УДК 681.3.00 : 007

Д.Э. Ситников[1], О.А. Романенко[2], Е.В. Титова[1], С.В. Титов[1]

[1]*Харьковская государственная академия культуры*
[2]*ООО Укргазтехкомплекс, Харьков*

# МЕТОД НАХОЖДЕНИЯ МИНИМИЗИРОВАННЫХ НАБОРОВ ПРИЗНАКОВ В БАЗАХ ДАННЫХ С ИСПОЛЬЗОВАНИЕМ ТЕОРИИ ПРИБЛИЖЁННЫХ МНОЖЕСТВ

*Интеллектуальный анализ данных направлен на обработку больших массивов информации с целью обнаружения в них скрытых закономерностей. Нахождение минимизированного набора признаков, адекватно описывающих множество, является очень важной задачей, т.к. ведет к сокращению размерности обрабатываемых данных. Статья посвящена описанию метода нахождения несократимых наборов признаков для приближенных множеств с использованием алгебры конечных предикатов.*

***интеллектуальный анализ данных, приближенное множество, неразличимые отношения, верхняя и нижняя аппроксимация, конечные предикаты, несократимые признаки, глобальный редакт, локальный редакт***

## Введение

**Постановка проблемы.** Современные методы интеллектуального анализа данных (Data Mining) в последние годы привлекают большое внимание специалистов, так как позволяют выявлять определенные закономерности и зависимости в больших информационных массивах. Эти закономерности в дальнейшем учитываются в процессе принятия решений, а также служат основой создания разнообразных корпоративных баз знаний и систем управления знаниями предприятия или организации.

Поскольку методы Data Mining направлены на обработку и анализ сверхбольших объемов информации, то одним из основных требований к ним является удовлетворительное время работы с современными базами и хранилищами данных. Это требование, безусловно, напрямую связано с проблемой ограничения размерности решаемой задачи, иными словами, с ограничением количества анализируемых признаков.

Таким образом, одной из проблем, стоящих на данном этапе перед разработчиками интеллектуальных систем анализа данных, является проблема нахождения минимальных наборов признаков, позволяющих адекватно описывать имеющиеся данные. Признаки, не входящие в данные наборы, могут считаться сократимыми и не учитываться при анализе.

**Анализ последних исследований и публикаций.** Теория приближенных множеств, разработанная польским ученым Здиславом Павлаком [1, 2], послужила базой для развития целого направления интеллектуального анализа данных, основанного на этой теории. Одним из основных понятий теории приближенных множеств является понятие "неразличимости" (или "неразличимого" отношения). Утверждается, что с разными элементами множества может ассоциироваться одинаковая информация, что делает невозможным точное определение принадлежности этих элементов к некоторому множеству (такое множество называют приближенным). Приближенное множество характеризуется двумя аппроксимациями: нижней, которая определяет элементы, которые однозначно принадлежат данному множеству, и верхней, которая определяет элементы, которые могут принадлежать данному множеству.

Пусть X – некоторое множество, являющееся подмножеством универсума U и I – отношение "неразличимости". Нижней аппроксимацией X является:

$$I_*(X) = \{x \in U : I(x) \subseteq X\},$$

а верхней аппроксимацией:

$$I^*(X) = \{x \in U : I(x) \cap X \neq 0\},$$

где I(x) определяет множество объектов, "неразличимых" с x (с точки зрения имеющейся информации) [1]. Под граничным регионом (boundary region) понимается $BN_I = I^*(X) - I_*(X)$. Таким образом, граничный регион содержит элементы, которые принадлежат верхней аппроксимации множества и не принадлежат нижней. Если $BN_I \neq 0$, то множество X является приближенным.

Для приближенных множеств вводится также понятие функции принадлежности:

$$\mu_X^I(x) = \frac{|X \cap I(x)|}{|I(x)|}.$$

Очевидно, что $0 \le \mu_X^I(x) \le 1$.

Функция принадлежности может быть использована для определения верхней, нижней аппроксимаций приближенного множества, а также граничного региона:

$$I_*(X) = \{x \in U : \mu_X^I(x) = 1\};$$

$$I^*(X) = \{x \in U : \mu_X^I(x) > 0\};$$

$$BN_I(X) = \{x \in U : 0 < \mu_X^I(x) < 1\}.$$

Наборы "неразличимых" элементов, с которыми ассоциирована одинаковая информация, Павлак называет "гранулами". В своей работе [3] он также вводит понятие редакта (от англ. reduction – уменьшение, сокращение). Под редактом понимается минимальный набор признаков, позволяющий различать "гранулы". Минимальность этого набора подразумевает невозможность его сокращения без потери различимости "гранул". Таким образом, редакт определяет набор признаков, адекватно описывающий приближенное множество. Нахождение редакта (редактов) для произвольного множества Павлак называет интересной, но довольно сложной задачей [3] и не приводит никаких способов ее решения. Один из подходов к решению этой задачи был предложен в [4].

**Цель статьи:** На основании теории приближенных множеств разработать метод нахождения минимизированных наборов признаков в базах данных с использованием алгебры конечных предикатов; предложить параметр, позволяющий оценивать найденные редакты.

## Понятие редакта по Павлаку

Рассмотрим понятие редакта, введенное З. Павлаком, на простом примере. Пусть существует фрагмент медицинской базы данных, содержащий информацию о 6-ти пациентах (табл. 1) [1].

Строки этой таблицы содержат значения атрибутов (симптомов заболевания), столбцы соответственно пациентов A1...A6. Из таблицы видно, что если у пациента очень высокая температура, то он на основании имеющейся информации (в терминах данной системы) является больным, и если у пациента нормальная температура, то он здоров.

Таблица 1

Фрагмент медицинской базы данных

| **Пациент** | A1 | A2 | A3 | A4 | A5 | A6 |
|---|---|---|---|---|---|---|
| **Головная боль** | Нет | Да | Да | Нет | Да | Нет |
| **Мышечная боль** | Да | Нет | Да | Да | Нет | Да |
| **Температура** | Высокая | Высокая | Очень высокая | Нормальная | Высокая | Очень высокая |
| **Грипп** | Да | Да | Да | Нет | Нет | Да |

Рассмотрим атрибут "Температура", который не несет однозначной информации о том, болен или нет пациент. Среди пациентов, имеющих высокую температуру, имеются как больные (A1 и A2), так и здоровые (A5). Отметим также, что объекты A2 и A5 "неразличимы" с точки зрения имеющейся информации ("Головная боль", "Мышечная боль", "Температура") и для них невозможно однозначно установить наличие гриппа. В данном случае мы имеем дело с приближенным множеством.

Исключим из рассмотрения атрибут "Головная боль" (табл. 2).

Таблица 2

Фрагмент медицинской базы данных

| **Пациент** | A1 | A2 | A3 | A4 | A5 | A6 |
|---|---|---|---|---|---|---|
| **Мышечная боль** | Да | Нет | Да | Да | Нет | Да |
| **Температура** | Высокая | Высокая | Очень высокая | Нормальная | Высокая | Очень высокая |
| **Грипп** | Да | Да | Да | Нет | Нет | Да |

Заметим, что данный набор атрибутов ("Мышечная боль" и "Температура") адекватно описывает приближенное множество. Пациенты, имеющие очень высокую температуру, однозначно больны, имеющие нормальную температуру, однозначно здоровы. Пациент A1 с точки зрения имеющейся информации болен (мышечная боль имеется, температура высокая). Пациенты A2 и A5 "неразличимы" между собой. Можно утверждать, что "гранульная" структура множества не нарушена. Следовательно, набор признаков {"Мышечная боль", "Температура"} является редактом.

При исключении из рассмотрения атрибута "Мышечная боль" (табл. 3) "гранульная" структура множества также не нарушается. Набор признаков {"Головная боль", "Температура"} также является редактом.

Таблица 3

Фрагмент медицинской базы данных

| **Пациент** | A1 | A2 | A3 | A4 | A5 | A6 |
|---|---|---|---|---|---|---|
| **Головная боль** | Нет | Да | Да | Нет | Да | Нет |
| **Температура** | Высокая | Высокая | Очень высокая | Нормальная | Высокая | Очень высокая |
| **Грипп** | Да | Да | Да | Нет | Нет | Да |

Заметим также, что атрибут "Температура" не может быть исключен без потери различимости "гранул" (при исключении этого атрибута становятся неразличимыми пациенты A1, A4 и A6).

Из данного примера понятно, что для множества может существовать не один редакт. Отметим также, что для множества может не существовать признаков, удаление которых не ведет к нарушению гранульной структуры. В этом случае редактом является весь набор атрибутов.

Назовем такой редакт редактом по Павлаку (глобальным редактом по Павлаку). Редакт по Павлаку содержит несократимый набор признаков. Далее введем понятие локального редакта, с помощью которого продемонстрируем предлагаемый метод нахождения минимизированных наборов признаков.

## Метод нахождения локальных редактов для приближенного множества

Для описания предлагаемого метода нахождения редактов приближенного множества обратимся к определению верхней и нижней аппроксимаций [1, 2].

Верхняя аппроксимация ($I^*$) состоит из элементов, которые могут принадлежать данному множеству. Рассматривая приведенный пример (табл. 1), мы не можем однозначно утверждать, больны или нет пациенты 2 и 5, поскольку относительно атрибутов "Головная боль", "Мышечная боль", "Температура" они "неразличимы" между собой, но имеют разный диагноз. Следовательно, они могут принадлежать множеству больных гриппом и должны быть включены в верхнюю аппроксимацию. Также в верхнюю аппроксимацию войдут пациенты 1, 3, 6.

Нижняя аппроксимация ($I_*$) состоит из элементов, о которых, можно однозначно (с точки зрения имеющейся информации) утверждать, что они входят в рассматриваемое множество X. В нашем примере это пациенты под номерами 1, 3 и 6. Пациент A2 в нижнюю аппроксимацию не входит, поскольку он "неразличим" с пациентом A5.

Таким образом, для данного примера:

$$I^* = \{1, 2, 3, 5, 6\};\ I_* = \{1, 3, 6\}.$$

Метод нахождения верхней и нижней аппроксимаций приближенного множества с использованием алгебры конечных предикатов был предложен в [4, 5]. Он позволяет находить аппроксимации заданного множества за один просмотр набора данных (т.е. имеет линейную сложность), при этом используя только булевы операции. Его суть заключается в следующем: пусть существует непустое множество объектов $U=\{a_1, a_2, \ldots, a_n\}$, называемое универсумом и некоторое множество X, заданное на универсуме (табл. 4).

Таблица 4

Множество X, заданное на универсуме

| Элементы / Атрибут (признак) | $a_1$ | $a_2$ | … | $a_n$ |
|---|---|---|---|---|
| $P_1$ | $\delta_{11}$ | $\delta_{12}$ | … | $\delta_{1n}$ |
| $P_2$ | $\delta_{21}$ | $\delta_{22}$ | … | $\delta_{2n}$ |
| … | … | … | … | … |
| $P_k$ | $\delta_{k1}$ | $\delta_{k1}$ | … | $\delta_{kn}$ |
| X | $\lambda_1$ | $\lambda_2$ | … | $\lambda_n$ |

Для табл. 4:

$\delta_{ij} = 1$, если элемент (объект, запись в базе данных) $a_i$ обладает признаком (атрибутом, свойством) $P_j$;

$\delta_{ij} = 0$, если элемент $a_i$ не обладает признаком $P_j$;

$\lambda_I = 1$, если элемент $a_i$ принадлежит множеству X;

$\lambda_I = 0$, если элемент $a_i$ не принадлежит множеству X.

В общем виде формулы для аппроксимаций записываются в следующем виде [5]:

$$I^* = (\lambda_1 \wedge P_1 * \delta_{11} \wedge P_2 * \delta_{21} \wedge \ldots \wedge P_k * \delta_{k1}) \vee$$
$$\vee(\lambda_2 \wedge P_1 * \delta_{12} \wedge P_2 * \delta_{22} \wedge \ldots \wedge P_k * \delta_{k2}) \vee \ldots$$
$$\ldots \vee (\lambda_n \wedge P_1 * \delta_{1n} \wedge P_2 * \delta_{2n} \wedge \ldots \wedge P_k * \delta_{kn}); \quad (1)$$

$$I_* = (\lambda_1 \vee P_1 * (1-\delta_{11}) \vee P_2 * (1-\delta_{21}) \vee \ldots$$
$$\ldots \vee P_k * (1-\delta_{k1})) \wedge (\lambda_2 \vee P_1 * (1-\delta_{12}) \vee$$
$$\vee P_2 * (1-\delta_{22}) \vee \ldots \vee P_k * (1-\delta_{k2})) \vee \ldots$$
$$\ldots \vee (\lambda_n \vee P_1 * (1-\delta_{1n}) \vee P_2 * (1-\delta_{2n}) \vee \ldots$$
$$\ldots \vee P_k * (1-\delta_{kn})), \quad (2)$$

где $P*\delta = P$, если $\delta = 1$ и $P*\delta = \overline{P}$, если $\delta = 0$ для всех признаков P.

Рассмотрим применение данного метода на примере. Элементы $a_1$, $a_2$, $a_3$, $a_4$, и $a_5$ описаны с помощью признаков $P_1$, $P_2$, $P_3$. Множество X включает в себя элементы $a_2$, $a_4$, $a_5$ (табл. 5).

Таблица 5

Фрагмент базы данных и множество X

| Элементы / Атрибут (признак) | $a_1$ | $a_2$ | $a_3$ | $a_4$ | $a_5$ |
|---|---|---|---|---|---|
| $P_1$ | 1 | 0 | 0 | 1 | 0 |
| $P_2$ | 0 | 0 | 1 | 1 | 1 |
| $P_3$ | 0 | 1 | 0 | 0 | 0 |
| X | 0 | 1 | 0 | 1 | 1 |

Аппроксимации множества X, найденные согласно формулам (1) и (2):

$$I^* = \left(\overline{P_1} \wedge \overline{P_2} \wedge P_3\right) \vee \left(P_1 \wedge P_2 \wedge \overline{P_3}\right) \vee \left(\overline{P_1} \wedge P_2 \wedge \overline{P_3}\right);$$
$$I_* = \left(\overline{P_1} \vee P_2 \vee P_3\right) \wedge \left(P_1 \vee \overline{P_2} \vee P_3\right).$$

Рассчитанные аппроксимации для множества X (табл. 6):

Таблица 6

Аппроксимации множества X

| Элементы / Атрибут (признак) | $a_1$ | $a_2$ | $a_3$ | $a_4$ | $a_5$ |
|---|---|---|---|---|---|
| $P_1$ | 1 | 0 | 0 | 1 | 0 |
| $P_2$ | 0 | 0 | 1 | 1 | 1 |
| $P_3$ | 0 | 1 | 0 | 0 | 0 |
| X | 0 | 1 | 0 | 1 | 1 |
| $I^*$ | 0 | 1 | 1 | 1 | 1 |
| $I_*$ | 0 | 1 | 0 | 1 | 0 |

Возможность ускоренного поиска аппроксимаций позволяет минимизировать начальное множество данных путём выявления локальных редактов.

**Определение.** Локальным редактом называется редакт, полученный путём исключения из множества атрибутов $P_1, P_2, \ldots, P_k$ одного или нескольких атрибутов.

Параметром, согласно которому можно оценить локальный редакт, будем считать граничный регион.

Из формул (1) и (2) видно, что при исключении одного из атрибутов, верхняя аппроксимация может только расти, а нижняя аппроксимация может только уменьшаться. Таким образом, граничный регион,

который по определению равен разности между верхней и нижней аппроксимациями, служит параметром, характеризующим изменения их значений.

Перед нахождением локального редакта зададимся максимально возможным отклонением граничного региона $\Delta(BN_I)$ (это значение зависит от решаемой задачи и устанавливается аналитиком). Будем считать, что для нашего маленького примера максимально возможное отклонение граничного региона должно составлять не более 1-го элемента. Исключим из таблицы 5 атрибут $P_1$ и рассчитаем новые аппроксимации. Получаем результаты, представленные в табл.. 7:

$$I_1^* = (\overline{P_2} \wedge P_3) \vee (P_2 \wedge \overline{P_3});\ I_{1*} = (P_2 \vee P_3) \wedge (\overline{P_2} \vee P_3).$$

Таблица 7

Аппроксимации множества X
для минимизированного набора признаков

| Элементы / Атрибут (признак) | $a_1$ | $a_2$ | $a_3$ | $a_4$ | $a_5$ |
|---|---|---|---|---|---|
| $P_2$ | 0 | 0 | 1 | 1 | 1 |
| $P_3$ | 0 | 1 | 0 | 0 | 0 |
| X | 0 | 1 | 0 | 1 | 1 |
| $I_1^*$ | 0 | 1 | 1 | 1 | 1 |
| $I_{1*}$ | 0 | 1 | 0 | 0 | 0 |

Из табл. 7 видно, что в данном случае мощность верхней аппроксимации не изменилась, зато мощность нижней аппроксимации уменьшилась на 1. Таким образом граничный регион изменился на 1 элемент $(\Delta(BN_I) = 1)$. Это удовлетворяет поставленному условию задачи, следовательно, получен локальный редакт (признаки $P_2$ и $P_3$).

Исключаем признак $P_2$ и рассчитываем новые аппроксимации. Получаем результаты, представленные в табл. 8:

$$I_2^* = (\overline{P_1} \wedge P_3) \vee (P_1 \wedge \overline{P_3}) \vee (\overline{P_1} \wedge \overline{P_3});$$
$$I_{2*} = (\overline{P_1} \vee P_3) \wedge (P_1 \vee P_3).$$

Таблица 8

Аппроксимации множества X
для минимизированного набора признаков

| Элементы / Атрибут (признак) | $a_1$ | $a_2$ | $a_3$ | $a_4$ | $a_5$ |
|---|---|---|---|---|---|
| $P_1$ | 1 | 0 | 0 | 1 | 0 |
| $P_3$ | 0 | 1 | 0 | 0 | 0 |
| X | 0 | 1 | 0 | 1 | 1 |
| $I_2^*$ | 1 | 1 | 1 | 1 | 1 |
| $I_{2*}$ | 0 | 0 | 0 | 0 | 0 |

Из табл. 8 видно, что и верхняя и нижняя аппроксимации изменились на 1 элемент. Таким образом, изменение граничного региона составило две единицы (два элемента) $(\Delta(BN_I) = 2)$. Данный вариант не удовлетворяет поставленному условию задачи и должен быть отброшен. Следует уточнить что данный вариант может быть отброшен и по другим причинам. Нижняя аппроксимация не включает ни одного объекта, а верхняя – включает все объекты, следовательно достигнута полная "неразличимость" между объектами и данный набор данных абсолютно не информативен.

Аналогичным образом, исключая $P_3$, находим оставшийся локальный редакт (табл. 9):

$$I_3^* = (\overline{P_1} \wedge \overline{P_2}) \vee (P_1 \wedge P_2) \vee (\overline{P_1} \wedge P_2);$$
$$I_{3*} = (\overline{P_1} \vee P_2) \wedge (P_1 \vee \overline{P_2}).$$

Таблица 9

Аппроксимации множества X
для минимизированного набора признаков

| Элементы / Атрибут (признак) | $a_1$ | $a_2$ | $a_3$ | $a_4$ | $a_5$ |
|---|---|---|---|---|---|
| $P_1$ | 1 | 0 | 0 | 1 | 0 |
| $P_2$ | 0 | 0 | 1 | 1 | 1 |
| X | 0 | 1 | 0 | 1 | 1 |
| $I_3^*$ | 0 | 1 | 1 | 1 | 1 |
| $I_{3*}$ | 0 | 1 | 0 | 1 | 0 |

В данном случае найден лучший локальный редакт, поскольку ни нижняя, ни верхняя аппроксимации не изменились, и как следствие не изменился граничный регион $(\Delta(BN_I) = 0)$. Таким образом, получен минимизированный набор признаков ($P_1$ и $P_2$), который содержит полную информацию о начальном наборе данных. Следует отметить, что чем меньше элементов содержит граничный регион, тем более четким (в отличие от приближенного) является рассматриваемое множество. Размеры граничного региона служат своего рода показателем "качества" данных (с точки зрения имеющейся информации).

Исключение одного или нескольких атрибутов (признаков) ведет в общем случае к не уменьшению мощности граничного региона.

Таким образом, данную процедуру исключения атрибутов необходимо завершить при достижении неудовлетворительной мощности граничного региона (такой, которая будет отличаться от первоначальной мощности более чем на некоторое заданное значение).

Лучший из локальных редактов $(\min \Delta(BN_I))$ будет являться глобальным редактом. Тут следует заметить, что классическое определение редакта по Павлаку не допускает отличий в аппроксимациях приближенного множества. Павлак утверждает [3], что если хотя бы одна из аппроксимаций, вследствие исключения того или иного признака изменилась, то данный набор признаков не является редактом.

С точки зрения предлагаемого метода можно утверждать, что если верхняя и/или нижняя аппроксимации изменились не более чем на заданное значение $(\Delta(BN_I) \le \varepsilon)$, то данный набор признаков является редактом.

## Выводы

В данной статье представлен метод нахождения минимизированных наборов признаков, позволяющих адекватно описывать заданное множество.

Основной идеей данного метода является нахождение так называемых локальных редактов, т.е. таких наборов атрибутов, которые наиболее близко описывают некоторое множество. Мерой близости описания является сравнение исходной мощности граничного региона с полученной. Этот подход к определению редакта (редактов) представляется более гибким, в отличие от классического определения по Павлаку. Если для некоторого приближенного множества может не существовать классического редакта (иными словами, количество признаков не может быть уменьшено и весь набор признаков является редактом), то при использовании предлагаемого метода могут быть получены минимизированные наборы, с достаточной точностью описывающие множество.

Безусловно, данная задача является одной из разновидностей задач классификации, иначе говоря, определения того принадлежит ли некоторый элемент данному множеству X или нет.

Также следует отметить, что предлагаемый метод использует для нахождения аппроксимаций и редактов только операции сравнения и булевы операции, что делает его привлекательным с вычислительной точки зрения.

Одним из перспективных направлений исследований представляется изучение возможности исключения более одного признака из анализируемого набора атрибутов. Также одной из задач является определение порядка исключения признаков, их "значимости" для описания множества.

## Список литературы

*1. Pawlak Z. Rough set approach to knowledge-based decision support // Proc. of the 14 European Conference on Operational Research Jerusalem, Israel, July 1995. – 121 p.*

*2. Pawlak Z.: Vaguenes and uncertainty: a Rough set perspective // Computational Intelligence, Volume 11 (Issue 2), May 1995. – P. 227-232.*

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

*4. Ситников Д.Э., Вильчинская О.С., Кравец Н.С., Титова Е.В. Определение минимального набора признаков, адекватно описывающих нечеткое множество // Вестник национального технического университета "ХПИ". – Х.: НТУ «ХПИ», 2002. – № 20. – С. 65-70.*

*4. Sitnikov D., Ryabov O. An algebraic approach to defining rough set approximations and generating logic rules// Zanasi, A.; Ebecken, N.; Brebbia, C.; (eds), Data Mining V. – Malaga, Spain, 2004. – P. 179-188.*

*5. Ситников Д.Э., Рябов О.А., Титова Е.В., Романенко О.А. Метод нахождения аппроксимаций приближенных множеств и построения логических правил на основе алгебры конечных предикатов // Системи обробки інформації. – Х.: ХУ ПС, 2007. – № 4 (62). – С.144-149.*

*Поступила в редколлегию 31.05.2007*

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