УДК 51 : 004.7

Д.Э. Ситников[1], П.Э. Ситникова[2], А.И. Коваленко[1]

[1] *Харьковская государственная академия культуры, Харьков*
[2] *Харьковский гуманитарный университет «Народная украинская академия», Харьков*

АЛГЕБРАИЧЕСКИЙ МЕТОД ИДЕНТИФИКАЦИИ ПРИБЛИЖЕННЫХ МНОЖЕСТВ В ИНФОРМАЦИОННЫХ СИСТЕМАХ

В данной статье рассматривается информационная система с отношением неразличимости. В рамках такой системы анализируются понятия приближенного множества, а также его верхней и нижней аппроксимаций. Рассмотрены классический топологический подход, предложенный З. Павлаком и новый алгебраический метод представления и обработки данных, основанный на описании объектов и их свойств с помощью системы унарных предикатов. В результате использования такого метода аппроксимации приближенные множества представляются в виде логических формул алгебры конечных предикатов.

***Ключевые слова**: Rough sets theory, приближенные множества, верхняя аппроксимация, нижняя аппроксимация, отношение неразличимости, информационная система.*

Введение

Способ представления знаний в информационной системе играет, как правило, большую роль. Наиболее известными способами представления знаний в системах индуктивного формирования понятий являются: продукционные правила, решающие деревья, исчисление предикатов и семантические сети.

При извлечении и обобщении знаний, хранящихся в реальных информационных массивах, возникают следующие основные проблемы:

1. Данные являются разнородными (количественными, качественными, структурными).

2. Реальные базы данных, как правило, велики, а потому алгоритмы экспоненциальной сложности для извлечения знаний из баз данных могут оказаться неприемлемыми.

3. Информация, содержащаяся в реальных массивах данных, может быть неполна, избыточна, искажена, противоречива, а также некоторые значения ряда атрибутов могут вовсе отсутствовать. Поэтому для построения классификационных правил следует использовать только существенные атрибуты.

В настоящее время для извлечения знаний из баз данных (Data Mining) теория приближенных множеств rough sets theory все чаще используется как теоретическая база и набор практических методов.

Приближенные множества – это множества с неопределенными границами, т.е. множества, которые не могут быть точно описаны доступным набором признаков.

Теория приближенных множеств была предложена Здиславом Павлаком в 1982 году и явилась новым математическим инструментом работы с неполной информацией. Важнейшими понятиями данной теории являются так называемые верхняя и нижняя аппроксимации приближенного множества, позволяющие оценить возможность или необходимость принадлежности элемента множеству с «размытыми» границами [1, 2].

Простая, но мощная концепция приближенных множеств стала базовой как в теоретических исследованиях – логике, алгебре, топологии, так и в прикладных – задачах искусственного интеллекта, приближенных рассуждениях, интеллектуальном анализе данных, теории принятия решений, обработке изображений и распознавании образов.

Концепция «приближенного множества» имеет дело с «несовершенством данных» (imperfection), относящимся к «гранулярности» информации (granularity). Эта концепция по своей природе является топологической и дополняет другие известные подходы, используемые для работы с неполной информацией, такие как нечеткие множества (fuzzy sets), методы Байеса (Bayesian reasoning), нейронные сети (neural networks), эволюционные алгоритмы (evolutionary algorithms), статистические методы анализа данных.

Цель данной работы: разработка алгебраического подхода к описанию приближенных множеств в информационной системе и его рассмотрение в контексте классических определений верхней и нижней аппроксимаций, данных З. Павлаком.

Некоторые математические предпосылки описания теории приближенных множеств

Говоря о математических предпосылках данной теории, приведем следующие определения.

Бинарное отношение множества A ко множеству B – это подмножество $A \times B$. Пусть R являет-

 © Д.Э. Ситников, П.Э. Ситникова, А.И. Коваленко

ся отношением, тогда мы можем записать $(x,y)\in R$ или xRy.

Подмножество $A\times A$ является бинарным отношением на множестве A. В частности $A\times A$ является универсальным отношением на A.

Пусть A – множество и R – отношение на этом множестве. Тогда подмножество элементов из A, состоящих в отношении R, можно записать как $R[A]=\{y,\forall x\in A,xRy\}$.

Если R – отношение A к B, то отношение, обратное R, записанное как R^{-1}, это отношение B к A, такое что $yR^{-1}x$ если и только если yRx.

Свойства отношения R на множестве A:

– отношение рефлексивно, если и только если xRx;

– нерефлексивно, если $\exists x\in A$, что $x\overline{R}x$;

– тождественно, если оно рефлексивно и xRy для $x,y\in A\to x=y$;

– симметрично, если xRy для $x,y\in A\to yRx$;

– несимметрично, если $\exists x,y$ такие, что xRy и $x\overline{R}y$;

– антисимметрично, если xRy и yRx для $x,y\in A\to x=y$;

– транзитивно, если xRy и yRz для $x,y,z\in A\to xRz$.

Рефлексивное, антисимметричное и транзитивное отношение на множестве является отношением частичного порядка на этом множестве. Если R – отношение частичного порядка на A, то упорядоченная пара (A,R) – частично упорядоченное множество.

Отношение на множестве A является отношением эквивалентности, если оно рефлексивно, симметрично и транзитивно.

Если R – отношение эквивалентности на множестве A и элемент $a\in A$. Тогда подмножество элементов a из множества A, связанных отношением R, $R[\{a\}]$ называется классом эквивалентности относительно a, что можно также записать $[a]$.

Концепция аппроксимации множеств

Базовой идеей теории приближенных множеств является концепция аппроксимации множеств. Рассмотрим эту концепцию более подробно.

База знаний описывается как упорядоченное множество $A=(U,R)$, где U является универсальным множеством (непустым множеством объектов), а R – отношением эквивалентности на этом множестве, называемое отношением неразличимости. Если $x,y\in U$ и xRy, то x и y неразличимы на множестве A. Пусть элементы из универсума разбиты на классы эквивалентности, определяемые как $[x]=\{y\in U,xRy\}$. Каждый класс эквивалентности, порожденный R, называется элементарным множеством из A.

Пусть X является подмножеством A. В случае, если множество X является объединением некоторых классов эквивалентности, то данное множество называется R-точным. В противном случае, множество X является R-неточным (R-грубым) и характеризуется двумя множествами – верхней и нижней аппроксимацией.

Нижняя аппроксимация описывается формулой:

$$A_*(X)=\{x\in U,[x]\subseteq X\},$$

т.е. нижнюю аппроксимацию составляют элементы, которые **точно** принадлежат X.

Верхняя аппроксимация описывается формулой:

$$A^*(x)=\{x\in U,[x]\cap X\neq\varnothing\}$$

т.е. верхнюю аппроксимацию составляют элементы, которые **возможно** принадлежат X.

Граничной областью (boundary region) множества X называется разность между верхней и нижней аппроксимацией

$$B(X)=A^*(X)-A_*(X).$$

Т.е. в граничную область входят элементы множества X, принадлежащие верхней аппроксимации и не принадлежащие нижней аппроксимации.

Например, пусть дано множество элементов $U=\{a_1,...,a_{10}\}$. Пусть дано разбиение данного множества на следующие классы эквивалентности: $\{a_1,a_2\},\{a_3,a_4,a_5\}\{a_6,a_7\}\{a_8\}\{a_9\}\{a_{10}\}$.

Целевое множество $X_1=\{a_1,a_2,a_6,a_7,a_9\}$ характеризуется верхней аппроксимацией $A^*(X_1)=\{a_1,a_2,a_6,a_7,a_9\}$ и нижней аппроксимацией $A_*=(X_1)=\{a_1,a_2,a_6,a_7,a_9\}$, а граничная область $B(X_1)=\varnothing$. Таким образом, очевидно, что данное множество является точным.

Рассмотрим другое целевое множество: $X_2=\{a_1,a_2,a_4,a_6\}$. Для него верхняя аппроксимация $A^*(X_2)=\{a_1,a_2,a_3,a_4,a_5,a_6,a_7\}$ и нижней аппроксимацией $A_*(X_1)=\{a_1,a_2\}$, а граничная область $B(X_2)=\{a_3,a_4,a_5,a_6,a_7\}$ и данное множество является приближенным.

Отношение неразличимости и аппроксимации в информационной системе

В [3] З. Павлак вводит определение информационной системы как пары $S=(U,A)$, где $U=\{x_1,x_2,...,x_n\}$ – непустое конечное множество

объектов, называемое обучающим множеством или универсумом, $A = \{a_1, a_2, ..., a_k\}$ – непустое конечное множество атрибутов. Для каждого атрибута известно множество V_a, которое называется множеством значений атрибута a.

С каждым подмножеством атрибутов $B \subseteq A$ может быть связано бинарное отношение IND(B), называемое отношением неразличимости:

$$IND(B) = \{(x, y) \in U \times U \mid \forall a \in B(a(x) = a(y))\}.$$

Если $(x, y) \in IND(B)$, то объекты x и y являются неразличимыми по отношению ко всем атрибутам множества B.

Рассмотрим на примере неразличимость объектов. Пусть даны сведения о некоторых пациентах с соответствующими значениями признаков: головная боль, температура, наличие гриппа.

Таблица 1

Фрагмент медицинской базы данных

	Головная боль	Температура	Грипп
1.	Да	Высокая	Да
2.	Да	Нормальная	Нет
3.	Нет	Высокая	Нет
4.	Да	Очень высокая	Да
5.	Да	Высокая	Нет
6.	Да	Очень высокая	Нет

Разбиение множества U в соответствии со значениями атрибута «Головная боль» имеет вид:

$$\{\{1, 2, 4,5,6\}, \{3\}\}.$$

Разбиение множества U в соответствии со значениями атрибута «Температура» имеет вид:

$$\{\{1, 3,5\}, \{2\}, \{4, 6\}\}.$$

Разбиение множества U в соответствии со значениями атрибутов «Головная боль» и «Температура» имеет вид:

$$\{\{1, 5\}, \{2\}, \{3\}, \{4,6\}\}.$$

Разбиение множества U в соответствии со значениями атрибутов «Грипп» имеет вид:

$$\{\{1, 4\}, \{2,3,5,6\}\}.$$

Элементы 1 и 5 являются *неразличимыми* с точки зрения теории множеств.

В [4] был описан алгебраический подход к определению приближенных множеств. В качестве характеристических функций для некоторых свойств объектов универсума предложено ввести множество унарных предикатов $P_1(t)$, $P_2(t)$, $..., P_k(t)$, которые принимают одно из значений $\{0,1\}$. При этом объект a_i имеет свойство P_j тогда и только тогда, когда $P_j(a_i) = 1$. Задачей является описать некоторое множество $X \subseteq U$ в терминах координат, а значит описать предикат $X(t)$, равный 1 тогда и только тогда, когда $t \in X$, в терминах предикатов $P_1(t), P_2(t), ..., P_k(t)$.

Из классического определения аппроксимации следует, что каждому классу эквивалентности может быть поставлен в соответствие предикат, принимающий значение 1, если элемент множества принадлежит этому классу, и значение 0 для всех других элементов множества U. Таким образом, имеется множество предикатов $P_1(t), P_2(t), ..., P_k(t)$, удовлетворяющих условиям:

$$\exists t \;\; P_i(t) \;\; i = 1, 2, ..., k;$$

$$\forall t \in U \quad P_1(t) \vee P_2(t) \vee ... \vee P_k(t);$$

$$\forall t \in U \quad (P_i(t) \wedge P_j(t)); \;\; i, j = 1, 2, ..., k; i \neq j.$$

Таким образом, нижней аппроксимацией множества X является объединение классов эквивалентности подмножеств X, т.е. дизъюнкция всех предикатов $P_i(t)$, для которых $\forall t \in U \; P_i(t) \rightarrow X(t)$. Верхней аппроксимацией является объединение всех классов эквивалентности, которые имеют непустое пересечение с X, т.е. дизъюнкция всех предикатов $P_i(t)$, для которых $\exists t \in U \; P_i(t) \wedge X(t)$.

В [5] предлагается метод, позволяющий получать аппроксимации предиката за один просмотр таблицы данных. Это особенно актуально для решения проблем Data Mining, когда приходится оперировать с большими объемами данных.

Метод заключается в том, что при нахождении точной верхней аппроксимации множества X рассматриваются те строки таблицы значений, на которых X обращается в 1 и записывается соответствующая дизъюнктивная нормальная форма, а для нахождения точной нижней аппроксимации этого множества рассматриваются, наоборот, только те строки, на которых предикат X обращается в 0 и записывается соответствующая конъюнктивная нормальная форма.

Рассмотрим пример, приведенный выше. Преобразуем таблицу значений таким образом, чтобы атрибуты принимали значение из множества $\{0,1\}$. Атрибут «Головная боль» принимает значения 1 (есть) или 0 (нет), обозначим его P_1 Для атрибута «Температура» сделаем такое преобразование:

Температура	P_2	P_3	P_4
Нормальная	1	0	0
Высокая	0	1	0
Очень высокая	0	0	1

Тогда из табл. 1 получим табл. 2.

Используя предложенный метод, запишем дизъюнктивную нормальную форму для строк 1 и 4 таблицы 2, где предикат X обращается в 1:

Таблица 2

Таблица значений,
представленная в координатах P_i

	P_1	P_2	P_3	P_4	X
1.	1	0	1	0	1
2.	1	1	0	0	0
3.	0	0	1	0	0
4.	1	0	0	1	1
5.	1	0	1	0	0
6.	1	0	0	1	0

$$A^* = (P_1 \wedge \overline{P_2} \wedge P_3 \wedge \overline{P_4}) \vee (P_1 \wedge \overline{P_2} \wedge \overline{P_3} \wedge P_4)$$

Теперь запишем конъюнктивную нормальную форму для строк, где предикат X обращается в 0 – 2, 3, 5, 6:

$$A_* = (\overline{P_1} \vee \overline{P_2} \vee P_3 \vee P_4) \wedge (P_1 \vee P_2 \vee \overline{P_3} \vee P_4) \wedge$$
$$\wedge (\overline{P_1} \vee P_2 \vee \overline{P_3} \vee P_4) \wedge (\overline{P_1} \vee P_2 \vee P_3 \vee \overline{P_4}).$$

Из приведенных формул получаем результат для верхней и нижней аппроксимаций:

Таблица 3

Верхняя и нижняя аппроксимация предиката X

	P_1	P_2	P_3	P_4	X	A^*	A_*
1.	1	0	1	0	1	1	0
2.	1	1	0	0	0	0	0
3.	0	0	1	0	0	0	0
4.	1	0	0	1	1	1	1
5.	1	0	1	0	0	1	0
6.	1	0	0	1	0	1	1

Очевидно, в результате нахождения верхней и нижней аппроксимации можно определить *неразличимые* элементы: 1 и 5, что совпадает с рассуждениями, приведенными выше.

Выводы

Данная работа является продолжением исследований в области теории приближенных множеств. Рассмотрен алгебраический подход к нахождению аппроксимаций приближенного множества, и дана его интерпретация с точки зрения классического топологического подхода З. Павлака. Алгебраический подход представляется удобным в тех случаях, когда задачу можно свести к представлению объектов и их свойств с помощью унарных предикатов.

Список литературы

1. *Pawlak Z. Vaguenes and uncertainty: a Rough set perspective / Z. Pawlak // Computational Intelligence. – May 1995. – Volume 11 (Issue 2). – P. 227-232.*
2. *Pawlak Z. Rough set approach to knowledge-based decision support / Z. Pawlak // European Journal of Operational Research. – 99 (1997). – P. 420-432.*
3. *Pawlak Z. Rough set approach to knowledge-based decision support / Z. Pawlak // Proc. of the 14 European Conference on Operational Research Jerusalem. – Israel, July 1995.*
4. *Sitnikov D. An algebraic approach to defining rough set approximations and generating logic rules / D. Sitnikov, O. Ryabov; A. Zanasi, N. Ebecken, C. Brebbia (eds) // Data Mining V. – Malaga, Spain, 2004. – P. 179-188.*
5. *Ситников Д.Э. Метод нахождения аппроксимаций приближенных множеств и построения логических правил на основе алгебры конечных предикатов / Д.Э. Ситников, О.А. Рябов, Е.В. Титова, О.А. Романенко // Системи обробки інформації. – Х.: ХУ ПС, 2007. – Вип. 4(62). – С. 144-149.*

Поступила в редколлегию 25.10.2013

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

АЛГЕБРАЇЧНИЙ МЕТОД ІДЕНТИФІКАЦІЇ НАБЛИЖЕНИХ МНОЖИН В ІНФОРМАЦІЙНИХ СИСТЕМАХ

Д.Е. Ситніков, П.Е. Ситнікова, А.І. Коваленко

У даній статті розглядається інформаційна система з відношенням нерозрізненості . В рамках такої системи аналізуються поняття наближеної множини , а також її верхньої та нижньої апроксимацій . Розглянуто класичний топологічний підхід , запропонований З. Павлаком і новий алгебраїчний метод подання та обробки даних, заснований на описі об'єктів і їх властивостей за допомогою системи унарних предикатів . У результаті використання такого методу апроксимації наближені множини представляються у вигляді логічних формул алгебри кінцевих предикатів.

***Ключові слова**: Rough sets theory, наближені множини, верхня апроксимація, нижня апроксимації, ставлення непомітності, інформаційна система.*

AN ALGEBRAIC APPROACH TO THE IDENTIFICATION OF ROUGH SETS IN INFORMATION SYSTEMS

D.E. Sitnikov, P.E. Sitnikova, A.I. Kovalenko

In this paper, an information system with the indiscernibility relation has been considered. In the framework of this system the concepts of a rough set and its upper and lower approximations have been analyzed. The authors have considered a classical topological approach suggested by Z. Pawlak and a new algebraic method for data representation and processing based on the description of objects and their properties with the help of a system of unary predicates. Because of using such a method approximations of a rough set can be represented in the form of finite predicates algebra formulae.

***Keywords:** Rough sets theory, Rough Sets, upper approximation, lower approximation, indiscernibility relation, information system.*