УДК 681.3.00:007

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

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

ОБОБЩЁННЫЙ ЛОГИКО-АЛГЕБРАИЧЕСКИЙ МЕТОД НАХОЖДЕНИЯ АППРОКСИМАЦИЙ ПРИБЛИЖЕННЫХ МНОЖЕСТВ И ГЕНЕРАЦИИ НА ИХ ОСНОВЕ ЛОГИЧЕСКИХ ПРАВИЛ

Теория приближенных множеств позволяет оперировать с объектами, являющимися "неразличимыми" с точки зрения имеющейся информации, что делает ее незаменимым средством интеллектуального анализа данных. Основными характеристиками приближенного множества являются нижняя и верхняя аппроксимации, нахождение которых представляет собой интересную и не всегда простую задачу. Статья представляет собой обобщение алгебраического метода нахождения аппроксимаций приближенного множества, предложенного авторами ранее.

приближенное множество, неразличимые отношения, верхняя и нижняя аппроксимация, конечные предикаты, булевы операции, логические правила, характеристические функции

Введение

Постановка проблемы. Технологии интеллектуального анализа данных (Data Mining) в последние время привлекают большое внимание специалистов в области информационных технологий и разработчиков программного обеспечения. Эти технологии позволяют выявлять в огромных массивах данных определенные закономерности, таким образом обобщать накопленные данные до такого уровня представления информации, который может быть охарактеризован как получение знаний. Полученные закономерности представляются в виде моделей, выделенных кластеров, ассоциативных правил и т.д.

Методы и алгоритмы Data Mining (большинство из которых были разработаны в рамках теории искусственного интеллекта) часто направлены на принятие решений на основании неопределённых условий. Одной из классических задач интеллектуального анализа данных является задача классификации новых объектов на основании имеющейся информации. Причем информация, связанная с данными объектами, не позволяет однозначно отнести их к некоторому множеству. В подобных случаях эффективным решением является использование теории приближенных множеств, которая оперирует такими понятиями, как "неразличимые" элементы и "неразличимые" отношения.

Приближенное множество характеризуется двумя аппроксимациями – нижней, которая определяет элементы, которые однозначно принадлежат данному множеству, и верхней, которая определяет элементы, которые могут принадлежать данному множеству. Для нахождения аппроксимаций приближенного множества могут использоваться различные методы, в том числе и алгебраический, ранее предложенный авторами [1, 2].

Одной из проблем, возникающих при нахождении аппроксимаций приближенных множеств, является проблема обработки небинарных типов данных.

Анализ последних исследований и публикаций. Классическое определение нижней и верхней аппроксимаций предназначено для топологического описания свойств приближенных множеств [3 – 5]. Авторы указанных работ следуют классическому топологическому пути определения приближенных концепций, начиная с представления "неразличимых" отношений, определяют некоторые специальные свойства для них и после этого выводят понятия нижней и верхней аппроксимаций.

В [1, 2] авторами был предложен алгебраический метод определения аппроксимаций приближенных множеств с использованием алгебры конечных предикатов [6].

Новые определения аппроксимаций были сравнены с классическими и показано, что из предлагаемых определений можно получить классические путём наложения некоторых ограничений. Также в [1, 2] было отмечено, что использование бинарных выражений позволяет быстро рассчитывать аппроксимации для приближенного множества и генерировать на их основе логические правила, что делает данный подход эффективным с вычислительной точки зрения.

Цель статьи. Предложить обобщение метода поиска аппроксимаций приближенных множеств, что позволит оперировать не только с бинарными признаками, но и с дискретными небинарными.

Обобщенный алгебраичеЯкий подход к нахождению аппрокЯимаций приближенных множеЯтв

Алгебраический подход к определению аппроксимаций нечётких множеств был разработан авторами в [1, 2] и заключался 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, …, 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, …, P_k.

Также в [1, 2] была представлена концепция *языка аппроксимации* состоящего из унарных предикатов P_1, P_2, …, P_k и булевых операций: *конъюнкции* ($\wedge$), *дизъюнкции* ($\vee$) и *отрицания*. Было указано что в общем случае язык аппроксимации может включать другие типы предикатов и операций.

Под множеством Φ понималось множество всех возможных формул, построенных с помощью булевых операций конъюнкции, дизъюнкции и отрицания и применяемых к предикатам P_1, P_2, …, P_k. Под множеством Λ понималось множество всех предикатов, полученных путем расчета всех формул из Φ. Было отмечено, что различные формулы могут относиться к одному и тому же предикату. Таким образом, если предикат X(t) принадлежит Λ, это означает что X(t) *может быть выражен в терминах координат* и множество X по отношению к предикату X(t) может быть названо *точным* в терминах данных *координат*. Если предикат X(t) не принадлежит к Λ, этот предикат не может быть точно описан в терминах данных координат, а может быть описан только *приблизительно*.

Также в [1, 2] были представлены определения верхней и нижней аппроксимаций X(t) и описаны их свойства.

Необходимо отметить, что характеристические функции $P_1, P_2, \ldots, P_k$, которые используются для описания объектов в [1, 2] могут принимать только одно из двух значений: 1 и 0 (объект либо обладает некоторым свойством либо нет).

В общем случае характеристические функции могут принимать значение из некоего конечного множества значений. Например, для функции P_1 определяющей свойство "цвет" можно сказать следующее: $P_1(a_i)=1$ означает что объект a_i красный, $P_1(a_i)=2$ – оранжевый, $P_1(a_i)=3$ – жёлтый и т.д. В этом случае такая функция может быть описана с помощью конечного предиката: $P_k^w(a_i)=1$, который означает, что значение свойства P_k для объекта a_i есть w, иначе данный конечный предикат равен нулю [6].

Справедливы следующие свойства:

1. Для любого объекта a_i свойство P_k обязательно принимает хотя бы одно значение w из множества возможных значений: $w \in W$, $W=\{1,2,...,m\}$. Таким образом:

$$P_k^1(a_i) \vee P_k^2(a_i) \vee \ldots \vee P_k^m(a_i) \equiv 1\,.$$

2. Для любого объекта a_i свойство P_k не может принимать одновременно два значения из множества возможных значений:

$$P_k^w(a_i) \,\&\, P_k^r(a_i) \equiv 0 \quad (w \in W,\ r \in W,\ w \neq r).$$

Принимая во внимание вышеперечисленное, можно определить язык аппроксимаций как язык, состоящий из множества дискретных характеристических функций и множества операций, включая булевы операции и операторы сравнения, которые дают 1 если два дискретных значения равны и 0 если они различны.

СвойЯтва обобщенных аппрокЯимаций

В [1, 2] доказаны свойства верхней и нижней аппроксимаций в случае, когда функции описывающие свойства объекта бинарные. Аналогичные свойства остаются действительными и в случае, когда характеристические функции принимают значение из конечного множества:

1. Множество нижних аппроксимаций не пустое для любых X(t).

Это следует из того, что предикат $0=P_k^w \,\&\, P_k^r, (w \neq r)$ всегда является нижней аппроксимацией для X(t).

2. Множество верхних аппроксимаций не пустое для любых X(t).

Это следует из предиката $1=P_k^1 \vee P_k^2 \vee \ldots \vee P_k^m$, который всегда является верхней аппроксимацией для X(t).

3. Для любого предиката X(t) не может существовать более чем одна точная нижняя аппроксимация.

4. Для любого предиката X(t) существует как минимум одна точная нижняя аппроксимация.

5. Для любого предиката X(t) существует единственная точная нижняя аппроксимация.

6. Для любого предиката X(t) не может существовать более чем одна точная верхняя аппроксимация.

7. Для любого предиката X(t) существует как минимум одна точная верхняя аппроксимация.

8. Для любого предиката X(t) существует единственная точная аппроксимация. Это следует из свойств 6 и 7.

Доказательства свойств 3 – 7 аналогичны приведенному в [1].

ПоиЯк точных аппрокЯимаций Я помощью диЯкретных характериЯтичеЯких функций

Обсудим далее метод поиска точных верхних и нижних аппроксимаций для случая, когда свойства объекта представлены небинарными значениями. Поиск всех возможных формул предикатов для расчёта аппроксимаций требует значительных затрат времени.

Несмотря на это, существует путь получения аппроксимаций, который позволяет быстро создавать необходимые формулы.

Рассмотрим следующую таблицу (табл. 1).

Таблица 1

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

	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

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

Предположим, что предикат X должен быть описан в терминах координат $P_1, P_2, \ldots, P_k$, которые являются случайными дискретными характеристическими функциями.

Найдём точную верхнюю аппроксимацию для X. Для этого рассмотрим столбцы таблицы, которые содержат 1 для предиката X и напишем соответствующий предикат в дизъюнктивной нормальной форме.

Приведем простейший пример.

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}. в соответствии с табл. 2.

Таблица 2

Пример координат и предиката 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

Для этого примера получаем следующую формулу расчета верхней аппроксимации:

$$I^* = (P_1^0 \wedge P_2^2 \wedge P_3^2) \vee (P_1^0 \wedge P_2^0 \wedge P_3^1) \qquad (4)$$

Рассмотрим теперь столбцы которые содержат 0 для предиката X и составим соответствующее выражение для нижней аппроксимации. Для данного примера:

$$I_* = (\overline{P_1^1} \vee \overline{P_2^0} \vee \overline{P_3^0}) \wedge (\overline{P_1^2} \vee \overline{P_2^1} \vee \overline{P_3^0}) \wedge$$
$$\overline{(\wedge P_1^0} \vee \overline{P_2^2} \vee \overline{P_3^2}) \qquad (5)$$

Формулы (4) и (5) дают следующий результат:

Таблица 3

Верхняя и нижняя аппроксимации предиката 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 общем случае предикаты I^* и I_* могут быть представлены следующим образом:

$$I_* = (\lambda_1 \wedge P_1^{\delta_{11}} \wedge P_2^{\delta_{21}} \wedge \ldots \wedge P_k^{\delta_{k1}}) \vee (\lambda_2 \wedge P_1^{\delta_{12}} \wedge P_2^{\delta_{22}} \wedge$$
$$\ldots \wedge P_k^{\delta_{k2}}) \vee \ldots \vee (\lambda_n \wedge P_1^{\delta_{1n}} \wedge P_2^{\delta_{2n}} \wedge \ldots \wedge P_k^{\delta_{nn}} \ldots), \qquad (6)$$

$$I_* = (\lambda_1 \vee \overline{P_1^{\delta_{11}}} \vee \overline{P_2^{\delta_{21}}} \vee \ldots \vee \overline{P_k^{\delta_{k1}}}) \wedge (\lambda_2 \vee \overline{P_1^{\delta_{12}}} \vee \overline{P_2^{\delta_{22}}} \vee \ldots$$
$$\vee \overline{P_k^{\delta_{k2}}}) \wedge \ldots \wedge (\lambda_n \vee \overline{P_1^{\delta_{1n}}} \vee \overline{P_2^{\delta_{2n}}} \vee \ldots \vee \overline{P_k^{\delta_{nn}}}), \qquad (7)$$

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 предиката P.

Покажем, что предикаты I^* и I_* являются точными верхней и нижней аппроксимациями. Очевидно, что предикат I^* является верхней аппроксимацией для предиката X. Если удалить из формулы (6) любую конъюнкцию где $\lambda_i=1$, результирующая формула будет уже не верхней аппроксимацией, так как предикат I^* будет иметь 0 в колонке, в которой X равен 1. (Мы предполагаем, что конъюнкции идентичные убранной также будут убраны).

Например, если убрать конъюнкцию $(P_1^0 \wedge P_2^2 \wedge P_3^2)$ из формулы (4), тогда $I^*(a_2)$ станет равным 0 в то время как $X(a_2)=1$. Это означает, что аппроксимация I^* не может быть улучшена и, таким образом, I^* является точной верхней аппроксимацией. Предикат I_* является нижней аппроксимацией для X. Если удалить из формулы (7) любую дизъюнкцию где $\lambda_i=0$, результирующая формула не будет нижней аппроксимацией так как предикат I_* будет принимать значение 1 в столбце, в котором X равен 0. (Предполагается, что будут убраны все повторяющиеся дизъюнкции, аналогичные уже убранной). Например, если убрать дизъюнкцию $(\overline{P_1^1} \vee \overline{P_2^0} \vee \overline{P_3^0})$ из формулы (5), то $I^*(a_1)=1$, а при этом $X(a_1)=0$.

Это означает, что аппроксимация не может быть улучшена и, таким образом, I_* является точной нижней аппроксимацией.

Лог, чеЯк, е прав, ла, оЯнованные на аппрокЯ, мац, ях

Рассмотрим пример генерации логических правил с помощью формул (4) и (5).

Используя традиционные концепции теории приближенных множеств, можно сказать, что правила основанные на точной верхней аппроксимации могут существовать в рамках набора данных, а правила, основанные на точной нижней аппроксимации, должны присутствовать в наборе данных.

Преобразуя выражение в правой части формулы (5), получим:

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

Теперь можно сформировать следующие точные правила:

A. Элемент принадлежит множеству X если

a) свойство P_1 не равно 1 или P_2 и P_3 не равны 0

И

b) свойство P_1 не равно 2 или свойство P_2 не равно 1 или свойство P_3 не равно 0

И

c) свойство P_1 не равно 0 или свойство P_2 не равно 2 или свойство P_3 не равно 2.

Это правило говорит о том, что если все из условий a), b) и c) выполняются, то элемент принадлежит множеству X.

Так как каждая характеристическая функция P_k (в данном случае P^1, P^2, P^3) принимает как минимум одно значение из множества возможных значений, то $\overline{P_k^i} = P_k^1 \vee P_k^2 \vee \ldots \vee P_k^{i-1} \vee P_k^{i+1} \vee \ldots \vee P_k^n$.

Используя это свойство можно преобразовать полученное правило и получить следующую формулу:

$$I_* = P_3^1 \vee (P_1^1 \wedge P_3^2) \vee (P_1^1 \wedge \overline{P_2^0}) \vee (P_1^2 \wedge P_3^2) \vee (P_1^2 \wedge \overline{P_2^1})$$
$$\vee (P_1^0 \wedge \overline{P_2^2}) \vee (P_1^0 \wedge P_3^0) \vee (P_3^2 \wedge P_2^2) \vee (P_2^2 \wedge P_3^0).$$

С помощью алгебраических преобразований появляется возможность избежать использования отрицаний и получить следующие точные правила из приведенной формулы:

B. Элемент принадлежит к множеству X если

a) свойство P_3 равно 1

ИЛИ

b) свойство P_1 равно 1 и P_3 равно 2

ИЛИ

c) свойство P_1 равно 1 и P_2 равно 1 или 2

ИЛИ

d) свойство P_1 равно 2 и P_3 равно 2

ИЛИ

e) свойство P_1 равно 2 и P_2 равно 0 или 2

ИЛИ

f) свойство P_1 равно 0 и P_2 равно 0 или 1

ИЛИ

g) свойство P_1 равно 0 и P_3 равно 0

ИЛИ

h) свойство P_2 равно 0 или 1 and P_3 равно 2

ИЛИ

i) свойство P_2 равно 2 и P_3 равно 0.

Эти правила показывают что, если одно из условий a), b), c), d), e), f), g), h) и i) выполняется, элемент принадлежит к множеству X.

Формула (4) для точных верхних аппроксимаций позволяет сформировать следующие приближённые правила:

C. Элемент может принадлежать к множеству X *если*

свойство P_1 принимает значение 0 И свойство P_2 принимает значение 2 И свойство P_3 принимает значение 3 для этого элемента.

D. Элемент может принадлежать множеству X *если*

свойство P_1 принимает значение 3 И свойство P_2 принимает значение 0 И свойство P_3 принимает значение 0 для этого элемента.

Логические правила позволяют давать ответ на вопрос, принадлежит или нет элемент заданному множеству на основе имеющейся информации о свойствах элемента. Подобные расчёты, основанные на применении логических операций, являются очень быстрыми и позволяют получать результирующие правила даже для большого числа элементов и характеристических функций.

Выводыя

Данная статья представляет обобщение алгебраического метода нахождения верхней и нижней аппроксимаций приближенного множества, предложенного авторами ранее. Обобщение метода заключается в рассмотрении случая, когда свойства объектов (записей в базе данных) представлены не только бинарными значениями. В статье показано, как алгебраический подход применяется в тех случаях, когда свойства объектов принимают случайное значение из некоторого конечного множества. Логические правила полученные 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. Предлагаемый метод использует только операции сравнения и булевы операции, что делает процесс поиска логических правил быстрым с вычислительной точки зрения. Необходимо отметить, что преобразование формулы конечных предикатов для точных аппроксимаций может быть осуществлено различными путями, поэтому, в дальнейшем было бы интересно иметь возможность искать наиболее простые и информативные результирующие правила.

Также представляется интересным выделение наиболее значимых признаков, подлежащих рассмотрению при генерации логических правил.

Сп, Яокяп, тературыя

1. *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. – № 7 (63). – С.91-95.*

2. *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.*

3. *Pawlak Z. Rough set // International Journal of Computer and Information Sciences. – 1982. – №11. – P. 341-356.*

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

5. *Pawlak Z. Rough sets, rough relations and rough functions // Fundamenta Informaticae. – 1996. – Vol. 27, № 2/3. – P. 103-108.*

6. *Шабанов-Кушнаренко Ю. П. Теория интеллекта. Математические средства. – Х.: Вища шк. Изд-во при Харьк. ун-те, 1984. – 143 с.*

Поступила в редколлегию 2.10.2007

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