УДК 004.93

Т. А. Зайко[1], А. А. Олейник[2], С. А. Субботин[3]

[1] Запорожский национальный технический университет, г. Запорожье, Украина, tzyakun@mail.ru;
[2] Запорожский национальный технический университет, г. Запорожье, Украина, olejnikaa@gmail.com;
[3] Запорожский национальный технический университет, г. Запорожье, Украина, subbotin@zntu.edu.ua

МЕТОД ОПРЕДЕЛЕНИЯ ИНДИВИДУАЛЬНОЙ ЗНАЧИМОСТИ ПРИЗНАКОВ ДЛЯ ИЗВЛЕЧЕНИЯ ЧИСЛЕННЫХ АССОЦИАТИВНЫХ ПРАВИЛ

Статья посвящена решению проблемы автоматизации поиска информативных признаков в базах транзакций. Предлагается оценивать индивидуальную значимость признаков исходя из их значимости для описания границ областей компактного расположения транзакций в пространстве признаков. Выполняется экспериментальное исследование свойств и характеристик предложенного метода отбора информативных признаков.

АССОЦИАТИВНОЕ ПРАВИЛО, ИНФОРМАТИВНОСТЬ, КЛАСТЕРНЫЙ АНАЛИЗ, ПРИЗНАК, ТРАНЗАКЦИОННАЯ БАЗА ДАННЫХ, ФАЗЗИФИКАЦИЯ

Введение

Для обеспечения надежности, долговечности и безопасности работы сложных технических объектов и систем целесообразной является разработка методов и средств своевременного выявления сбоев в их работе, для чего, как правило, возникает необходимость извлечения новых знаний об исследуемых объектах и процессах [1], которые целесообразно представлять в виде ассоциативных правил [2], представляющих собой импликации вида $X \to Y$.

Однако признаки, описывающие исследуемые объекты или процессы, как правило, имеют различную информативность [3], поэтому с целью извлечения интересных ассоциативных правил, адекватно описывающих исследуемые зависимости, целесообразно учитывать индивидуальную значимость признаков. Поскольку выходной параметр в транзакционных базах данных, используемых для извлечения ассоциативных правил, является не заданным, применение известных методов отбора информативных признаков [3, 4] является затруднительным и нецелесообразным. Это обусловливает необходимость разработки нового метода отбора признаков на основе данных, представленных в виде баз транзакций.

Целью настоящей работы является создание метода определения индивидуальной значимости признаков для извлечения численных ассоциативных правил на основе транзакционных баз данных.

1. Постановка задачи отбора признаков из транзакционных баз данных

Пусть задана база транзакций D: $D = \left\{ T_1, T_2, ..., T_{N_D} \right\}$, в которой каждый элемент T_j, $j = 1,2,..., N_D$, содержит информацию о некоторых взаимосвязанных событиях, где $N_D = |D|$ – количество элементов (транзакций) в наборе данных D; $T_j = \left(tid_j, item_j \right)$; tid_j – идентификатор j-й транзакции T_j; $item_j = \left\{ t_{1j}, t_{2j}, ..., t_{N_{item_j} j} \right\} \subseteq I$ – список элементов, входящих в транзакцию T_j; t_{ij} – i-й элемент списка $item_j$, $i = 1,2,..., N_{item_j}$; $N_{item_j} = \left| item_j \right|$ – количество элементов множества $item_j$; $I = \left\{ \tau_1, \tau_2, ..., \tau_{N_I} \right\}$ – множество возможных переменных (признаков), которые могут входить в список элементов $item_j$ каждой транзакции T_j, $j = 1,2,..., N_D$, набора данных D; τ_a – a-й элемент множества I, $a = 1,2,..., N_I$; $N_I = |I|$ – количество элементов множества I.

Тогда задача отбора информативных признаков (исключения малозначимых признаков) [3–5] из заданной базы транзакций D может быть поставлена следующим образом: выделить комбинацию признаков $I^* \subset I$ из исходного набора данных D, в которой каждый a-й признак характеризуется оценкой индивидуальной значимости w_a не ниже заданного порогового значения $w_{h.}$:

$$w_a \geq w_{h.}.$$

2. Метод определения индивидуальной значимости признаков

Поскольку выходной параметр в транзакционных базах данных, как правило, не задан, в разработанном методе предлагается оценивать индивидуальную значимость признаков исходя из их значимости (важности) для описания границ областей компактного расположения транзакций в пространстве признаков. Следовательно, для определения индивидуальной значимости признаков предлагается выполнять кластерный анализ, в результате которого выделять группы (кластеры) компактно расположенных транзакций в пространстве признаков $\tau_a \in I$. При этом признаки

предварительно нормируются с целью приведения значений всех признаков к одному диапазону, что устранит влияние величины граничных значений признака на его индивидуальную значимость. После нормирования признаков выполняется разбиение диапазона их значений на некоторые интервалы для последующего отбора признаков и поиска ассоциативных правил, для чего целесообразно выполнять фаззификацию исходной базы транзакций D, выделяя таким образом нечеткие термы входных признаков. Преобразование вида $D \to FuzzyD$ целесообразно выполнять следующим образом. Вначале определяются функции принадлежности μ_a для каждого численного a-го признака $\tau_a \in I$.

Функции принадлежности могут быть заданы экспертом, исходя из его знаний и опыта относительно исследуемого объекта или процесса [6–8]. Однако использование субъективной информации и некоторых допущений при преобразовании ее в степени принадлежности нечетких множеств в некоторых случаях может привести к неприемлемым результатам такого преобразования, вследствие чего синтезированная база ассоциативных правил не будет содержать интересные правила, а новые знания, выделенные на основе построенной таким образом базы ассоциативных правил, будут необъективно отражать исследуемые объекты или процессы.

Поэтому в случае неуверенности в корректности выбора функций принадлежности экспертом целесообразно использовать оптимизационные методы построения функций принадлежности [6–8], основанные на параметрической идентификации параметров нечетких моделей по имеющейся входной информации, содержащей как входные значения признаков, так и выходные. При поиске ассоциативных правил входная информация, представленная в виде транзакционной базы данных D, как правило, не содержит значения выходных переменных. Поэтому применять такие методы затруднительно.

Следовательно, для фаззификации базы транзакций с целью последующего извлечения численных ассоциативных правил целесообразно использовать известные функции принадлежности [6–8], параметры которых предлагается выбирать исходя из идеи поддержки нечеткости знаний, т.е. таким образом, чтобы обеспечивалось пересечение соседних интервалов разбиений признаков (так, чтобы в результате фаззификации численное значение признака могло быть отнесено к нескольким термам).

В качестве функций принадлежности целесообразно использовать такие функции, которые позволяют ограничивать интервал значений признаков:

– трапециевидную функцию:

$$\mu_{ak} = \begin{cases} 0, \ \tau_a < p_{0ak}; \\ (\tau_a - p_{0ak})/(p_{1ak} - p_{0ak}), \ p_{0ak} \le \tau_a < p_{1ak}; \\ 1, \ p_{1ak} \le \tau_a \le p_{2ak}; \\ (p_{3ak} - \tau_a)/(p_{3ak} - p_{2ak}), \ p_{2ak} < \tau_a \le p_{3ak}; \\ 0, \ \tau_a > p_{3ak}, \end{cases}$$

где значения μ_{ak} – функция принадлежности a-го признака $\tau_a \in I$ k-му терму; p_{0ak}, p_{1ak}, p_{2ak}, p_{3ak} – параметры функции принадлежности, которые определяются таким образом, чтобы обеспечивалось некоторое пересечение интервалов, и признак имел возможность быть отнесенным к нескольким термам в силу нечеткости знаний;

– П-образную функцию:

$$\mu_{ak} = \mu S_{ak} \cdot \mu Z_{ak},$$

где μS_{ak} и μZ_{ak} – S-образная и Z-образная линейные функции принадлежности, соответственно:

$$\mu S_{ak} = \begin{cases} 0, \ \tau_a < p_{1ak}; \\ \dfrac{\tau_a - p_{1ak}}{p_{2ak} - p_{1ak}}, \ p_{1ak} \le \tau_a \le p_{2ak}; \\ 1, \ \tau_a > p_{2ak}, \end{cases}$$

$$\mu Z_{ik} = \begin{cases} 1, \ \tau_a < p_{1ak}; \\ \dfrac{p_{2ak} - \tau_a}{p_{2ak} - p_{1ak}}, \ p_{1ak} \le \tau_a \le p_{2ak}; \\ 0, \ \tau_a > p_{2ak}; \end{cases}$$

– треугольную функцию:

$$\mu_{ak} = \begin{cases} 0, \ \tau_a < p_{0ak}; \\ \dfrac{\tau_a - p_{0ak}}{p_{1ak} - p_{0ak}}, \ p_{0ak} \le \tau_a < p_{1ak}; \\ \dfrac{p_{2ak} - \tau_a}{p_{2ak} - p_{1ak}}, \ p_{1ak} \le \tau_a < p_{2ak}; \\ 0, \ \tau_a \ge p_{2ak}. \end{cases}$$

Исходя из особенностей решаемой задачи и исследуемых объектов или процессов, можно использовать и другие функции принадлежности [6–8]: сплайн-функцию, S-образную и Z-образную кривые, сигмоидную функцию, функцию Гаусса, колоколообразную функцию и другие.

Для определения значений параметров p_{0ak}, p_{1ak}, p_{2ak}, p_{3ak} функций принадлежности μ_{ak} выполняется разбиение каждого численного a-го признака $\tau_a \in I$ на некоторое количество интервалов $N_{int.a}$ с последующим определением границ полученных интервалов: $\Delta_{ak} = [l_{ak}; r_{ak})$, где l_{ak} и r_{ak} – соответственно левая и правая границы k-го интервала Δ_{ak} a-го признака τ_a.

Количество интервалов $N_{int.a}$, например, может задаваться пользователем как параметр метода. В таком случае ширина Δ_{ak} каждого из диапазонов

разбиения a-го признака τ_a определяется как отношение диапазона $\Delta_a = [\tau_{a\min}; \tau_{a\max}]$ его значений к заданному количеству интервалов $N_{int.a}$.

Кроме того, количество интервалов $N_{int.a}$, а также их границы могут быть определены с помощью метода, описанного ниже и не требующего участия пользователя в разбиении интервала значений признаков.

После определения видов функций принадлежности μ_{ak} и их параметров они применяются к фактическим значениям входных аргументов $\tau_a \in I$ в каждой транзакции T_j базы данных D. В результате такого применения четкому значению признака τ_a ставятся в соответствие степени его принадлежности к нечетким множествам, вследствие чего образуются транзакции с нечеткими значениями признаков $\tau_a \in I : T_j \to FuzzyT_j$. Полученные таким образом нечеткие транзакции $FuzzyT_j$ образуют нечеткую транзакционную базу правил $FuzzyD$.

После фаззификации базы транзакций и разбиения диапазона значений численных признаков на некоторые интервалы выполняется кластерный анализ – построение групп компактно расположенных транзакций в пространстве признаков.

В результате кластеризации выделяется $N_{kl.}$ кластеров. Для определения значимости каждого элемента $\tau_a \in I$ будем оценивать его влияние для отнесения транзакции к каждому из кластеров. Очевидно, чем меньше ширина диапазона изменения значений a-го признака во множестве транзакций кластера K_b ($b = 1,2,...,N_{kl.}$), тем выше его значимость в данном кластере. Ширину диапазона будем оценивать как среднеквадратическое отклонение [9]:

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

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

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

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

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

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

– как отношение ранга Rg_a к сумме рангов всех признаков:

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

– как отношение ранга Rg_a к максимальному значению рангов:

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

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

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

Поэтому значимость признака τ_a будем вычислять по одной из формул:

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

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

– нормированное значение величины $N_{int.a}$:

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

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

С целью анализа и исследования эффективности предложенного метода оценим его вычислительную сложность O – количество элементарных операций, необходимых для решения конкретной задачи.

Определение индивидуальной значимости w_a признаков τ_a в базе транзакций D связано с необходимостью их группировки во множестве компактно расположенных транзакций. Рассмотрим подход, учитывающий границы интервалов разбиения признаков в кластерах. Выше описано, что такой подход связан с необходимостью сортировки каждого признака τ_a. Поэтому вычислительная сложность данного этапа непосредственно связана с вычислительной сложностью используемого метода сортировки и может быть определена как $O_2\left(|I|\cdot O_{sort.}\right)$.

Эффективными методами сортировки являются методы, использующие деревья (турнирная сортировка, сортировка посредством поискового дерева), метод пирамидальной сортировки, метод быстрой сортировки К. Хоара, вычислительная сложность которых составляет $O_{sort.} = O_{sort.}\left(N_D \log_2\left(N_D\right)\right)$ [10].

Следовательно, вычислительная сложность предложенного метода может быть оценена следующим образом:

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

Оценка вычислительной сложности O показывает, что количество элементарных операций, необходимых для отбора признаков с помощью предложенного метода, линейно зависит от количества $|I|$ признаков $\tau_a \in I$ в транзакционной базе данных D, а также пропорциональна величине $N_D \log_2\left(N_D\right)$, где N_D – количество транзакций в D. Такая оценка позволяет сделать вывод о том, что предложенный метод является вычислительно эффективным, поскольку зависимость его элементарных операций от размера входных данных является полиномиальной.

3. Эксперименты и результаты

С целью проведения экспериментов по исследованию свойств и характеристик предложенного метода отбора информативных признаков он был программно реализован на языке программирования C#.

Экспериментальное исследование разработанного метода выполнялось на основе тестовых данных, представленных в виде транзакционных баз данных D. Результаты экспериментов приведены в табл. 1, где используются следующие обозначения: N_D – количество транзакций T_j в базе D; $|I|$ – количество элементов (признаков), $\tau_a \in I$, из которых могли формироваться транзакции; $\left|T_j\right|$ – среднее количество признаков в транзакциях базы D; $\left|I^*\right|$ – количество отобранных признаков; t – время отбора признаков.

Таблица 1

Результаты экспериментальных исследований

№	Характеристики базы транзакций D			Результаты отбора признаков	
	N_D	$\|I\|$	$\overline{\left\|T_j\right\|}$	$\left\|I^*\right\|$	t
1	10000	100	10	23	0,37
2	10000	1000	10	31	1,72
3	50000	100	10	27	1,43
4	50000	1000	20	39	8,21
5	100000	5000	10	32	12,19
6	100000	10000	20	45	20,57

Как видно из таблицы, время работы предложенного метода существенно зависит от количества транзакций T_j и количества элементов (признаков) в базе D, что подтверждает оценку вычислительной сложности $O = O\left(|I|\cdot N_D \log_2\left(N_D\right)\right)$ метода.

После отбора информативных признаков выполнялось построение численных ассоциативных правил на основе баз транзакций, из которых были исключены малоинформативные признаки. Результаты экспериментов показали, что время поиска ассоциативных правил существенно сократилось, что обусловлено снижением размерности исходных транзакционных баз данных.

Таким образом, результаты экспериментов показали, что разработанный метод позволяет извлекать из баз транзакций информативные признаки, используя при этом априорную информацию об их значимости, что сокращает пространство поиска и время извлечения ассоциативных правил.

Выводы

В работе решена актуальная задача автоматизации поиска информативных признаков в транзакционных базах данных.

Научная новизна работы заключается в том, что предложен метод отбора информативных признаков, предполагающий вычисление информативности каждого признака в транзакционной базе данных D, а также выделение интервалов разбиения признаков без необходимости задания количества интервалов разбиений, что позволяет уменьшить степень участия пользователя и влияние его субъективных оценок на результаты процесса извлечения ассоциативных правил, что в свою очередь снижает вероятность извлечения ассоциативных правил, некорректно описывающих исследуемые объекты или процессы.

Практическая ценность полученных результатов заключается в том, что на основе предложенного метода разработано программное обеспечение, позволяющее выполнять отбор информативных признаков.

Работа выполнена в рамках госбюджетной научно-исследовательской темы Запорожского нацио-

нального технического университета «Интеллектуальные информационные технологии автоматизации проектирования, моделирования, управления и диагностирования производственных процессов и систем» (номер государственной регистрации 0112U005350).

Список литературы: 1. *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. **2.** *Zhang C.* Association rule mining: models and algorithms / C. Zhang, S. Zhang. – Berlin : Springer-Verlag. – 2002. – 238 p. **3.** *Субботін С. О.* Неітеративні, еволюційні та мультиагентні методи синтезу нечіткологічних і нейромережних моделей: монографія / С. О. Субботін, А. О. Олійник, О. О. Олійник ; під заг. ред. С. О. Субботіна. – Запоріжжя : ЗНТУ, 2009. – 375 с. **4.** *Guyon I.* An Introduction to Variable and Feature Selection / I. Guyon, A. Elisseeff // Journal of Machine Learning Research. – 2003. – №3. – P. 1157–1182. **5.** *Dash M.* Feature Selection for Classification / M. Dash, H. Liu // Intelligent Data Analysis. – 1997. – №1. – P. 131–156. **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.** *Zadeh L.* Fuzzy sets / L. Zadeh // Information and Control. – 1965. – № 8. – P. 338–353. **8.** *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. **9.** *Айвазян С. А.* Прикладная статистика: Исследование зависимостей / С. А. Айвазян, И. С. Енюков, Л. Д. Мешалкин. – М.: Финансы и статистика, 1985. – 487 с. **10.** *Кнут Д.* Искусство программирования. В 3-х томах. Т. 2 Сортировка и поиск / Д. Кнут. – М. : Вильямс, 2007. – 824 с.

Поступила в редколлегию 09.07.2013

УДК 004.93

Метод визначення індивідуальної значущості ознак для видобування чисельних асоціативних правил / Т. А. Зайко, А. О. Олійник, С. О. Субботін // Біоніка інтелекту: наук.-техн. журнал. 2013. – № 2 (81). – С. 61-65.

Розглянуто задачу визначення індивідуальної значущості ознак у транзакційних базах даних. Розроблено метод відбору інформативних ознак, що передбачає обчислення інформативності кожної ознаки в транзакційній базі даних, а також виділення інтервалів розбиття ознак без необхідності завдання кількості інтервалів розбиттів. Проведено експерименти з рішення тестових задач відбору ознак.

Табл. 1. Бібліогр.: 10 найм.

UDC 004.93

Method for determining the individual significance of features for extraction of quantitative association rules / T. A. Zayko, A. O. Oliinyk, S. A. Subbotin // Bionics of Intelligence: Sci. Mag. – 2013. – № 2 (81). – P. 61-65.

The problem of determining the individual significance of features in transactional databases is considered. A method of feature selection is developed, the proposed method involves the calculation of informativeness of each feature in the transactional database, as well as the selection of intervals of features without the need to specify the number of partition intervals. The experiments on test problems of feature selection are conducted.

Tab. 1. Ref.: 10 items.