УДК 004.93

Т.А. ЗАЙКО, асп., ЗНТУ, Запорожье,
А.А. ОЛЕЙНИК, канд. техн. наук, доц., ЗНТУ, Запорожье,
С.А. СУББОТИН, д-р техн. наук, проф., ЗНТУ, Запорожье

УПРОЩЕНИЕ ТРАНЗАКЦИОННЫХ БАЗ ДАННЫХ НА ОСНОВЕ ЧЕТКИХ ПРОДУКЦИЙ

Рассмотрена задача упрощения транзакционных баз данных. Предложен метод сокращения баз транзакций на основе четких продукций. Разработанный метод позволяет исключить неинформативные признаки и избыточные экземпляры из заданных массивов данных, что, в свою очередь, позволяет понизить структурную и параметрическую сложность синтезируемых диагностических моделей. Библиогр. 14 назв.

Ключевые слова: транзакционные базы данных, четкие продукции, модель, признак, транзакция, экземпляр.

Постановка проблемы и анализ литературы. Разработка интеллектуальных систем неразрушающего контроля качества, технического и медицинского диагностирования, распознавания образов связана с необходимостью обработки больших объемов информации [1, 2]. Зачастую такая информация может представляться в виде баз транзакций, где каждая транзакция представляет собой список значений некоторых из возможных признаков, характеризующих исследуемые объекты или процессы [3, 4].

Использование избыточных данных при синтезе диагностических моделей может привести к построению моделей, обладающих низкими обобщающими способностями, а также высокой структурной и параметрической сложностью, что повлечет увеличение затрат памяти ЭВМ на хранение моделей и увеличение времени вычислений на обработку большого объема данных. Следовательно, перед осуществлением синтеза диагностических моделей целесообразным является сокращение обучающей выборки путем исключения из нее избыточной информации.

Известные методы редукции данных [2 – 6], как правило, предназначены либо для отбора признаков, либо для отбора экземпляров и часто не учитывают взаимосвязи сочетаний некоторых значений признаков, которые также могут быть исключены из исходной выборки. Поэтому актуальной является разработка нового метода сокращения обучающей выборки, позволяющего выполнять редукцию признаков, экземпляров, термов признаков и формировать множество данных с меньшим количеством элементов по сравнению с исходной выборкой.

Для редукции обучающей выборки в настоящей работе предлагается использовать четкие продукции, извлекаемые с помощью методов ассоциативных правил [3, 4, 7 – 11], поскольку извлечение таких правил из выборок данных позволяет существенно сокращать объемы информации и выполнять обобщение данных, преобразовывать значения признаков в некоторые диапазоны значений, оценивать степень влияния признаков на выходной параметр, а также уровень их взаимосвязи между собой, в том числе взаимосвязи некоторых значений признаков.

Цель статьи – разработка метода упрощения транзакционных баз данных на основе четких продукций.

Метод упрощения транзакционных баз данных на основе четких продукций. 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 объектах или процессах, где $N_D=|D|$ – число экземпляров (элементов) в наборе данных *D*. Элементы T_j представляют собой множество значений вида: $T_j=\{\tau_{1j},\tau_{2j},...,\tau_{N_I j},y_j\}$, где $\tau_{aj}=[\tau_{aj\min};\tau_{aj\max}]$ – значение *a*-го признака τ_a для элемента T_j; τ_a – *a*-й признак множества $I=\{\tau_1,\tau_2,...,\tau_{N_I}\}$, $a=1,2,...,N_I$; I – множество признаков, которыми описываются элементы T_j, набора данных *D*; $N_I=|I|$ – число признаков в выборке *D*; $\tau_{aj\min}$ и $\tau_{aj\max}$ – минимальное и максимальное значения из диапазона возможных значений признака τ_a; y_j – значение выходного параметра для элемента T_j. Тогда задача сокращения размерности обучающей выборки заключается в уменьшении числа её экземпляров $N'_D<N_D$ и описывающих их признаков $N'_I<N_I$, с сохранением возможности построения диагностических моделей с приемлемыми способностями к аппроксимации исследуемых зависимостей.

В разработанном методе сокращения размерности обучающей выборки для редукции данных предлагается извлекать ассоциативные правила. Информация об интересности выявленных правил используется для оценивания степени влияния признаков на выходной параметр, а также взаимосвязи некоторых значений признаков между собой.

На начальном этапе для заданной выборки *D* выполняется редукция её экземпляров. Для этого дискретизируются значения признаков (диапазон значений $\Delta_a=[\tau_{a\min};\ \tau_{a\max}]$ каждого признака τ_a

разбивается на $N_{\text{int}.a}$ интервалов). После дискретизации выполняется преобразование $D \to D_1'$, в результате которого значения исходных признаков τ_a заменяются номерами интервалов значений признаков, выделенных в процессе дискретизации: $\tau'_{aj} = n(\tau_{aj})$, где τ_{aj} и τ'_{aj} – значения a-го признака для j-го экземпляра в выборках D и D_1', соответственно; $n(\tau_{aj})$ – номер интервала значений признака τ_a, в который попадает его значение τ_{aj} для j-го экземпляра.

Полученные в результате преобразования $D \to D_1'$ экземпляры T_j' и T_k' с одинаковыми значениями признаков τ'_{aj} и τ'_{ak}, $a = 1,2,\ldots,N_I$ считаются эквивалентными и избыточными. Поэтому в выборке D_1' последовательно для каждых двух эквивалентных экземпляров T_j' и T_k' следует оставить один экземпляр T_j', а другой – исключить: $D_1' = D_1' \setminus T_k'$.

После выполнения этапа редукции экземпляров происходит выявление неинформативных признаков с последующим их исключением из выборки. Для редукции признаков τ_a из выборки D_1' будем извлекать ассоциативные правила $AR_l \in RB$ (RB – база извлеченных ассоциативных правил), оценивать их интересность и интересность каждого терма признаков, на основе чего будем делать вывод об информативности каждого признака. Для этого вначале извлекаются численные ассоциативные правила $AR_l : X_l \to Y_l$ [3, 4, 7 – 11], затем выполняется оценивание интересности I_{AR_l} каждого из выявленных правил. В качестве оценок интересности правил возможно использовать критерии (1) – (5) [3, 4, 7–11]:

$$I_{\text{AR}_l} = \text{supp}(X_l \to Y_l) + \text{supp}(\overline{X_l} \to \overline{Y_l}), \quad (1)$$

$$I_{\text{AR}_l} = \frac{\text{supp}(X_l \to Y_l)}{\text{supp}(X_l)\text{supp}(Y_l)}, \quad (2)$$

$$I_{\text{AR}_l} = \frac{\text{conf}(X_l \to Y_l)}{\text{conf}(\overline{X_l} \to Y_l)}, \quad (3)$$

$$I_{\text{AR}_l} = \frac{\text{supp}(X_l \to Y_l)\text{supp}(\overline{X_l} \to \overline{Y_l})}{\text{supp}(X_l \to \overline{Y_l})\text{supp}(\overline{X_l} \to Y_l)}, \quad (4)$$

$$I_{\text{AR}_l} = \text{supp}(X_l \to Y_l) - \text{supp}(X_l)\text{supp}(Y_l), \quad (5)$$

где supp(A) – поддержка множества A, определяемая как отношение числа элементов T_j, содержащих A, к общему числу экземпляров N_D в наборе данных D; $\mathrm{conf}(A)$ – достоверность множества A, рассчитываемая как отношение поддержки импликации A к поддержке ее левой части.

Используя информацию об интересности I_{AR_l} извлеченных ассоциативных правил, выполняется оценивание интересности термов $\Delta\tau_{ak}$, $k=1,2,\ldots,N_{\mathrm{int}.a}$ каждого признака τ_a, $a=1,2,\ldots,N_I$. Интересность термов $\Delta\tau_{ak}$ предлагается определять по одной из следующих формул (6) – (8):

$$I_{\Delta\tau_{ak}} = \frac{1}{N_{\Delta\tau_{ak}}} \sum_{\substack{l:\mathrm{AR}_l\in\mathrm{RB},\\ \Delta\tau_{ak}\in\mathrm{AR}_l}} I_{\mathrm{AR}_l}, \tag{6}$$

$$I_{\Delta\tau_{ak}} = \min_{\substack{l:\mathrm{AR}_l\in\mathrm{RB},\\ \Delta\tau_{ak}\in\mathrm{AR}_l}} \{I_{\mathrm{AR}_l}\}, \tag{7}$$

$$I_{\Delta\tau_{ak}} = \max_{\substack{l:\mathrm{AR}_l\in\mathrm{RB},\\ \Delta\tau_{ak}\in\mathrm{AR}_l}} \{I_{\mathrm{AR}_l}\}, \tag{8}$$

где $N_{\Delta\tau_{ak}}$ – число ассоциативных правил $AR_l \in RB$, содержащих терм $\Delta\tau_{ak}$: $\Delta\tau_{ak} \in AR_l$. Информативность I_a признаков τ_a будем оценивать исходя из оценок интересностей термов, входящих в соответствующий признак (9) – (11):

$$I_a = \frac{1}{N_{\mathrm{int}.a}} \sum_{k=1}^{N_{\mathrm{int}.a}} I_{\Delta\tau_{ak}}, \tag{9}$$

$$I_a = \max_{k=1,2,\ldots,N_{\mathrm{int}.a}} \{I_{\Delta\tau_{ak}}\}, \tag{10}$$

$$I_a = \min_{k=1,2,\ldots,N_{\mathrm{int}.a}} \{I_{\Delta\tau_{ak}}\}. \tag{11}$$

Признаки τ_a с низкими значениями информативности I_a исключаются из выборки D_1'.

С целью выполнения этапа сокращения избыточных термов из выборки D_2' извлекаются ассоциативные правила и выявляются взаимосвязи между различными интервалами $\Delta\tau_{ak}$ и $\Delta\tau_{bm}$ признаков.

В результате извлечения ассоциативных правил из выборки D_2' синтезируется база правил RB_2 вида $AR_l : X_l \to Y_l$ с уровнем

достоверности $\mathrm{conf}(X_l \to Y_l)$, не ниже минимально приемлемого $\mathrm{minconfidence}$. Поэтому из транзакций (экземпляров) T'_{2j} выборки D'_2 можно исключить термы $\Delta\tau_{ak} \in X_l$ при наличии в этих же транзакциях термов $\Delta\tau_{bm} \in Y_l$, входящих в консеквенты Y_l правил базы RB_2 (12):

$$T'_{3j} = T'_{2j} \setminus \bigcup_{\substack{\Delta\tau_{ak} \in X_l, \\ \exists(\Delta\tau_{bm} \subset T'_{2j}) \in Y_l, \\ (X_l \to Y_l) \in \mathrm{RB}_2}} (\tau_a \in \Delta\tau_{ak}). \qquad (12)$$

Путем исключения избыточных термов из выборки D'_2 выполняется преобразование $D'_2 \to D'_3$ и формирование выборки D'_3 сокращенной размерности. Таким образом полученное разбиение пространства признаков D'_3 содержит существенно меньшее число элементов $\Delta\tau_{ak}$ по сравнению с исходной выборкой D, характеризуется более высокими обобщающими свойствами и позволяет понизить структурную и параметрическую сложность синтезируемых диагностических моделей.

Для выполнения экспериментального исследования предложенного метода сокращения размерности обучающей выборки на основе ассоциативных правил он был программно реализован на языке C#. Выборка для проведения экспериментов содержала информацию о характеристиках сырья и параметрах технологического процесса изготовления кондитерской продукции для 3284 партий изделий (наблюдений), описывающихся с помощью 43 признаков. Далее эта выборка сокращалась путем применения предложенного метода, а также различных методов сокращения обучающих множеств (методы отбора признаков и методы отбора экземпляров [1, 2, 5, 6, 12 – 14]).

Результаты экспериментов показали, что предложенный метод упрощения баз транзакций на основе ассоциативных правил позволяет формировать множество данных с меньшим количеством элементов по сравнению с исходной выборкой, а также строить на его основе диагностические модели с высокими значениями показателей обобщения и интерпретабельности.

Выводы. В работе решена актуальная задача упрощения баз транзакций для построения диагностических моделей.

Научная новизна работы заключается в том, что предложен метод упрощения транзакционных баз данных на основе четких продукций, который предполагает выполнение этапов редукции экземпляров, признаков и избыточных термов, для оценивания информативности

признаков использует информацию об извлеченных ассоциативных правилах и позволяет формировать разбиение пространства признаков с меньшим количеством экземпляров по сравнению с исходной выборкой, что, в свою очередь, позволяет синтезировать более простые и удобные для восприятия диагностические модели.

Список літератури: **1**. *Denton T*. Advanced automotive fault diagnosis / *T. Denton*. – London: Elsevier, 2006. – 271 p. **2**. *Sobhani-Tehrani E*. Fault diagnosis of nonlinear systems using a hybrid approach / *E. Sobhani-Tehrani, K. Khorasani*. – New York: Springer, 2009. – 265 p. – (Lecture notes in control and information sciences; № 383). **3**. *Koh Y. S*. Rare Association Rule Mining and Knowledge Discovery / *Y.S. Koh, N. Rountree*. – New York: Information Science Reference. – 2009. – 320 p. **4**. *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. **5**. *Lee J.A.* Nonlinear dimensionality reduction / *J.A. Lee, M. Verleysen*. – New York: Springer, 2007. – 308 p. **6**. *Abonyi J*. Cluster analysis for data mining and system identification / *J. Abonyi, B. Feil*. – Basel: Birkhäuser, 2007. – 303 p. **7**. *Ayubi S*. An algorithm to mine general association rules from tabular data / *S. Ayubi, M.K. Muyeba, A. Baraani-dastjerdi, J.A. Keane* // Information Sciences. – 2009. – Vol. 179. – № 20. – P. 3520-3539. **8**. *Verlinde H*. Fuzzy versus quantitative association rules: a fair data-driven comparison / *H. Verlinde, M.D. Cock, R. Boute* // IEEE Transactions on Systems, Man and Cybernetics. – 2006. – Vol. 36. – № 3. – P. 679-684. **9**. *Sohn S. Y*. Searching customer patterns of mobile service using clustering and quantitative association rule / *S.Y. Sohn, Y. Kim* // Expert Systems With Applications. – 2008. – Vol. 34. – № 2. – P. 1070-1077. **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**. *Zhang C*. Association rule mining: models and algorithms / *C. Zhang, S. Zhang*. – Berlin: Springer-Verlag. – 2002. – 238 p. **12**. *Guyon I.* An introduction to variable and feature selection / *I. Guyon, A. Elisseeff* // Journal of machine learning research. – 2003. – № 3. – P. 1157–1182. **13**. *Jensen R*. Combining rough and fuzzy sets for feature selection: thesis ... doctor of philosophy / *Jensen Richard*. – Edinburgh: University of Edinburgh, 2005. – 221 p. **14**. *McLachlan G*. Discriminant Analysis and Statistical Pattern Recognition / *G. McLachlan*. – New Jersey: John Wiley & Sons. – 2004. – 526 p.

Bibliography (transliterated): **1**. *Denton T*. Advanced automotive fault diagnosis / *T. Denton*. – London: Elsevier, 2006. – 271 p. **2**. *Sobhani-Tehrani E*. Fault diagnosis of nonlinear systems using a hybrid approach / *E. Sobhani-Tehrani, K. Khorasani*. – New York: Springer, 2009. – 265 p. – (Lecture notes in control and information sciences; № 383). **3**. *Koh Y.S*. Rare Association Rule Mining and Knowledge Discovery / *Y.S. Koh, N. Rountree*. – New York: Information Science Reference. – 2009. – 320 p. **4**. *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. **5**. *Lee J.A.* Nonlinear dimensionality reduction / *J.A. Lee, M. Verleysen*. – New York: Springer, 2007. – 308 p. **6**. *Abonyi J*. Cluster analysis for data mining and system identification / *J. Abonyi, B. Feil*. – Basel: Birkhäuser, 2007. – 303 p. **7**. *Ayub, S*. An algorithm to mine general association rules from tabular data / *S. Ayubi, M. K. Muyeba, A. Baraani-dastjerdi, J. A. Keane* // Information Sciences. – 2009. – Vol. 179. – № 20. – P. 3520-3539. **8**. *Verlinde H*. Fuzzy versus quantitative association rules: a fair data-driven comparison / *H. Verlinde, M. D. Cock, R. Boute* // IEEE Transactions on Systems, Man and Cybernetics. – 2006. – Vol. 36. – № 3. – P. 679-684. **9**. *Sohn S.Y.* Searching customer patterns of mobile service using clustering and quantitative association rule / *S.Y. Sohn, Y. Kim* // Expert Systems With Applications. – 2008. – Vol. 34. – № 2. – P. 1070-1077. **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**. *Zhang C*. Association rule mining: models and algorithms / *C. Zhang, S. Zhang*. – Berlin: Springer-Verlag. – 2002. – 238 p. **12**. *Guyon, I.* An introduction to variable and feature selection / *I. Guyon, A. Elisseeff* // Journal of machine learning research. – 2003. – № 3. – P. 1157–1182. **13**. *Jensen R*. Combining rough and fuzzy sets for feature selection: thesis ... doctor of philosophy / *Jensen Richard*. – Edinburgh: University of Edinburgh, 2005. – 221 p. **14**. *McLachlan G*. Discriminant Analysis and Statistical Pattern Recognition / *G. McLachlan*. – New Jersey: John Wiley & Sons. – 2004. – 526 p.

Поступила (received) 25.04.2014

Статью представил д-р техн. наук, проф., декан Запорожского национального университета Гоменюк С.И.

Zayko Tatiana, postgraduate student
Zaporizhzhya National Technical University
Zhukovsky str., 64, Ukraine, Zaporizhzhya, 69063
tel: +380-97-355-61-55; e-mail: tzyakun@mail.ru

Oliinyk Andrii, Ph.D., Associate Professor
Zaporizhzhya National Technical University
Zhukovsky str., 64, Ukraine, Zaporizhzhya, 69063
tel: +380-98-256-38-93; e-mail: olejnikaa@gmail.com

Subbotin Sergey, Dr. Sci. Tech., Professor
Zaporizhzhya National Technical University
Zhukovsky str., 64, Ukraine, Zaporizhzhya, 69063
tel: +380-612-95-27-66; subbotin.csit@gmail.com
ORCID ID 0000-0001-5814-8268.