УДК 004.93

Т. А. Зайко, А. А. Олейник, С. А. Субботин
Запорожский национальный технический университет
ул. Жуковского, 64, 69063 Запорожье, Украина

Факторный анализ на основе ассоциативных правил

Рассмотрена задача факторного анализа в транзакционных базах данных. Предложен метод факторного анализа на основе ассоциативных правил. Разработанный метод предполагает извлечение правил из заданных баз транзакций, что позволяет получить оценки эквивалентности термов признаков, исключить избыточные признаки, сократив тем самым пространство поиска и уменьшив время анализа, а также сформировать группы качественно близких признаков. Проведены эксперименты по решению тестовых задач факторного анализа.

Ключевые слова: *ассоциативное правило, база транзакций, признак, терм признака, факторный анализ.*

Введение

Решение задач диагностирования, автоматической классификации и прогнозирования связано с необходимостью обработки больших объемов информации [1, 2]. С целью сокращения времени такой обработки, а также синтеза моделей исследуемых объектов или процессов, обладающих высокими способностями к аппроксимации и обобщению, в некоторых случаях необходимо выполнять сокращение объема обрабатываемых данных [3, 4]. Для выявления взаимосвязей между различными признаками, описывающими исследуемые объекты или процессы, применяют методы факторного анализа [1, 4–9], позволяющие сократить число параметров, необходимых для описания данных [4].

Однако известные методы факторного анализа выдвигают ряд требований к обрабатываемым данным: однородность обучающей выборки, большое количество экземпляров выборки (не менее чем в два раза превышающее количество признаков) [4–7]. Кроме того, такие методы не предназначены для обработки бинарных или категориальных данных, а также данных, представленных в виде набора транзакций, описывающих каждый исследуемый объект или процесс как последовательность значений некоторых из его возможных характеристик [4–9].

Целью настоящего исследования является разработка метода факторного анализа на основе ассоциативных правил, позволяющего выявлять факторные группы признаков в транзакционных базах данных.

Постановка задачи факторного анализа данных, представленных в виде баз транзакций

Пусть задана база транзакций *D*:

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

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

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

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; $t_{ij} = \langle \tau_{ij}; v(\tau_{ij}) \rangle$ — *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*; 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 $\tau \in I$.

Метод факторного анализа на основе ассоциативных правил

Поскольку использование известных методов факторного анализа [4–7] для обработки данных, представленных в виде баз транзакций, не представляется возможным, в настоящей работе предлагается метод факторного анализа на основе ассоциативных правил, использование которого предполагает извлечение правил [10–12] из заданных транзакционных баз данных. Это позволяет получить оценки эквивалентности термов признаков, исключить избыточные признаки, сократив тем самым пространство поиска и уменьшив время анализа, а также сформировать группы качественно близких признаков.

Разработанный метод факторного анализа в транзакционных базах данных на основе ассоциативных правил состоит из следующих этапов:

— инициализации;
— извлечения ассоциативных правил и построения базы правил;
— выделения термов признаков;
— определения эквивалентности термов и признаков;
— поиска 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 [10–12] $D \to \text{БП}$, 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. Далее выполняется упрощение синтезированной базы ассоциативных правил БП [1, 4, 10–12], объединяя по возможности некоторые правила.

После этого на основе построенной базы правил БП выделяются термы признаков $\tau_a \in I$. Для этого анализируется каждое ассоциативное правило из базы БП ($\text{АП}_l \in \text{БП}$), 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*-м 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:

$$\alpha_M = \frac{\sum_{l=1}^{N_{\text{БП}}} \beta_{Ml}}{N_{\text{БП}}},$$

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 $\tau_b \in I$; $N_{\text{БП}}$ — количество правил в базе БП; β_{Ml} — величина, определяющая наличие связи между термами признаков картежа M в l-м правиле $\text{АП}_l \in \text{БП}$ синтезированной базы правил БП:

$$\beta_{Ml} = \begin{cases} 1, \text{ если в } l\text{-м правиле базы правил БП содержатся термы } \Delta\tau_{ak} \text{ и } \Delta\tau_{bm}; \\ 0, \text{ в противном случае.} \end{cases}$$

После определения эквивалентности термов α_M определяется эквивалентность признаков. Будем считать, что признаки тем эквивалентнее, чем они содержат больше эквивалентных термов. Для оценивания эквивалентности a-го и b-го признаков определяется величина γ_{ab}:

$$\gamma_{ab} = \frac{\sum_{m=1}^{N_{\Delta\tau_b}} \sum_{k=1}^{N_{\Delta\tau_a}} \alpha_M}{N_{\Delta\tau_b} N_{\Delta\tau_a}}, \quad M = \langle a, b, k, m \rangle.$$

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:

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

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-м правиле $\text{АП}_l \in \text{БП}$ базы правил БП 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:

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

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 приведенное выше 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.

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 $O_1 + O_2 + ... + O_{N_{эт.}}$, характеризующих вычислительную сложность каждого из $N_{эт.}$ соответствующего метода. Вычислительная сложность этапа извлечения ассоциативных правил $O_{АП}$ определяется как $O_{АП} = O_{АП}\left(|I| \cdot N_D \log_2(N_D) + |I|^2\right)$, что связано с необходимостью сортировки признаков τ_a, а также требует выполнения некоторых других операций [10–12].

Этап выделения термов признаков предполагает анализ каждого ассоциативного правила в синтезированной базе правил для каждого признака. Поэтому вы-

числительная сложность данного этапа составит $O_{\text{ТП}} = O_{\text{ТП}}\left(N_{\text{БП}}|I|\right)$. Поскольку $N_{\text{БП}} << N_D$, оценка $O_{\text{ТП}}$ может быть определена как $O_{\text{ТП}} = O_{\text{ТП}}\left(N_D|I|\right)$.

При определении эквивалентности признаков для каждого *a*-го и *b*-го признаков τ_a и τ_b вычисляется величина γ_{ab} как сумма эквивалентностей каждого из *k*-го и *m*-го термов соответствующих признаков, на что потребуется $O_{\text{ЭП}} = O_{\text{ЭП}}\left(|I|^2\right)$ операций.

Формирование групп $\Psi = \left\{\Psi_1, \Psi_2, ..., \Psi_{N_\Psi}\right\}$ 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*: $O_{\text{ФГ}} = O_{\text{ФГ}}\left(|I|^2\right)$. При использовании эволюционного подхода на данном этапе требуется оценить N_H сгенерированных хромосом на каждой из $N_{\text{итер.}}$ итераций. При этом используется целевая функция, зависящая от величины γ, рассчитанной на предыдущем этапе. Следовательно, вычислительная сложность этапа формирования групп $\Psi = \left\{\Psi_1, \Psi_2, ..., \Psi_{N_\Psi}\right\}$ при использовании эволюционного поиска составит $O_{\text{ФГ}} = O_{\text{ФГ}}\left(N_H N_{\text{итер.}}\right)$. Учитывая, что $N_H \sim |I|$ и $N_{\text{итер.}} \sim |I|$, оценим величину $O_{\text{ФГ}}$ как $O_{\text{ФГ}} = O_{\text{ФГ}}\left(|I|^2\right)$.

Таким образом, оценка вычислительной сложности предложенного метода факторного анализа на основе ассоциативных правил может быть определена следующим образом:

$$O = O\left(|I| \cdot N_D \log_2\left(N_D\right) + |I|^2\right) + O\left(N_D|I|\right) + O\left(|I|^2\right) + O\left(|I|^2\right) = O\left(|I| \cdot N_D \log_2\left(N_D\right) + |I|^2\right).$$

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.

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.

Эксперименты и результаты

С целью экспериментального исследования предложенного метода факторного анализа на основе ассоциативных правил решался ряд задач факторного анализа на специально сгенерированных тестовых данных. Для исследования свойств и характеристик предложенного метода факторного анализа, а также сравнения его с известными аналогами использовались специально сгенерированные тестовые данные. Разработанный метод сравнивался с методом главных компонент PCA (Principal Component Analysis) [8] и методом дискриминантного анализа Фишера FDA (Fisher Discriminant Analysis) [9]. 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. Кроме того, важно отметить, что при создании тестовых выборок некоторые признаки зависели друг от друга, образовывая, таким образом, группы эквивалентности.

Влияние количества экземпляров в исходной выборке на время факторного анализа при использовании различных методов отображено в виде графиков, представленных на рис. 1. При проведении экспериментов для построения указанных графиков количество признаков в выборке было постоянным и составляло $|I| = 25$ шт.

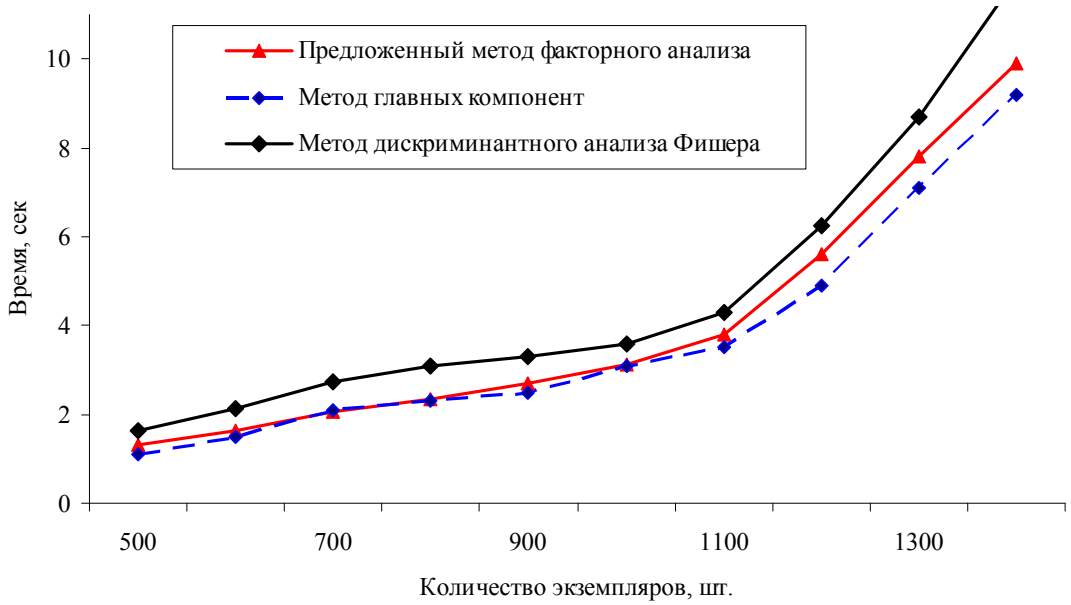

Рис. 1. График зависимости времени функционирования метода от количества экземпляров в исходной выборке

Из рис. 1 видно, 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 несущественно превышает время факторного анализа с помощью метода PCA и подтверждает оценку вычислительной сложности как функцию от величины $N_D \log_2(N_D)$.

На рис. 2 отображены результаты экспериментов по исследованию зависимости количества сгенерированных факторных групп от количества признаков в исходной выборке (количество экземпляров при проведении экспериментов составляло $N_D = 5000$ шт.).

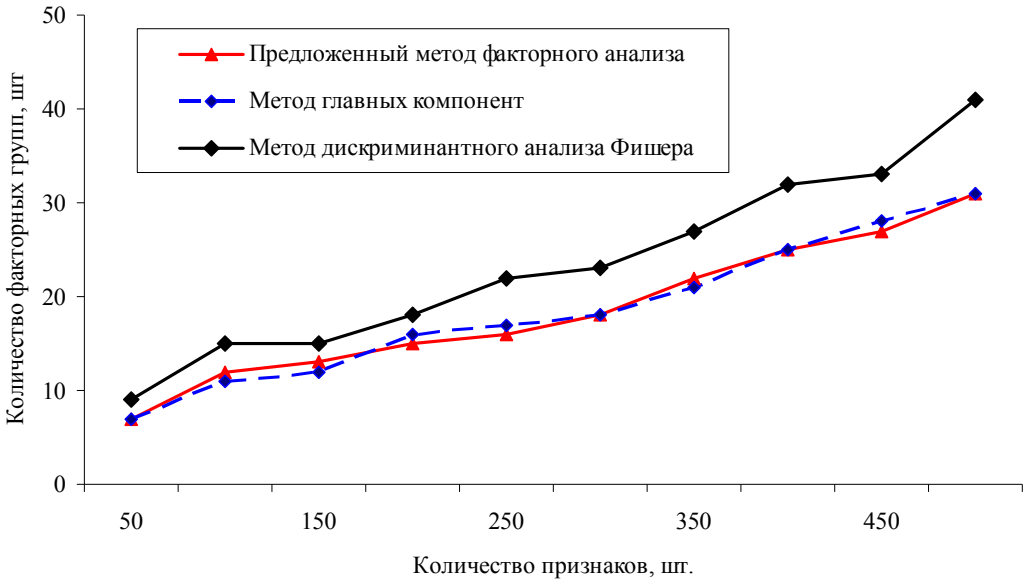

Рис. 2. График зависимости количества сгенерированных факторных групп от количества признаков в исходной выборке

Кривые, изображенные на рис. 2 и построенные по результатам применения различных методов факторного анализа, показывают, что количество синтезированных факторных групп пропорционально количеству признаков в обучающей выборке, что обусловлено внутренними взаимосвязями между признаками и структурой выборки. Как видно, метод FDA синтезировал большее количество факторных групп, выделяя некоторые взаимозависимые наборы признаков в разные группы. Метод PCA и предложенный метод показали схожие результаты.

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.

Выводы

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.

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).

1. *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.

2. *Прогрессивные* технологии моделирования, оптимизации и интеллектуальной автоматизации этапов жизненного цикла авиадвигателей: монография / [А.В. Богуслаев, Ал.А. Олейник, Ан.А. Олейник и др.]; под ред. Д.В. Павленко, С.А. Субботина. — Запорожье: ОАО «Мотор Сич», 2009. — 468 с.

3. *Барсегян А.А.* Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP: уч. пособ. / А.А. Барсегян. — СПб: BHV, 2007. — 384 с.

4. *Рассел С.* Искусственный интеллект: современный подход / С. Рассел, П. Норвиг. — М.: Вильямс, 2006. — 1408 с.

5. *Иберла К.* Факторный анализ / К. Иберла. — М.: Статистика. — 1980. — 398 с.

6. *Mulaik S.A.* Foundations of Factor Analysis / S.A. Mulaik. — Boca Raton, Florida: CRC Press. — 2009. — 548 p.

7. *Rummel R.J.* Applied Factor Analysis / R.J. Rummel. — Evanston: Northwestern University Press. — 1988. — 617 p.

8. *Jolliffe I.T.* Principal Component Analysis / I.T. Jolliffe. — Berlin: Springer-Verlag. — 2002. — 489 p.

9. *McLachlan G.* Discriminant Analysis and Statistical Pattern Recognition / G. McLachlan. — New Jersey: John Wiley & Sons. — 2004. — 526 p.

10. *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.

11. *Adamo J.-M.* Data Mining for Association Rules and Sequential Patterns: Sequential and Parallel Algorithms / Adamo J.-M. — New York: Springer-Verlag. — 2001. — 259 p.

12. *Koh Y.S.* Rare Association Rule Mining and Knowledge Discovery / Y.S. Koh, N. Rountree. — New York: Information Science Reference. — 2009. — 320 p.

Поступила в редакцию 04.11.2013