УДК 681.3.00:007

Д.Э. Ситников, О.А. Романенко, Е.В. Титова

Харьковская государственная академия культуры, Харьков

НАХОЖДЕНИЕ МИНИМИЗИРОВАННЫХ НАБОРОВ ПРИЗНАКОВ ДЛЯ ОПИСАНИЯ ДИСКРЕТНЫХ ОБЪЕКТОВ В БАЗАХ ДАННЫХ

Нахождение зависимостей в данных в виде логических правил является одной из типичных задач Интеллектуального анализа данных (Data Mining), при решении которой весьма актуальным является ограничение количества признаков, входящих в результирующее правило. Определение важности того или иного признака позволяет сократить количество анализируемых данных и как следствие, сократить количество логических условий в найденных правилах.

Ключевые слова: *логические правила, мощность аппроксимации, значимость признака, нечёткие множества, верхняя и нижняя аппроксимации, важность признака.*

Введение

Постановка проблемы. Современные методы Интеллектуального Анализа Данных (Data Mining) позволяют выявлять определенные закономерности и зависимости (скрытые знания) в огромных массивах обрабатываемых данных. К наиболее типичным задачам Data Mining относят: описание классов с целью построения их моделей, анализ ассоциаций, классификация, регрессионный анализ, кластерный анализ, агрегирование (обобщение), построение зависимостей, обнаружение отклонений.

Данные методы направлены на обработку сверхбольших объемов информации и, следовательно, задача ограничения количества как анализируемых признаков, так и признаков, входящих в результирующую модель, стоит достаточно остро.

Нахождение зависимостей в данных в виде логических правил является одной из типичных задач Data Mining, при решении которой также весьма актуальным является ограничение количества признаков, входящих в результирующее правило. Таким образом, проблема оценки значимости признаков объектов в базах данных при построении логических правил является весьма актуальной. В зависимости от установленной важности признаки могут подвергаться сокращению.

Анализ последних исследований и публикаций. Теория приближенных множеств, предложенная Z. Pawlak [1], является эффективным математическим инструментом, позволяющим решать задачу классификации новых объектов на основании имеющейся информации. Причем информация, связанная с данными объектами, не позволяет однозначно отнести их к некоторому множеству.

Приближенное множество характеризуется его верхней и нижней аппроксимациями. Нижняя аппроксимация содержит все элементы, про которые можно сказать, что они точно принадлежат заданному множеству. Верхняя аппроксимация содержит элементы, которые могут принадлежать множеству. Разница между верхней и нижней аппроксимацией определяется как граничный регион, размер которого характеризует "приближенность" рассматриваемого множества.

Для определения верхней и нижней аппроксимаций приближенного множества авторами ранее был предложен алгебраический подход [2]. Данный подход позволяет использовать для нахождения аппроксимаций только логические операции, что существенно увеличивает скорость генерации классификационных логических правил. Поскольку методы Data Mining направлены на обработку и анализ сверхбольших объемов информации, то предложенный подход имеет большие преимущества при обработке данных на ЭВМ. Алгебраический подход получил свое развитие в работах [3, 4], где был предложен метод поиска аппроксимаций приближенных множеств для небинарных признаков.

Одной из концепций, связанных с теорией приближенных множеств, является концепция редактов [5 – 7]. Редакт определяется как минимальное подмножество атрибутов, которые описывают приближенное множество без потери информации (иными словами, описывают его с такой же степенью "приближенности", как и полный набор признаков). Поиск редактов является интересной и сложной проблемой по мнению Z. Pawlak [5]. Следует отметить, что для произвольного приближенного множества может не существовать ни одного редакта (количество редактов равно нулю), а может существовать один и более редактов.

В случае существования нескольких редактов для данного приближенного множества встает вопрос определения *Важности* признаков, составляющих редакт. Авторы предполагают, что более важные признаки формируют более значимые редакты, что позволяет сравнивать полученные минимальные наборы.

Используя такое преимущество алгебраического подхода, как минимальное время генерации аппроксимаций, авторами был предложен метод определения *Важности* признака объекта базы данных [8]. Данный метод основан на подсчете изменения граничного региона при удалении того или иного признака. Чем больше увеличивается граничный регион, тем более важным считается удаленный признак.

Цель статьи. Предложить метод определения *Важности* атрибута в небинарном случае, а также алгоритм, позволяющий установить определенный порядок исключения признаков для нахождения наиболее значимых редактов.

Определение аппроксимаций дляянебинарныхяпризнаков

Метод определения аппроксимаций для небинарных признаков детально описан в [3, 4].

Его суть заключается в следующем. Предположим, что есть конечное непустое множество объектов $U=\{a_1, a_2, \ldots, a_n\}$, называемое универсум. Множество унарных предикатов (функций принимающих одно из значений из множества $\{0,1\}$) определённое на U: $P_1(t)$, $P_2(t)$, …, $P_k(t)$, было названо координатами.

Предикаты $P_1, P_2, \ldots, P_k$ могут быть интерпретированы как характеристические функции для некоторых свойств объектов универсума. В данном случае объект a_i имеет свойство P_j только в случае если $P_j(a_i) = 1$.

Следуя базовым концепциям теории приближенных множеств, необходимо описать некоторое множество $X \subseteq U$ в терминах координат. Так как существует отношение "один-к-одному" между всеми предикатами, определёнными на U, вместо $X \subseteq U$ мы можем рассматривать предикат X(t), который равен 1 тогда и только тогда, когда $t \in X$. Таким образом, необходимо дать описание предиката X(t) в терминах предикатов $P_1, P_2, \ldots, P_k$ (табл. 1)

Таблица 1

Предикаты $P_1, P_2, \ldots, P_k$ и предикат X, определённые на универсуме (в общем виде)

Элементы / Признаки	a_1	a_2	…	a_n
P_1	δ_{11}	δ_{12}	…	δ_{1n}
P_2	δ_{21}	δ_{22}	…	δ_{2n}
…	…	…	…	…
P_k	δ_{k1}	δ_{k2}	…	δ_{kn}
X	λ_1	λ_2	…	λ_n

Здесь $\delta_{1j} \in \{0,1..m_1\}$, $\delta_{2j} \in \{0,1..m_2\}$,...$\delta_{kj} \in \{0,1..m_k\}$, $\lambda_j \in \{0,1\}$, если $\delta_{ij}=w$ то $P_i(a_j)=w$, если $\lambda_j = 1$ то $X(a_j) = 1$, если $\lambda_j = 0$ то $X(a_j) = 0$.

В общем случае расчет аппроксимаций для такого набора данных производится по формулам [3]:

$$\begin{aligned} I_* = &(\lambda_1 \,\&\, P_1^{\delta_{11}} \,\&\, P_2^{\delta_{21}} \,\&\, \ldots \,\&\, P_k^{\delta_{k1}}) \vee \\ &\vee(\lambda_2 \,\&\, P_1^{\delta_{12}} \,\&\, P_2^{\delta_{22}} \,\&\, \ldots \,\&\, P_k^{\delta_{k2}}) \vee \ldots \\ &\vee(\lambda_n \,\&\, P_1^{\delta_{1n}} \,\&\, P_2^{\delta_{2n}} \,\&\, \ldots \,\&\, P_k^{\delta_{nn}}), \end{aligned} \quad (1)$$

$$\begin{aligned} I_* = &(\lambda_1 \vee \overline{P_1^{\delta_{11}}} \vee \overline{P_2^{\delta_{21}}} \vee \ldots \vee \overline{P_k^{\delta_{k1}}}) \,\& \\ \&(&\lambda_2 \vee \overline{P_1^{\delta_{12}}} \vee \overline{P_2^{\delta_{22}}} \vee \ldots \vee \overline{P_k^{\delta_{k2}}}) \,\& \ldots \\ \&(&\lambda_n \vee \overline{P_1^{\delta_{1n}}} \vee \overline{P_2^{\delta_{2n}}} \vee \ldots \vee \overline{P_k^{\delta_{nn}}}), \end{aligned} \quad (2)$$

где $P_k^{\delta_{ij}}=1$ если $P_k(a_i)= \delta_{ij}$, иначе $P_k^{\delta_{ij}}=0$, и $\overline{P_k^{\delta_{ij}}}=0$ если $P_k(a_i)=\delta_{ij}$, иначе $\overline{P_k^{\alpha_n}}=1$ для любого P.

Рассмотрим конкретный пример работы алгоритма (табл. 2).

Таблица 2

Фрагмент базы данных

Элементы / Признаки	a_1	a_2	a_3	a_4	a_5
P_1	1	0	2	0	0
P_2	0	2	0	0	2
P_3	0	2	1	1	2
X	0	1	0	1	0

Предположим что характеристические функции P_1, P_2, P_3, описывающие свойства объектов a_1, ...,a_5, могут принимать значения из множества $\{0,1,2\}$.Используя формулы (1) и (2), получаем следующие формулы для аппроксимаций:

$$I^* = (P_1^0 \,\&\, P_2^2 \,\&\, P_3^2) \vee (P_1^0 \,\&\, P_2^0 \,\&\, P_3^1) \quad (3)$$

$$I_* = (\overline{P_1^1} \vee \overline{P_2^0} \vee \overline{P_3^0}) \& (\overline{P_1^2} \vee \overline{P_2^0} \vee \overline{P_3^1}) \& (\overline{P_1^0} \vee \overline{P_2^2} \vee \overline{P_3^2}) \quad (4)$$

Рассчитываем значение аппроксимаций (табл. 3):

Таблица 3

Аппроксимации множества X

Элементы / Признаки	a_1	a_2	a_3	a_4	a_5
P_1	1	0	2	0	0
P_2	0	2	0	0	2
P_3	0	2	1	1	2
X	0	1	0	1	0
I^*	0	1	0	1	1
I_*	0	0	0	1	0

Таким образом, нижняя аппроксимация I_* содержит один элемент $\{a_4\}$, верхняя аппроксимация I^* содержит три элемента $\{a_2, a_4, a_5\}$.

Важность атрибутов

Под *Важностью* признака P_i понимается некоторая численная величина, показывающая, насколько возрастает (в процентном соотношении) "приближенность" множества X при удалении признака P_i [8]:

$$V(P_i) = \frac{\Delta(BN_I)}{M(X)} * 100\%, \quad (5)$$

где $(\Delta(BN_I))$ – изменение граничного региона (после удаления признака P_i; $M(X)$ – мощность множества X.

Если $V(P_i) \geq \text{minDeterioration}$, то признак P_i является значимым, если $V(P_i) < \text{minDeterioration}$, то признак P_i не является значимым.

В данном случае minDeterioration – заданное пороговое значение, величина которого задается аналитиком и зависит от конкретных данных и решаемой задачи.

Продемонстрируем вышесказанное на примере.

Мощность граничного региона для начальной таблицы получаем равной 2 (верхняя аппроксимация содержит 3 объекта, нижняя аппроксимация содержит 1 объект: 3–1=2). Исключаем признак P_1 и рассчитываем новые аппроксимации (табл. 4):

Таблица 4

Аппроксимации множества X
для минимизированного набора признаков

Элементы / Признаки	a_1	a_2	a_3	a_4	a_5
P_2	0	2	0	0	2
P_3	0	2	1	1	2
X	0	1	0	1	0
I^*	0	1	1	1	1
I_*	0	0	0	0	0

Теперь мощность граничного региона равна 4 (верхняя аппроксимация содержит 4 объекта, нижняя аппроксимация пуста). Таким образом, *Важность* атрибута P_1 будет равна:

$$V(P_1) = \frac{(4-2)}{2} * 100\% = 100\%.$$

Рассчитаем важность признака P_2 (табл. 5):

Таблица 5

Аппроксимации множества X
для минимизированного набора признаков

Элементы / Признаки	a_1	a_2	a_3	a_4	a_5
P_1	1	0	2	0	0
P_3	0	2	1	1	2
X	0	1	0	1	0
I^*	0	1	0	1	1
I_*	0	0	0	1	0

Мощность граничного региона равна 2, *Важность* признака P_2:

$$V(P_2) = \frac{(2-2)}{2} * 100\% = 0\%.$$

Рассчитываем *Важность* признака P_3 (табл. 6):

Таблица 6

Аппроксимации множества X
для минимизированного набора признаков

Элементы / Признаки	a_1	a_2	a_3	a_4	a_5
P_1	1	0	2	0	0
P_2	0	2	0	0	2
X	0	1	0	1	0
I^*	0	1	0	1	1
I_*	0	0	0	1	0

Важность атрибута P_3:

$$V(P_3) = \frac{(2-2)}{2} * 100\% = 0\%.$$

Таким образом, можно утверждать, что признак P_1 является важным, а признаки P_2 и P_3 не важны, т.е. их исключение не приводит к «ухудшению» данных.

Методы поиска редактов

Информация о *Важности* атрибутов позволяет исключать некоторые атрибуты (пары атрибутов) из начального набора данных. Большой интерес представляет задача нахождения минимальных наборов признаков, адекватно описывающих исходное множество (так называемых редактов). В общем случае для произвольного приближенного множества может существовать более одного редакта. Выбор "лучшего" редакта из множества допустимых может осуществляться согласно различным критериям, однако, задача получения минимального набора наиболее *Важных* признаков представляет большой интерес.

Таким образом, используя такой критерий, как *Важность* того или иного признака, можно предложить следующий метод нахождения редактов.

Первым этапом поиска локальных редактов является определение *Важности* каждого из атрибутов, входящих в исходные данные и сортировка атрибутов по возрастанию *Важности*.

Здесь можно выделить два подхода. Первый заключается в том, чтобы получить все возможные редакты для данного приближенного множества, второй заключается в получении заданного количества наиболее значимых локальных редактов.

В первом случае необходимо организовать поиск всех возможных поднаборов атрибутов, из которых исключен(ы) атрибут(ы) с *Важностью* меньшей, чем допустимая (minDeterioration). Следует учесть, что расчет *Важности* атрибутов следует проводить в сравнении с начальным (неизмененным) набором данных.

Пусть в отсортированном по мере возрастания *Важности* списке атрибутов $P_1, P_2, ..., P_k, P_{k+1}, ...P_n$ первые k атрибутов не преодолевают порог minDeterioration. Тогда общее количество искомых наборов будет: $\sum_{m=1}^{k} C_n^m$ (полный перебор). Однако, следует отметить, что если исключено j атрибутов (j<k) и изменение граничного региона превосходит допустимый порог $\Delta(BN_I)$ >minDeterioration, то перебор вариантов по этому направлению заканчивается.

Общая работа алгоритма заканчивается тогда, когда для всех комбинаций атрибутов не осталось редактов, из которых можно было бы исключить хоть один атрибут.

В рассмотренном выше примере единственным вариантом является попытка исключить атрибуты P_2 и P_3, т.к. они имеют нулевую важность (табл. 7).

Таблица 7

Аппроксимации множества X
для минимизированного набора признаков

Элементы / Признаки	a_1	a_2	a_3	a_4	a_5
P_1	1	0	2	0	0
X	0	1	0	1	0
I^*	0	0	0	0	0
I_*	0	0	0	0	0

Получаем что важность набора атрибутов P_2 и P_3 (при условии их совместного исключения) равна:

$$V(P_2, P_3) = \frac{(2-0)}{2} * 100\% = 100\%,$$

из чего следует, что эти атрибуты вместе исключать нельзя.

Таким образом, получаем, что локальными редактами, которые можно использовать в дальнейшем для обработки различными алгоритмами Data Mining, являются наборы данных, полученные в таблицах 5 и 6.

Во втором случае (когда необходимо получить заданное количество наиболее значимых локальных редактов), описанный выше метод поиска модифицируется. *Важность* каждого признака пересчитывается после исключения того или иного атрибута из рассмотрения. Таким образом, на каждом шаге сокращается признак, имеющий наименьшую *Важность*. Если существуют признаки с одинаковой *Важностью*, то исключается любой из них.

Перебор прекращается, если найдено необходимое количество редактов (из полученных редактов выбираются те, количество атрибутов в которых минимально).

Следует отметить, что этот вариант поиска может не дать в результате ни одного редакта, содержащего только значимые признаки от P_1 до P_k. В этом случае применяется первый вариант с попыткой получения всех возможных локальных редактов.

Выводыя

Предлагаемый метод определения *Важности* признака позволяет сократить количество атрибутов, участвующих в описании приближенного множества. Благодаря этому, классификационные правила, получаемые на основе аппроксимаций, приобретают более краткий и понятный для аналитика вид.

Предлагаемый алгоритм поиска локальных редактов можно использовать для двух случаев: когда необходимо получить все редакты для данного приближенного множества и когда необходимо найти определенное количество редактов. Предлагаемый алгоритм относится к алгоритмам ограниченного перебора, однако, принимая во внимание такое преимущество алгебраического подхода по нахождению аппроксимаций, как использование только операций сравнения и булевых операций, можно утверждать, что время работы алгоритма будет невелико даже на больших массивах данных.

Списокялитературыя

1. Pawlak Z. Rough set approach to knowledge-based decision support / Z. Pawlak // Proc. Of the 14 European Conference on Operational Research. – Jerusalem, Israel, 1995.

2. Sitnikov D. An algebraic approach to defining rough set approximations and generating logic rules / D. Sitnikov, O. Ryabov // Data Mining V. – Malaga, Spain, 2004. – P. 179-188.

3. Sitnikov D. A generalized algebraic approach to finding rough set approximations and generating logic rules / D. Sitnikov, O. Ryabov, O. Titova, O. Romanenko // Data Mining VIII. – WIT Press., 2007. – P. 3-12.

4. Ситников Д.Э. Обобщенный логико-алгебраический метод нахождения аппроксимаций приближенных множеств и генерации на их основе логических правил / Д.Э. Ситников, О.А. Романенко, С.В. Титов, Е.В. Титова // Збірник наукових праць Харківського університету Повітряних Сил. – Х.: ХУПС, 2007. – Вип. 5(15). – С. 115-119.

5. Pawlak Z. Rough set approach to knowledge-based decision support / Z. Pawlak // European Journal of Operational Research, 99. – 1997. – P. 420-432.

6. Sitnikov D. An approach to finding reduced sets of information features describing discrete objects based on rough sets theory / D. Sitnikov, O. Titova, O. Romanenko, O. Ryabov // Data Mining IX. – WIT Press., 2008. – P. 3-11.

7. Ситников Д.Э. Метод нахождения минимизированных наборов признаков в базах данных с использованием теории приближённых множеств / Д.Э. Ситников, О.А. Романенко, Е.В. Титова, С.В. Титов // Системи обробки інформації. – Х.: ХУПС, 2007. – Вип. 7(65). – С. 91-95.

8. Романенко О.А. Метод оценки значимости признаков объектов в базах данных с использованием теории приближённых множеств / О.А. Романенко, Е.В. Титова, А.А. Усань // Збірник наукових праць Харківського університету повітряних сил. – Х.: ХУПС, 2008. – Вип. 1(16). – С. 94-96.

Поступила в редколлегию 16.06.2009

Рецензент: д-р техн. наук, проф. И.В. Гребенник, Харьковский национальный университет радиоэлектроники, Харьков.

ЗНАХОДЖЕННЯ МІНІМІЗОВАНИХ НАБОРІВ ОЗНАК ДЛЯ ОПИСУ ДИСКРЕТНИХ ОБ'ЄКТІВ В БАЗАХ ДАНИХ

Д.Е. Ситніков, О.А. Романенко, О.В. Тітова

Знаходження залежностей в даних у вигляді логічних правил є одним з типових завдань Інтелектуального аналізу даних (Data Mining), при рішенні якої вельми актуальним є обмеження кількості ознак, що входять в результуюче правило. Визначення важливості тієї або іншої ознаки дозволяє скоротити кількість аналізованих даних і як наслідок, скоротити кількість логічних умов в знайдених правилах.

Ключові слова: *логічні правила, потужність апроксимації, значущість ознаки, нечіткі множини, верхня і нижня апроксимації, важливість ознаки.*

FINDING MINIMAL SETS OF FEATURES FOR DESCRIBING DISCRETE INFORMATION OBJECTS IN DATABASE

D.E. Sitnikov, O.A. Romanenko, E.V. Titova

Finding of dependences in information as logical rules is one of typical tasks of the Intellectual analysis of data (Data Mining), at the decision of which very actual is limitation of amount of signs, included in a resulting rule. Determination of importance of one or another sign allows to shorten the amount of the analysed information and as a result, to shorten the amount of logical terms in the found rules.

Keywords: *logical rules, power of approximation, meaningfulness of sign, fuzzy sets, overhead and lower approximations, importance of sign..*