdot N_D \log_2(N_D) + |I|^2\right)$, что связано с необходимостью сортировки признаков τ_a, а также требует выполнения некоторых других операций [5–12].

Этап выделения термов признаков предполагает анализ каждого ассоциативного правила в синтезированной базе правил для каждого признака. Поэтому вычислительная сложность данного этапа составит $O_{\mathrm{T}} = O_{\mathrm{T}}\left(N_{\mathrm{RB}}|I|\right)$. Поскольку $N_{\mathrm{RB}} << N_D$, оценка O_{T} может быть определена как $O_{\mathrm{T}} = O_{\mathrm{T}}\left(N_D|I|\right)$. При определении эквивалентности признаков для каждого *a*-го и *b*-го признаков вычисляется величина γ_{ab} как сумма эквивалентностей термов соответствующих признаков, на что потребуется $O_{\mathrm{E}} = O_{\mathrm{E}}(|I|^2)$ операций.

Формирование групп $\Psi = \{\Psi_1, \Psi_2, ..., \Psi_{N_\Psi}\}$ близких признаков предполагает поиск признаков $\tau_a \in I$ и $\tau_b \in I$ по максимуму оценки их эквивалентности γ_{ab}. Поэтому вычислительная сложность данного этапа также квадратично зависит от количества признаков $|I|$ в базе транзакций *D*: $O_{\mathrm{F}} = O_{\mathrm{F}}(|I|^2)$. При использовании эволюционного подхода на данном этапе требуется оценить N_H сгенерированных хромосом на каждой из N_{it} итераций. При этом используется целевая функция, зависящая от величины γ, рассчитанной на предыдущем этапе. Следовательно, вычислительная сложность этапа формирования групп $\Psi = \{\Psi_1, \Psi_2, ..., \Psi_{N_\Psi}\}$ при использовании эволюционного поиска составит $O_{\mathrm{F}} = O_{\mathrm{F}}\left(N_H N_{\mathrm{it}}\right)$. Учитывая, что $N_H \sim |I|$ и $N_{it} \sim |I|$, оценим величину O_{F} как $O_{\mathrm{F}} = O_{\mathrm{F}}(|I|^2)$.

Таким образом, оценка вычислительной сложности предложенного метода факторного анализа может быть определена следующим образом (13):

$$O = O_{\mathrm{AR}} + O_{\mathrm{T}} + O_{\mathrm{E}} + O_{\mathrm{F}} = O\left(|I| \cdot N_D \log_2(N_D) + |I|^2\right) + \\ + O\left(N_D |I|\right) + O\left(|I|^2\right) + O\left(|I|^2\right) = O\left(|I| \cdot N_D \log_2(N_D) + |I|^2\right). \quad (13)$$

Полученная оценка вычислительной сложности позволяет оценить разработанный метод как вычислительно эффективный, поскольку количество элементарных операций, необходимых для факторного анализа, полиномиально зависит от характеристик исходных данных, представленных в виде базы транзакций *D*, содержащей $|I|$ признаков и N_D транзакций.

6. ЭКСПЕРИМЕНТЫ И РЕЗУЛЬТАТЫ

С целью экспериментального исследования предложенного метода факторного анализа на основе ассоциативных правил решалась задача медицинского диагностирования нейро-артритических аномалий [11], а также ряд задач факторного анализа на специально сгенерированных тестовых данных.

Транзакционная база данных, представляющая собой информацию о состоянии здоровья детей, рожденных от родителей, пострадавших от аварии на Чернобыльской АЭС [11], содержала значения диагностических критериев формирования различных заболеваний, а также установленные диагнозы. С целью выявления взаимосвязей между различными заболеваниями и наборами тех или иных показателей выполнялся факторный анализ базы транзакций. База транзакций содержала $N_D = |D| = 344$ записей, каждая из которых представляла информацию о конкретном пациенте и могла характеризоваться несколькими из $N_I = |I| = 69$ признаков (численные данные лабораторных исследований, антропометрические показатели, другие характеристики пациента, а также набор установленных диагнозов). Каждая запись содержала в среднем 14 признаков.

Результаты проведения экспериментов позволили выявить группы взаимосвязанных характеристик пациента и различных заболеваний. Это позволит выполнять диагностирование некоторых болезней на ранних стадиях (при наличии некоторых факторов, входящих в одну группу с соответствующим диагнозом), 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.

В результате применения предложенного метода было выделено семь групп признаков, характеризующих здоровье пациента. Так, например, выявлено, что взаимосвязанными являются такие показатели и диагнозы: эмоциональная лабильность, диспептический синдром, уменьшение концентрации 4-пиридоксиновой кислоты, ацетонемическая рвота, уратурия в период новорожденности, нейро-артритические аномалии. Выявленные факторы и зависимости позволят своевременно выявлять сложно диагностируемые болезни типа «нейро-артритическая аномалия», а также предпринимать необходимые действия для предотвращения нежелательных переходов от латентной формы к открытой форме заболевания.

Важно отметить, что известные методы факторного анализа (например, метод главных компонент PCA (*Principal Component Analysis*) [6] и метод дискриминантного анализа Фишера FDA

(*Fisher Discriminant Analysis*) [7]) не могут быть применены для решения указанной задачи, поскольку входные данные представлены в виде транзакционной базы данных, содержащей транзакции с различным количеством признаков без указания значений выходных параметров. Результаты решения задачи медицинского диагностирования нейро-артритических аномалий позволяют сделать вывод о целесообразности применения разработанного метода для факторного анализа данных, представленных в виде баз транзакций.

Для исследования свойств и характеристик предложенного метода факторного анализа, а также для сравнения его с известными аналогами использовались специально сгенерированные тестовые данные. Разработанный метод сравнивался с методом главных компонент PCA [6] и методом дискриминантного анализа Фишера FDA [7]. Существующие методы поиска групп взаимосвязанных признаков позволяют выделять факторные группы на основе выборок, представляющие собой прямоугольные таблицы чисел, содержащие значения всех признаков для всех экземпляров. Поэтому тестовые выборки генерировались таким образом, чтобы они могли являться исходными данными как для известных методов факторного анализа, так и для разработанного. Также, важно отметить, что при создании тестовых выборок некоторые признаки зависели друг от друга, образовывая, таким образом, группы эквивалентности.

На рис.2 приведен график зависимости времени функционирования различных методов факторного анализа от количества признаков в исходной выборке. При этом количество экземпляров в выборке было постоянным и составляло $N_D = 1000$ шт. График, изображенный на рис.2, подтверждает квадратичную зависимость оценки вычислительной сложности O предложенного метода от количества признаков $|I|$ и характеризует его как вычислительно эффективный метод факторного анализа. Кроме того, из рис.2 видно, что предложенный метод при невысоких значениях количества признаков быстрее выполняет факторный анализ входных данных, при $|I| = 50$ времена работы методов несущественно отличаются.

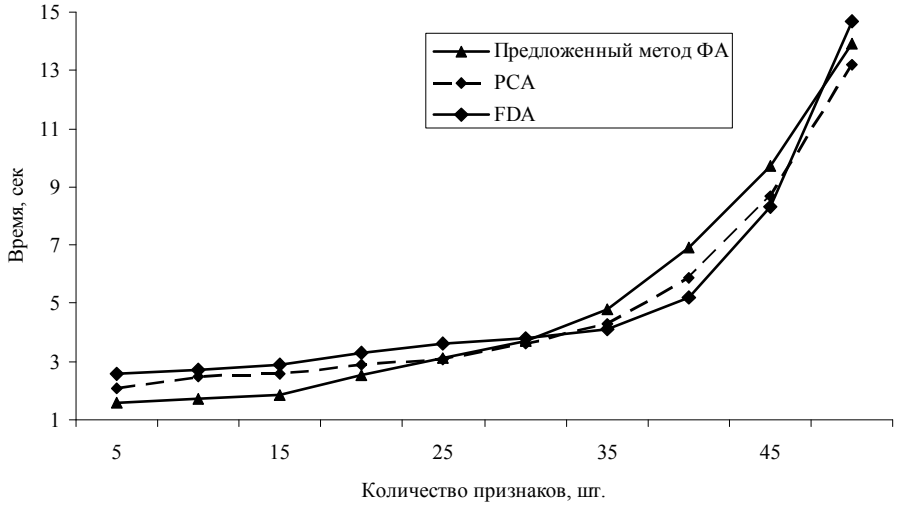

Рис.2. График зависимости времени функционирования метода от количества признаков в исходной выборке.

Также проводились эксперименты по исследованию влияния количества экземпляров в исходной выборке на время факторного анализа при использовании различных методов. Результаты экспериментов показали, что наиболее эффективным является метод PCA, который предназначен для решения задач факторного анализа данных, представленных в виде таблиц значений всех признаков для всех экземпляров, однако, в отличие от предложенного метода, не позволяет обрабатывать транзакционные базы данных. Время функционирования разработанного метода факторного анализа на основе ассоциативных правил не существенно превышает время факторного анализа с помощью метода PCA и подтверждает оценку вычислительной сложности как функцию от величины $N_D \log_2 \left(N_D \right)$.

Зависимость ошибки распознавания от количества признаков в исходной выборке приведена на рис.3 (количество экземпляров в выборке составляло $N_D = 1000$ шт.).

Как видно из рис.3, при незначительном количестве признаков (до 25) предложенный метод показывал несколько лучшие результаты по сравнению с существующими. При более высоких значениях известные методы показывали более приемлемые результаты. Это вызвано тем, что предложенный метод факторного анализа разрабатывался для обработки данных, представленных в виде баз транзакций. Но, как отмечено выше, для проведения экспериментов по сравнению предложенного метода с существующими аналогами использовались выборки данных в виде таблиц значений всех признаков для всех экземпляров. Тем не менее, результаты, полученные с помощью предложенного метода, не существенно отличаются от результатов известных методов, что позволяет сделать вывод о целесообразности применения разработанного метода на практике для факторного анализа транзакционных баз данных.

Таким образом, результаты экспериментов показали, что разработанный метод позволяет выполнять факторный анализ на основе баз транзакций. Сравнение предложенного метода с существующими аналогами подтвердило целесообразность его применения на практике.

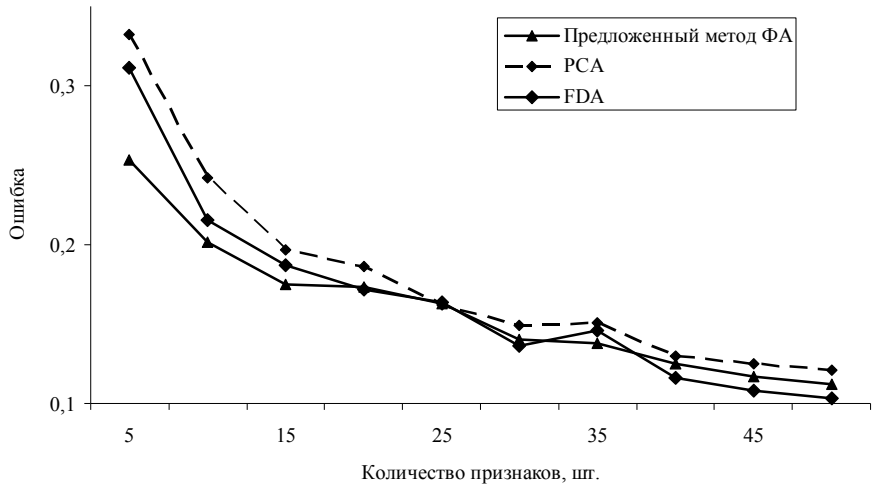

Рис.3. График зависимости ошибки распознавания от количества признаков в исходной выборке.

7. ЗАКЛЮЧЕНИЕ

В работе решена актуальная задача автоматизации факторного анализа в транзакционных базах данных. Научная новизна работы заключается в том, что предложен метод факторного анализа на основе ассоциативных правил, который предполагает извлечение правил из заданных баз транзакций, в результате чего выполняется обобщение данных, и, соответственно, исключение из дальнейшего рассмотрения избыточных признаков, что позволяет сократить пространство поиска и время выполнения факторного анализа. В разработанном методе определение эквивалентности признаков для формирования факторных групп выполняется исходя из частоты их совместного попадания в ассоциативные правила синтезированной базы правил, что позволяет оценивать тесноту связи между различными признаками (качественными, количественными), не выдвигать требований к входным данным и выполнять факторный анализ в транзакционных базах данных. Практическая ценность полученных результатов заключается в том, что на основе предложенного метода решена практическая задача выделения факторных групп для медицинского диагностирования нейро-артритических аномалий.

СПИСОК ЛИТЕРАТУРЫ

[1] Интеллектуальные информационные технологии проектирования автоматизированных систем диагностирования и распознавания образов: монография / С. А. Субботин, А. А. Олейник, Е.А. Гофман, С.А. Зайцев, А.А. Олейник; под ред. С.А. Субботина. – Харьков: ООО “Компания Смит”, 2012.

[2] *Mulaik S.A*. Foundations of Factor Analysis. Boca Raton, Florida: CRC Press, 2009.

[3] *Rummel R.J.* Applied Factor Analysis. Evanston: Northwestern University Press, 1988.

[4] *Иберла К*. Факторный анализ. М.: Статистика, 1980.

[5] *Zhang C., Zhang S*. Association rule mining: models and algorithms. Berlin: Springer, 2002.

[6] *Jolliffe I.T*. Principal Component Analysis. Berlin: Springer, 2002.

[7] *McLachlan G*. Discriminant Analysis and Statistical Pattern Recognition. New Jersey : John Wiley & Sons, 2004.

[8] *Gkoulalas-Divanis A*., *Verykios S*. Association Rule Hiding for Data Mining. New York: Springer-Verlag, 2010.

[9] Encyclopedia of artificial intelligence / Eds.: J.R. Dopico, J.D. de la Calle, A.P. Sierra. New York: Information Science Reference, 2009.

[10] *Зайко Т.А., Олейник А.А, Субботин С.А*. Определение индивидуальной значимости признаков для извлечения численных ассоциативных правил // Материалы III-го научно-исследовательского семинара Искусственный интеллект и его приложения, Магнитогорск, 2012. С. 105–108.

[11] *Зайко Т.А., Олейник А.А., Жихарева Н.В., Субботин С.А*. Диагностирование нейро-артритических аномалий на основе ассоциативных правил // Бионика интеллекта. 2012. № 2 (79). С. 53–57.

[12] *Зайко Т.А., Олейник А.А, Субботин С.А*. Вычисление пороговых значений поддержки при извлечении ассоциативных правил из больших массивов данных // Материалы Международной научно–практической конференции Информационные процессы и технологии. Информатика. Севастополь, 2013. С. 53–54.

[13] *Zhao Y., Zhang C., Cao L*. Post-mining of association rules: techniques for effective knowledge extraction. New York: Information Science Reference, 2009.

[14] *Dubois D.A., Hullermeier E., Prade H*. Systematic Approach to the Assessment of Fuzzy Association Rules // Data Mining and Knowledge Discovery. 2006. Vol. 13. P. 167–192.

[15] *Субботін С.О., Олійник А.О., Олійник О.О*. Неітеративні, еволюційні та мультиагентні методи синтезу нечіткологічних і нейромережних моделей: монографія. Запоріжжя: ЗНТУ, 2009.

[16] The Practical Handbook of Genetic Algorithms / ed. L. D. Chambers. Florida: CRC Press, 2000.

[17] *Gen M., Cheng R*. Genetic algorithms and engineering design. New Jersey: John Wiley & Sons, 1997.

[18] *Haupt R., Haupt S.* Practical Genetic Algorithms.New Jersey: John Wiley & Sons, 2004.

Рукопись получена 29.08.2013,
финальная версия – 12.12.2013.