УДК 681.3.00:007

О.А. Романенко[1], Е.В. Титова[2], А.А. Усань[2]

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

МЕТОД ОЦЕНКИ ЗНАЧИМОСТИ ПРИЗНАКОВ ОБЪЕКТОВ В БАЗАХ ДАННЫХ С ИСПОЛЬЗОВАНИЕМ ТЕОРИИ ПРИБЛИЖЕННЫХ МНОЖЕСТВ

Современные методы интеллектуального анализа данных направлены на обработку больших информационных массивов с целью нахождения в них скрытых закономерностей. Одной из распространенных форм представления найденных зависимостей являются логические правила. Определение значимости того или иного признака позволяет сократить количество анализируемых данных и как следствие, сократить количество логических условий в найденных правилах. Предлагаемый метод определения значимости признака позволяет сократить количество атрибутов, участвующих в описании приближенного множества.

Ключевые слова: логические правила, мощность аппроксимации, значимость признака, нечёткие множества, конечные предикаты, верхняя и нижняя аппроксимации, интеллектуальный анализ данных.

Введение

Постановка проблемы. Современные методы интеллектуального анализа данных (Data Mining) позволяют обнаруживать в больших информационных массивах скрытые закономерности – так называемые знания. Полученные зависимости могут представляться в виде моделей: выделенных кластеров, ассоциативных правил и т.д. Виды получаемых моделей зависят от методов их создания. На основании полученных зависимостей осуществляется глубинный анализ данных, делаются прогнозы относительно будущих ситуаций, находятся источники малообъяснимых явлений.

Несмотря на обилие методов Data Mining, приоритет постепенно все более смещается в сторону методов поиска логических закономерностей в данных в виде правил "Если …, то …". С их помощью решаются задачи классификации, распознавания образов, прогнозирования, установления ассоциаций между признаками объектов в БД. Кроме того, результаты работы данных методов выражаются в понятной форме и легко интерпретируются. Одной из проблем, стоящих перед разработчиками методов Data Mining является проблема ограничения количества признаков, входящих в результирующее правило. "Набор операторов "ЕСЛИ – ТО" иногда бывает столь же непонятным, как и нейронная сеть, особенно если список условий длинный и сложный" [1].

Таким образом, проблема оценки значимости (важности) признаков объектов в базах данных при построении логических правил является весьма актуальной.

Анализ последних исследований и публикаций. Теория приближённых множеств, предложенная Павлаком [2], послужила базой для развития целого направления интеллектуального анализа данных, основанного на этой теории. Согласно Павлаку, приближенное множество характеризуется двумя аппроксимациями – нижней и верхней. Нижняя аппроксимация определяет элементы, которые однозначно принадлежат данному множеству, верхняя аппроксимация определяет элементы, которые могут принадлежать данному множеству. Логические правила, получаемые с помощью верхней и нижней аппроксимаций, могут быть использованы для классификации новых элементов, т.е. для определения принадлежит или нет элемент множеству, что является классической задачей интеллектуального анализа данных.

В [3, 4] авторами был предложен алгебраический метод определения аппроксимаций приближенных множеств с использованием алгебры конечных предикатов [5], который использует только операции сравнения и булевы операции, что делает процесс поиска логических правил быстрым с вычислительной точки зрения. В [6] предлагается метод нахождения минимизированных наборов признаков, позволяющих адекватно описывать приближенное множество, однако оценки значимости того или иного признака не производится.

Цель статьи – предложить метод по определению значимости признаков объектов в базе данных на основании теории приближённых множеств.

Метод оценки значимости признаков с использованием теории приближенных множеств

Метод нахождения аппроксимаций приближенного множества для случая, когда признаки объектов принимают значения из некоторого конечного множества (небинарны), предложен в [4].

К достоинствам данного подхода следует отнести: реальный механизм нахождения аппроксимаций, легкий перерасчет аппроксимаций в случае изменения исходных данных, простота программной реализации.

Рассмотрим пример нахождения верхней и нижней аппроксимаций, предложенный в [4]:

Таблица 1

Верхняя и нижняя аппроксимации множества X

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

В табл. 1: a_1, ...,a_5 – объекты (записи) в базе данных; P_1, P_2, P_3 – признаки (атрибуты) объектов, принимающие значения из некоторого множества {0,1,2}; X – приближенное множество, характеризующееся верхней аппроксимацией I^* и нижней аппроксимацией I_*.

В общем случае аппроксимации I^* и I_* могут быть представлены следующим образом:

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

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

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 признака P.

Процедура получения логических правил из аппроксимаций достаточно проста и подробно описана в [4].

Для оценки значимости того или иного признака введем понятие мощности аппроксимации.

Под *мощностью аппроксимации* будем понимать число элементов, входящих в аппроксимацию (по аналогии с мощностью множества).

Для нашего примера:

Мощность верхней аппроксимации равна 3 (элементы a_2, a_4, a_5). Мощность нижней аппроксимации равна 1 (элемент a_4).

Аксиома 1. Чем меньше разность между мощностью нижней аппроксимации множества X и мощностью самого множества X, тем более точно нижняя аппроксимация описывает исходное множество.

Аксиома 2. Чем меньше разность между мощностью верхней аппроксимации множества X и мощностью самого множества X, тем меньше неопределённости в исходных данных.

Разница между верхней и нижней аппроксимациями ("граничный регион" [2]) служит своего рода показателем "качества" данных (с точки зрения имеющейся информации).

Очевидно, что при удалении того или иного признака P_1, P_2, …, P_k нижняя аппроксимация может только уменьшиться (в общем случае не увеличиться), а верхняя аппроксимация может увеличиться (в общем случае не уменьшиться). Действительно, при уменьшении количества признаков, количество имеющейся информации об элементах множества X уменьшается. Следовательно, количество неразличимых элементов может только увеличиваться (не уменьшаться), т.е. множество X становиться более приближенным (менее четким).

Определим *значимость признака* P_i как некоторую численную величину, показывающую насколько возрастает (в процентном соотношении) "приближенность" множества X при удалении признака P_i:

$$V(P_i) = \frac{((I_n^* - I_{*n}) - (I_o^* - I_{*o}))}{M(X)} * 100\%, \quad (3)$$

где I_o^* – старое значение верхней аппроксимации; I_{*o} – старое значение нижней аппроксимации; I_n^* – новое значение верхней аппроксимации; I_{*n} – новое значение нижней аппроксимации; M(X) – мощность множества X.

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

Рассмотрим описанный выше метод на примере (табл. 4).

Таблица 4

Фрагмент базы данных

Элементы / Признаки	a_1	a_2	a_3	a_4	a_5
P_1	0	0	0	0	0
P_2	0	1	1	0	0
P_3	0	1	0	1	1
P_4	1	0	1	1	1
X	1	0	0	1	0

Формулы для аппроксимаций будут выглядеть следующим образом:

$$I^* = (\overline{P}_1 \wedge \overline{P}_2 \wedge \overline{P}_3 \wedge P_4) \vee (\overline{P_1} \wedge \overline{P}_2 \wedge P_3 \wedge P_4); \quad (4)$$

$$I_* = (P_1 \vee \overline{P}_2 \vee \overline{P}_3 \vee P_4) \wedge P_1 \vee \overline{P}_2 \vee P_3 \vee \overline{P}_4) \wedge \wedge (P_1 \vee P_2 \vee \overline{P}_3 \vee \overline{P}_4). \quad (5)$$

Рассчитанные по формулам (4) и (5) аппроксимации – в табл. 5, значимость признака P_2 – в табл. 6.

Таблица 5

Верхняя и нижняя аппроксимации X

I^*	1	0	0	1	1
I_*	1	0	0	0	0

Таблица 6

Фрагмент базы данных (удален признак P_2)

Элементы / Признаки	a_1	a_2	a_3	a_4	a_5
P_1	0	0	0	0	0
P_3	0	1	0	1	1
P_4	1	0	1	1	1
X	1	0	0	1	0

Верхняя и нижняя аппроксимации:

$$I^* = (\overline{P}_1 \wedge \overline{P}_3 \wedge P_4) \vee (\overline{P_1} \wedge P_3 \wedge P_4); \quad (6)$$

$$I_* = (P_1 \vee \bar{P}_3 \vee P_4) \wedge (P_1 \vee P_3 \vee \bar{P}_4) \wedge (P_1 \vee \bar{P}_3 \vee \bar{P}_4). \quad (7)$$

Рассчитанные по формулам (6) и (7) аппроксимации – в табл. 7.

Таблица 7

Верхняя и нижняя аппроксимации X

I^*	1	0	0	1	1
I_*	0	0	0	0	0

Мощности верхних и нижних аппроксимаций:

$$I^*_o = 3,\ I_{*o} = 1,\ I^*_n = 3,\ I_{*n} = 0.$$

Значимость признака P_2:

$$V(P_2) = \frac{((3-0)-(3-1))}{2} * 100\% = 50\%.$$

Таким образом, можно сделать вывод, что признак P_2 является значимым. Следует отметить, что для данного примера не задавалось значение minDeterioration, однако возрастание "приближенности" множества X на 50% для столь малого набора данных, безусловно, стоит считать значимым.

Определим значимость признака P_1, исключив его и рассчитав новые аппроксимации приближенного множества X (табл. 8).

Таблица 8

Фрагмент базы данных (удален признак P_1)

Элементы / Признаки	a_1	a_2	a_3	a_4	a_5
P_2	0	1	1	0	0
P_3	0	1	0	1	1
P_4	1	0	1	1	1
X	1	0	0	1	0

Верхняя и нижняя аппроксимации рассчитываются по следующим формулам:

$$I^* = (\bar{P}_2 \wedge \bar{P}_3 \wedge P_4) \vee (\bar{P}_2 \wedge P_3 \wedge P_4), \quad (8)$$

$$I_* = (\bar{P}_2 \vee \bar{P}_3 \vee P_4) \wedge (\bar{P}_2 \vee P_3 \vee \bar{P}_4) \wedge (P_2 \vee \bar{P}_3 \vee \bar{P}_4). \quad (9)$$

Значения аппроксимаций – в табл. 9.

Таблица 9

Верхняя и нижняя аппроксимации X

I^*	1	0	0	1	1
I_*	1	0	0	0	0

Мощности аппроксимаций: $I^*_o = 3$, $I_{*o} = 1$, $I^*_n = 3$, $I_{*n} = 1$. Значимость признака P_1:

$$V(P_1) = \frac{((3-1)-(3-1))}{2} * 100\% = 0\%.$$

В данном случае можно сделать вывод, что признак P_1 является малозначимым.

Выводы

Предлагаемый метод определения значимости 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.

Безусловно, одним из дальнейших направлений исследования является определение порядка исключения признаков, а также возможности исключения нескольких признаков одновременно.

Список литературы

1. Едельстайн Г. Интеллектуальные средства анализа, интерпретации и представления данных в информационных хранилищах // Компьютеруик. – 1996. – № 16. – С. 32-35.

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

3. 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.

4. Ситников Д.Э., Романенко О.А., Титов С.В., Титова Е.В. Обобщенный логико-алгебраический метод нахождения аппроксимаций приближенных множеств и генерации на их основе логических правил // Збірник наукових праць ХУ ПС. – Х.: ХУПС, 2007. – № 5(15). – С. 115-119.

5. Шабанов-Кушнаренко Ю.П. Теория интеллекта. Математические средства. – Х.: Вища шк., 1984. – 143 с.

6. Ситников Д.Э., Романенко О.А., Титова Е.В., Титов С.В. Метод нахождения минимизированных наборов признаков в базах данных с использованием теории приближённых множеств // Системи обробки інформації. – Х.: ХУ ПС, 2007. – № 7(65). – С. 91-95.

Поступила в редколлегию 5.02.2008

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

МЕТОД ОЦІНКИ ЗНАЧУЩОСТІ ОЗНАК ОБ'ЄКТІВ В БАЗАХ ДАНИХ З ВИКОРИСТАННЯМ ТЕОРІЇ НАБЛИЖЕНИХ МНОЖИН

О.А. Романенко, Е.В. Тітова, А.А. Усань

Сучасні методи інтелектуального аналізу даних направлені на обробку великих інформаційних масивів з метою знаходження в них прихованих закономірностей. Однією з поширених форм представлення знайдених залежностей є логічні правила. Визначення значущості тієї або іншої ознаки дозволяє скоротити кількість аналізованих даних і як наслідок, скоротити кількість логічних умов в знайдених правилах. Пропонований метод визначення значущості ознаки дозволяє скоротити кількість атрибутів, що беруть участь в описі наближеної множини.

Ключові слова: логічні правила, потужність апроксимацій, значущість ознаки, нечіткі множини, кінцеві предикати.

METHOD OF ESTIMATION OF MEANINGFULNESS OF SIGNS OF OBJECTS IN DATABASES WITH THE USE OF THEORY OF CLOSE SETS

O.A. Romanenko, E.V. Titova, A.A. Usan'

The modern methods of intellectual analysis of data are directed on treatment of large informative arrays with the purpose of finding in them of the hidden conformities to the law. One of widespread forms of presentation of the found dependences are logical rules. Determination of meaningfulness of one or another sign allows to shorten the amount of analysable information and as a result, to shorten the amount of logical terms in the found rules. The offered method of determination of meaningfulness of sign allows to shorten the amount of attributes, participating in description of close great number.

Keywords: logical rules, power of approximation, meaningfulness of sign, fuzzy sets, eventual predicates.