УДК 004.7

ОБОБЩЕНИЕ ПОНЯТИЙ ПО ПРИЗНАКАМ: МЕТОДЫ ТЕОРИИ ПРИБЛИЗИТЕЛЬНЫХ МНОЖЕСТВ

Марьин С.А., Ситников Д.Э., Коваленко А.И.
Харьковская государственная академия культуры

В статье рассматривается ряд вопросов связанных с индуктивным обобщением понятий в современных интеллектуальных и в экспертных системах. Достаточно подробно исследован логический подход к определению списка признаков, по которым можно построить такое обобщение. Для этого исследуется функция различимости, матрица различимости и срез в рамках теории приблизительных множеств. Рассмотрен пример, иллюстрирующий описанный метод.
Ключевые слова: теория приблизительных множеств, обобщение понятий, экспертная система, информационная система, отношение неразличимости, пространство аппроксимаций, пространство приближений, эквивалентность, нижнее приближение для множества, верхнее приближение для множества, обобщенное решение.

Постановка проблемы. При построении современных экспертных систем, систем моделирующих мышление человека, часто приходиться решать задачи обобщения понятий. Обычно под термином обобщение понимается «процесс получения и обработки знаний, дающих полные объяснения существующим фактам, а также способность к классификации и предсказанию новых фактов» В данной статье рассматривается процесс получения обобщенных знаний на приблизительных множествах З. Павлака [5].

Анализ последних исследований и публикаций. Задача обобщения, в общем виде, была сформулирована Михальским. Процесс обобщения – это процесс, при котором рассматривается одно обобщенное понятие вместо того, чтобы рассматривать некоторое множество объектов или единичный конкретный объект [8]. По утверждению Михальского, все многообразие множества моделей обобщения разбиваются на два класса: первый – модели обобщения по выборкам, второй – модели обобщения по данным. Процесс обобщения базируется на выявлении наиболее характерных частей описания исходных объектов, при условии, что каждый из них представляется некоторым множеством признаков. Если исследуемый объект точно попадает в объем понятия, тогда для данного понятия он называется положительным. Если не попадает – отрицательным объектом.

К теперешнему моменту известно несколько подходов, в рамках которых, можно провести подобное обобщение. К таким подходам относят: индукцию решающих деревьев [7], продукционные правила [6; 9], различного типа нейронные сети, теорию приближенных множеств [3; 4] и т.д.

Выделение нерешенных ранее частей общей проблемы. Реальные данные имеют следующий ряд характеристик, которые включают неполноту, большой объем исходных данных, их противоречивость. Поэтому, перечисленные выше подходы, не гарантируют получения хороших результатов, и это причина дальнейшего совершенствования методов и подходов к построению правил обобщения. Среди всех перечисленных подходов была выбрана теория приближенных множеств.

Формулирование целей статьи. Целью данной статьи является создание алгоритма и усовершенствование математического аппарата обобщения на базе теории приближенных множеств З. Павлака.

Изложение основного материала исследования. Понятия, непосредственно связанные с информационными системами, были даны в статьях З. Павлака [1; 3]. Итак, информационная система – это два значения $S = (U, A)$, где $U = \{x_1,x_2,...,x_n\}$ – непустое конечное множество объектов, которое представляет собой обучающее множество или универсум, а $A = \{a_1,a_2,...,a_k\}$ – непустое конечное множество атрибутов. Решающая система (или решающая таблица) представляет собой информационную систему, имеющую следующий вид: $S = (U, A \cup \{d\})$, где $d \notin A$ – это выделенный атрибут или решение, или решающий атрибут, A – условные атрибуты.

Предположим, что имеем обучающее множество U. На этом множестве введем отношение неразличимости или эквивалентности на U: $IND(A) \subseteq U \times U$IND(A). Для двух упорядоченных значений $AS = (U,IND(U))$ дадим название – пространство аппроксимации или пространство приближений. Отметим, что если $(x,y) \in IND(A)$, тогда x и y неразличимы по значениям атрибутов из A в AS. Дадим название классам эквивалентности по отношению к IND. Тогда, класс эквивалентности – это элементарное множество, или атом в AS. Пусть, множество классов эквивалентности будет обозначено в виде множества $\{X_1^A, X_2^A, ..., X_m^A\}$.

Составное множество AS – это конечное объединение одного элементарного множества или нескольких элементарных множеств в AS. Несколько составных множеств в AS (семейство) обозначим как $Def\ (AS)$.

Пусть, $X \subseteq Y$. Самое большое составное множество в AS, содержащееся в X, будем называть нижним приближением множества X. Самое маленькое множество в AS, содержащееся в X, будем называть верхним приближением X. Верхнее приближение множества X обозначим как $\overline{AX}$, нижнее приближение множества X обозначим как $\underline{AX}$.

Для определения нижнего приближения для множества X необходимо использовать следующую формулу: $\underline{A}X = \bigcup_{X_i^A \subseteq X} X_i^A$. Тогда, нижнее приближение множества X – это объединение классов эквивалентности отношения неразличимости, которые входят в множество X.

Для определения верхнего приближения для множества X необходимо использовать следую-

щую формулу: $AX = \bigcup_{X_i^A \cap X = \varnothing} X_i^A$. Верхнее приближение множества X – это объединение классов эквивалентности, часть объектов которых относится к X.

Определим формулу для граничной или недостоверной области множества X. Она будет иметь следующий вид: $BN_A(X) = \underline{A}X \setminus \overline{A}X$. Эта формула определяет область множества X в AS, которая состоит из объектов, которые мы не можем уверенно отнести к X.

Множество $U \setminus \overline{A}X$ включает только отрицательные объекты для X. Множество $POS_A(D) = \underline{A}C_1 \cup ... \cup \underline{A}C_{r(d)}$ состоит исключительно из объектов, которые относятся к положительной области решающей системы S. Эти объекты точно относятся к одному из классов решения.

Пара $< \underline{A}X, \overline{A}X >$ является образующей для приближенного множества X. Совпадение двух приближений верхнего и нижнего образуют обычное множество X.

Классификация объектов может быть проведена на основе использования приблизительных множеств следующим образом.

Выберем отрицательную область множества. Пусть $NEG_A(X) = U \setminus \overline{A}(X)$. В таком случае можно сформировать решающие правила:

$$Описание\ POS_A(X) \to X,$$
$$Описание\ NEG_A(X) \to X,$$
$$Описание\ BN_A(X) \to \text{возможно}\ X,$$

где описание множества включает набор характерных атрибутов. Выше в тексте было дано отношение неразличимости на множестве объектов U. Такое отношение имеет связь не только со всем множеством условных атрибутов A, но и с любым подмножеством атрибутов $B \subseteq C$. Подобное отношение будем обозначать далее как $IND(B)$ и будем называть отношением неразличимости по B. Пару объектов x и y, полностью удовлетворяющих ему, не возможно отличить друг от друга при помощи атрибутов, взятых из B. Дадим формальное определение отношению неразличимости. Отношение неразличимости по B можно определить следующим образом:

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

Два любых объекта будут принадлежать одному классу эквивалентности, если не существует возможности отличить их друг от друга на базе использования данного подмножества атрибутов. Базируясь на $IND(B)$, можно дать два понятия. Первое, понятие верхнего приближения множества X по B. Второе, понятие нижнего приближения множества X по B.

Нижнее приближение множества X по B – это объединение классов эквивалентности отношения неразличимости, которые входят в X. Другими словами, нижнее приближение множества X по B – это $\underline{B}X = \bigcup_{X_i^B \subseteq X} X_i^B$. Верхнее приближение множества X по B – это объединение классов эквивалентности, часть объектов которых относится к X. Иными словами, нижнее приближение – это $\overline{B}X = \bigcup_{X_i^B \cap X = \varnothing} X_i^B$ BX.

Определение недостоверной или граничной области будет выглядеть следующим образом: $BN_B(X) = \underline{B}X \setminus \overline{B}X$. Полное название этой области – граничная или недостоверная область множества X относительно B. Она включает некоторое множество объектов, которые мы не можем уверенно отнести к X на основе информации, содержащейся в атрибутах из B.

Положительная область решающей системы S относительно B – это множество $POS_B(X)$, которое включает объекты, точно относящиеся к одному из классов решения на основе информации, содержащейся в атрибутах из B.

Рассмотрим пример обучающего множества (см. табл. 1). В этом примере множество объектов представляют собой множество описаний претендентов на получение работы. Обучающее множество U включает девять различных объектов. Каждый из которых в качестве характеристики имеет следующее множество атрибутов A={*Диплом, Опыт, Рекомендации*}. Все эти атрибуты относятся к условным атрибутам. Пополним этот список дополнительным атрибутом *«Решение»*. У него будет всего два значения: *«Принять»* и *«Отказать»*.

Таблица 1

Решающая система

S	Диплом	Опыт	Рекомендации	Решение
x_1	ХАИ	Средний	Отличные	Принять
x_2	ХАИ	Малый	Удовлетворительные	Отказать
x_3	ХПИ	Малый	Хорошие	Отказать
x_4	ХИРЭ	Высокий	Удовлетворительные	Принять
x_5	ХИРЭ	Средний	Хорошие	Отказать
x_6	ХИРЭ	Высокий	Отличные	Принять
x_7	ХПИ	Малый	Отличные	Отказать
x_8	ХИРЭ	Средний	Хорошие	Принять
x_9	ХПИ	Малый	Отличные	Отказать

Легко заметить, что в табл. 1 объекты x_5 и x_8, а также x_7 и x_9 имеют одни и те же значения условий, но первая пара имеет также и различные решения.

Проиллюстрируем на данном примере введенное понятие отношения неразличимости. Выберем несколько непустых подмножеств условных атрибутов: *{Диплом}*, *{Опыт, Рекомендации}* и *{Диплом, Опыт, Рекомендации}* и рассмотрим, как с их помощью определяется отношение неразличимости.

$$IND(\{Диплом\}) = \{x_1, x_2\}, \{x_4, x_5, x_6, x_8\}, \{x_3, x_7, x_9\}).$$
$$IND(\{Диплом, Рекомендации\}) = \{\{x_1\}, \{x_2\}, \{x_3\}, \{x_4\}, \{x_5, x_8\}, \{x_6\}, \{x_7, x_9\}).$$
$$IND(\{Диплом, Опыт, Рекомендации\}) = \{\{x_1\}, \{x_2\}, \{x_3\}, \{x_4\}, \{x_5, x_8\}, \{x_6\}, \{x_7, x_9\}).$$

Далее построим приближения для множества $X = \{x : Решение(x) = Принять\}$, используя всё множество условных атрибутов A. Тогда получим $\underline{A}X = \{x_1, x_4, x_6\}$, $\overline{A}X = \{x_1, x_4, x_5, x_6, x_8\}$, $BN_A = \{x_5, x_8\}$ и $U \setminus \overline{A}X = \{x_2, x_3, x_7, x_9\}$. Из того, что граничная область не пуста, следует, что множество X можно определить лишь приближенно (рис. 1).

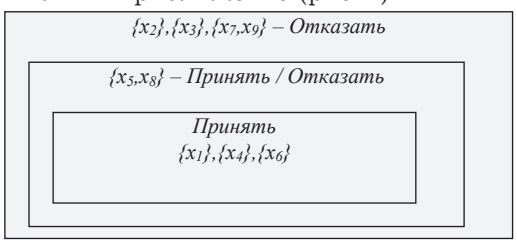

Рис. 1. Приближенное определение множества X

Определим положительную область для решающей таблицы S (табл. 2). $POS_A(d) = \underline{AC}_{Принять} \cup \underline{AC}_{Отказать} = \{x_1, x_2, x_3, x_4, x_6, x_7, x_9\} \neq U$. Это является следствием того, что для объектов x_5 и x_8 невозможно найти однозначное решение.

Итак, введем понятие обобщенного решения. Определим функцию $\partial_B : U \to P(V_d)$, называемую обобщенным решением S на множестве атрибутов $B \subseteq A$, следующим образом:

$$\partial_B(x) = \{v \in V_d : \exists x' \in U(x' IND(B) x \wedge d(x') = v)\}.$$

Т. е. обобщенным решением S для объекта x является множество решений объектов, входящих в тот же класс эквивалентности, что и объект x. Так для приведенного примера получим:

$$\partial_A(x_1) = \partial_A(x_4) = \partial_A(x_6) = \{Принять\},$$

$$\partial_A(x_2) = \partial_A(x_3) = \partial_A(x_7) = \partial_A(x_9) = \{Отказать\},$$

$$\partial_A(x_5) = \partial_A(x_8) = \{Принять, Отказать\}.$$

Пусть, обобщенные решения ∂_A системы S называются просто обобщенным решением S. Решающая таблица S называется непротиворечивой (детерминированной), если $|\partial_A(x)| = 1$ для любого объекта $x \in U$, в противном случае S называется противоречивой (недетерминированной). Просто заметить, что решающая таблица непротиворечива тогда и только тогда, когда $POS_A(d) = U$.

Помимо одинаковых объектов, которые могут несколько раз встречаться в таблице, другой причиной увеличения размеров таблицы часто являются несущественные условные атрибуты или зависимость одних условных атрибутов от других. Теперь, когда было введено одно из основных понятий решающих систем, обобщенное решение, определим понятие среза решающей таблицы.

Срезом относительно решения d таблицы $S = \{U, A \cup \{d\}\}$ называется минимальное подмножество атрибутов $B \subseteq A$, которое позволяет сохранить обобщенное решение для всех объектов обучающего множества, т. е. $\partial_A(x) = \partial_B(x)\ \forall x \in U$. В дальнейшем при рассмотрении решающих таблиц срез относительно решения будем называть просто срезом. Множество всех срезов решающей системы будем обозначать $RED(A,d)$.

Для нахождения срезов решающей таблицы используются матрица различимости и функция различимости. Матрицей различимости относительно решения d таблицы $S = \{U, A \cup \{d\}\}$ называется матрица $M(S) = c_{ij}^d$ размерности $n \times n$, элементы которой принимают следующие значения $c_{ij}^d = \varnothing$, если $d(x_i) = d(x_j)$, в противном случае $c_{ij}^d = \{a \in C : a(x_i) \neq a(x_j)\} - \{d\}$ для $i, j = 1,...,n$. То есть элементы матрицы различимости содержат информацию об условных атрибутах, значения которых отличаются для объектов разных классов.

Функцией различимости f_s для решающей системы $S = \{U, A \cup \{d\}\}$ будем называть логическую функцию от k булевых переменных $\hat{a}_1, \hat{a}_2, ..., \hat{a}_k$ соответствующих атрибутам $a_1, a_2, ..., a_k$, которая определяется как

$$f_S(\hat{a}_1, \hat{a}_2, ..., \hat{a}_k) = \wedge\{\vee \hat{c}_{ij}^d : 1 \leq j < i \leq n,\ \hat{c}_{ij}^d \neq \varnothing\}, \text{ где } \hat{c}_{ij}^d = \{\hat{a} : a \in \hat{c}_{ij}^d\}.$$

Множество всех простых импликант функции f_s определяет множество всех срезов решающей таблицы S [2]. Заметим, что импликантой логической функции f называется любая конъюнкция литер (переменных или их отрицаний), такая, что если эти литеры истинны на некоторой интерпретации, то функция f также истинна на этой интерпретации. Простой импликантой называется импликанта, никакая собственная часть которой не является импликантой.

Рассмотрим пример построения матрицы и функции различимости, соответствующих решающей системе, представленной в табл. 2, а затем найдем её срез. Для краткости обозначим условные атрибуты «длительность выпуска товара», «конкуренция» и «тип товара» как a, b и t соответственно. Матрица различимости приведена в табл. 3. Она является симметричной и имеет пустые элементы главной диагонали.

Таблица 2

Решающая система

S	Длительность выпуска товара	Конкуренция	Тип товара	Прибыль
x_1	Большая	да	*Swr*	Падает
x_2	Большая	нет	*Swr*	Падает
x_3	большая	нет	*Hwr*	Падает
x_4	средняя	да	*Swr*	Падает
x_5	средняя	да	*Hwr*	Падает
x_6	средняя	нет	*Hwr*	Растет
x_7	средняя	нет	*Swr*	Растет
x_8	малая	да	*Swr*	Растет
x_9	малая	нет	*Hwr*	Растет
x_{10}	малая	нет	*Swr*	Растет

Таблица 3

Матрица различимости

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
x_1	∅									
x_2	∅	∅								
x_3	∅	∅	∅							
x_4	∅	∅	∅	∅						
x_5	∅	∅	∅	∅	∅					
x_6	*a,b,t*	*a,t*	*a*	*b,t*	∅	∅				
x_7	*a,b*	*a*	*a,t*	*b*	*b,t*	∅	∅			
x_8	*a*	*a,b*	*a,bj*	*a*	*a,t*	∅	∅	∅		
x_9	*a,b,t*	*a,t*	*a*	*a,b,t*	*a,b*	∅	∅	∅	∅	
x_{10}	*a,b*	*a*	*a J*	*a,b*	*a,b,t*	∅	∅	∅	∅	∅

Матрице различимости соответствует следующая функция различимости:

$$f_S(a,b,t) = (a \vee b \vee t)(a \vee b)a(a \vee b \vee t)(a \vee b)$$
$$(a \vee t)a(a \vee b)(a \vee t)a$$
$$a(a \vee t)(a \vee b \vee t)a(a \vee t)$$
$$(b \vee t)ba(a \vee b \vee t)(a \vee b).$$

После упрощения получим $f_s(a,b,t) = ab$. Обозначение ab является краткой записью $a \wedge b$. Таким образом, срез решающей системы включает два атрибута: «длительность выпуска товара» и «конкуренция».

Заметим, что каждая строка в этой записи функции различимости соответствует некоторому столбцу матрицы различимости. Так, например, пятый объект можно отличить от седьмого по одному из атрибутов «наличие конкуренции» или «тип товара», а от девятого по одному из атрибутов «длительность выпуска товара» или «наличие конкуренции». Если из функции различимости взять только элементы, соответствующие одному из столбцов матрицы различимости (например, i-му столбцу), то мы получим функцию различимости относительно объекта

x_i. Множество всех первичных импликант этой функции определяет набор всех срезов относительно объекта x_i. Эти срезы содержат необходимую информацию, позволяющую отличить объект x_i от объектов, принадлежащих другим классам решения.

Выводы. Таким образом, в статье затронут ряд вопросов связанных с походом к решению задачи обобщения понятий по признакам или, другими словами, индуктивному обобщению. Во-первых, был подробно рассмотрен целый ряд понятий. К этим понятиям относятся: отношение эквивалентности, пространство аппроксимации, класс эквивалентности, составное множество, верхнее и нижнее приближение, решающая таблица, срез, обобщенное решение и т.д. Во-вторых, рассмотрен подход к нахождению срезов: матрица различимости, функции различимости. В-третьих, приведен пример нахождения конкретного среза – построение матрицы и функции различимости, а затем нахождение среза.

Список литературы:

1. Agrawal R. et al. Fast algorithms for mining association rules // Proc. 20th int. conf. very large data bases, VLDB. – 1994. – T. 1215. – C. 487-499.
2. Уэно Х. Представление и использование знаний // под ред. Х. Уэно, М. Исидзука; пер. с яп. И. А. Иванова. – М.: Мир, 1989. – 222 с.
3. Pawlak Z. Rough sets / Z. Pawlak // International Journal of Computer and Information Sciences. – 1982. – 11. – P. 341-356.
4. Pawlak Z. Rough set approach to knowledge-based decision support / Z. Pawlak // European Journal of Operational Research. – 1997. – 99. – P. 420–432.
5. Вагин В. Н. Достоверный и правдоподобный вывод в интеллектуальных системах / В. Н. Вагин, Е. Ю. Головина, А. А. Загорянская, В. М. Фомина; под ред. В. Н. Вагина, Д. А. Поспелова; изд. 2-е испр. и доп. – Москва: Физматлит, 2008. – 712 с.
6. Люггер Дж. Ф. Искусственный интеллект: стратегии и методы решения сложных проблем / Дж. Ф. Люггер. – Москва: «Вильямс», 2003. – 864 с.
7. Финн В. К. Об интеллектуально анализе данных // Новости искусственного интеллекта. – 2004. – № 3. – С. № -18.
8. Dietterich T. G. Machine Learning: An Artificial Intelligence Approach / Ed. by R. S. Michalski, J. Carbonell, T. M. Mitchell. – PaloAlto: Tioga, 1983. – P. 41–82.
9. Уотерман Д. Построение экспертных систем // пер. с англ.; под ред. Ф. Хейес-Рота, Д. Уотермана, Д. Лената – М.: Мир, 1987. – 442 с.

Мар'їн С.О., Ситніков Д.Е., Коваленко А.І.
Харківська державна академія культури

УЗАГАЛЬНЕННЯ ПОНЯТЬ ЗА ОЗНАКАМИ: МЕТОДИ ТЕОРІЇ ПРИБЛИЗНИХ МНОЖИН

Анотація
У статті розглядається ряд питань пов'язаних з індуктивним узагальненням понять в сучасних інтелектуальних та експертних системах. Проведено докладне дослідження логічного підходу до визначення переліку ознак, за яким будується таке узагальнення. Для рішення цієї задачи досліджується функція розрізнення, матриця розрізнення та зріз. Розглянуто приклад, який ілюструє використовуваний метод і його переваги.
Ключові слова: теорія приблизних множин, узагальнення понять, експертна система, інформаційна система, ставлення непомітності, простір апроксимацій, простір наближень, еквівалентність, нижнє наближення для безлічі, верхнє наближення для безлічі, узагальнене рішення.

Marin S.A., Sitnikov D.E., Kovalenko A.I.
Kharkiv State Academy of Culture

ROUGTH SET THEORY: CONCEPT OF GENERALIZATION

Summary
The article discusses a number of issues related to the inductive generalization of concepts in intelligent and expert systems. We studied logical approach to the definition of attributes list on which we can construct a generalization. To investigate the distinctiveness function, the matrix section and legibility. An example illustrating the method used and its advantages.
Keywords: rough sets theory, generalization of concepts, expert system, information system, attitude indistinguishable space approximations, approximation space, equivalence, the lower approximation of the set upper approximation for the set, a generalized solution.