УДК 004.94 DOI: 10.31673/2412-4338.2018.044046

Шатохина Н.К., Гололобов Д.А., Котомчак А.Ю., Сударева О.В.
Государственный университет телекоммуникаций, Киев

ЭВРИСТИЧЕСКИЙ МЕТОД ОБОБЩЕНИЯ НЕЧЕТКИХ ПРОДУКЦИЙ

Рассмотрена задача индуктивного обобщения базы знаний продукционного типа. Предложен эвристический метод решения этой задачи с целью, с одной стороны, сохранения знаний, предлагаемых экспертами, и, с другой стороны, сокращения общего объема базы знаний и возможностью выбора наиболее вероятных утверждений из базы. Такой подход позволяет эффективно получать, возможно, не точное решение, но за приемлемое время. Результаты работы могут быть применены для формирования и сжатия баз знаний, используемых в экспертных системах.

***Ключевые слова**: экспертная система, нечеткая продукция, база знаний, коэффициент уверенности, правила вывода.*

Шатохіна Н. К., Гололобов Д. О., Котомчак О. Ю., Сударєва О. В.
Державний університет телекомунікацій, Київ

ЕВРИСТИЧНИЙ МЕТОД УЗАГАЛЬНЕННЯ НЕЧІТКИХ ПРОДУКЦІЙ

Процес узагальнення заснований на порівнянні описів вихідних об'єктів, заданих набором характеристичних значень, і вибором найбільш характерних фрагментів цих описів. Цей процес називається індуктивним формуванням поняття. Завдання індуктивного формування понять, що також називається завданням узагальнення понять за допомогою атрибутів, полягає у побудові концепції, яке базується на аналізі вибірки, що дозволяє використовувати певне правило розпізнавання для правильного поділу всіх позитивних і негативних об'єктів набору досліджуваного зразка.

У статті розглядається проблема індуктивного узагальнення бази знань продукційного типу. Важливою вимогою повинно бути, з одного боку, збереження знань, запропонованих експертами, а з іншого – зменшення загального обсягу бази знань. Крім того, необхідно також забезпечити можливість вибору найбільш вірогідних тверджень з бази. Представлено евристичний метод розв'язання проблеми узагальнення нечітких продукцій. Метод дозволяє замінити повний пошук продукцій бази знань більш швидкою процедурою, яка враховує частоту виникнення фактів у продукціях експертних висновків. Запропоновано метод представлення продукцій у вигляді булевих функцій, за допомогою якого було призначено числові ваги для управління їх обробкою. Розроблено етапи процесу узагальнення нечітких продукцій. Результати роботи можуть бути використані для побудови бази знань експертних систем з метою зменшення розміру бази знань за допомогою індуктивних узагальнених тверджень, представлених у вигляді нечітких продукцій, з отриманням майже мінімального набору продукцій, виключаючи взаємні фрагменти.

***Ключові слова**: експертна система, нечітка продукція, база знань, коефіцієнт впевненості, правила виведення*

Shatokhina N. K., Hololobov D. O., Kotomchak O. Yu., Sudarieva O. V.
State University of Telecommunications, Kyiv

HEURISTIC METHOD OF GENERALIZATION OF FUZZY PRODUCTIONS

The generalization process is based on the comparison of descriptions of the original objects, given by a set of characteristic values, and the selection of the most characteristic fragments of these descriptions. This process is also called inductive formation of concept. The task of the inductive formation of concepts, also called the task of generalizing concepts by means of attributes, is to build a concept based on the analysis of a sample, which allows using a certain recognition rule to correctly separate all positive and negative objects of the set of the sample under consideration.

This paper addresses the problem of inductive generalization of a knowledge base of a production type. An important requirement should be, on the one hand, preserving the knowledge offered by experts, and, on the other hand, reducing the total volume of the knowledge base. In addition, the possibility of choosing the most likely allegations from the base must also be ensured. The paper presents a heuristic method for solving the problem of generalization of fuzzy productions. The method allows replacing a complete search of knowledge base productions with the faster procedure that takes into account the frequency of occurrence of facts in the productions of expert opinions. A method for presenting products in the form of Boolean functions has been proposed, with which the numerical weights have been assigned to control their processing. The stages of the generalization process have been developed. The results of the work can be used to build knowledge bases of expert systems in order to reduce the size of the knowledge base by inductive summarizing statements presented in the form of fuzzy productions, with obtaining an almost minimal set of productions excluding reciprocal fragments.

***Keywords**: expert system, fuzzy products, knowledge base, confidence coefficient, assurance rules.*

1. Введение

Формирование базы знаний (БЗ) является сложным и важным этапом проектирования экспертной системы (ЭС). По содержанию БЗ состоит из рассуждений экспертов (фактов), общее число которых может быть достаточно велико, некоторые рассуждения могут представлять собой частные факты. Поэтому возникают задачи обобщения знаний, т.е. задачи получения более общих фактов по нескольким частным с условием сохранения знаний исходного множества. В силу природы данной задачи получение точного решения является принципиально трудной задачей. Имеются попытки решения этой задачи [1, 2], не гарантирующие получения минимального объема БЗ, однако и не гарантирующие и сохранения исходных знаний исходной БЗ.

В данной работе рассмотрена *актуальная* задача получения приближенного решения, которое позволяет решать задачу обобщения исходного множества фактов с сохранением знаний, заложенных экспертами в БЗ.

2. Анализ литературных данных и постановка проблемы

Пусть задано множество U (универсум) и непересекающееся с U множество Y. Элементы $(a, b, c,\ldots)\in U$ называются элементарными фактами, а элементы $(x, y,...) \in Y$ – целевыми фактами. Произвольные подмножества $(W, Q, Z,\ldots) \subset U$ – это подмножество элементарных фактов, соответствующие утверждениям экспертов.

Формулой (продукционным правилом или продукцией) называется выражение вида $(W\rightarrow y, h)$, где W – некоторое подмножество U, $y\in Y$, а h – некоторый числовой коэффициент, называемый коэффициентом уверенности (КУ), и $0<h\leq 1$.

Отдельная формула языка задает некоторую закономерность, которая в зависимости от состояния предметной области (ПО) формально описывается понятием интерпретации - нечетким множеством $I=(\{a, \mu_a\})$, $a\in U$, $\mu_a\in[0, 1]$. Каждое a соответствует некоторому высказыванию о состоянии ПО, а μ_a – степени уверенности в этом высказывании. То состояние ПО, для которого данная закономерность справедлива, описывается понятием модели соответствующей формулы.

Моделью формулы $(\{a_1, a_2,\ldots, a_n\}\rightarrow y, h)$ называется такая интерпретация $I=(\{a, \mu_a\})$, для которой выполняется условие $\mu_y\geq\min(\mu_{a1},\ldots, \mu_{an}, h)$. В частности, при $n=0$ $\mu_y\geq h$.

Множество всех моделей формулы f обозначается Mod (f) (указываются только те факты, КУ которых отличен от 0).

Под базой знаний понимается некоторое конечное множество формул. При этом предполагается, что БЗ описывает те состояния ПО, в которых одновременно выполняются

все формулы, включенные в БЗ. Иными словами, множество моделей БЗ является Mod (T) = ∩ Mod (f_i) по всем $f_i \in T$.

Формула f является логическим следствием БЗ T ($T|$-f), если Mod (T)$\subseteq M$od (f), т.е. логическими следствиями БЗ являются те формулы, которые выполняются всегда, когда выполняются все формулы БЗ.

Установить отношение логического следования непосредственно по введенному определению невозможно в силу бесконечности множества интерпретаций. Традиционно выход из такой ситуации состоит в попытке построения чисто синтаксических правил, конечным образом описывающих бесконечные множества объектов. Такие правила, позволяющие строить для заданного набора формул всевозможные его логические следствия, называются правилом вывода. Совокупность правил вывода для заданной ЭС определяет ее механизм вывода (МВ). Факт выводимости формулы f из БЗ T с помощью системы С правил вывода обозначается через $T||\underset{c}{=}f$.

Система C называется непротиворечивой, если для любых T и f из того, что $T||\underset{c}{=}f$, следует $T|$-f. Система называется полной, если выполняется обратное, т.е. $T|$-f из следует $T||\underset{c}{=}f$. Непротиворечивость системы гарантирует, что все заключения ЭС правильны, а полнота – что пользователь получает все правильные заключения.

Будем использовать следующую систему правил вывода:

$$(W\rightarrow \text{у}, k)||\underset{c}{=}(Q\rightarrow \text{у}, h) \text{ тогда и только тогда, когда } h\leq k \text{ и } W\subseteq Q.$$

Как видим, правило формулируется для продукций ($W\rightarrow$у, k) с одним единственным целевым фактом у, поэтому при записи правил нет необходимости указывать явным способом целевой факт, т.е. (W, k), где $W\subseteq U$ и $k\subseteq$ [0, 1].

Индуктивное построение БЗ по примерам является в некотором смысле обратной задачей: по известным примерам заключений (логическим следствиям) требуется построить БЗ, т.е. набор формул, для которых каждый исходный пример является логическим следствием. Такой набор формул называется индуктивным обобщением заданного множества примеров.

Индуктивным обобщением (ИО) заданного набора формул E называется такой набор формул S, что $S|$-E.

Для фиксированного E существует много различных ИО. В качестве ИО для E всегда можно взять само E. Естественным основанием для сравнения различных наборов формул является само понятие ИО.

Набор формул S называется более общим, чем набор $S'(S \succ S')$, если $S \mid -S'$.

Из более общего набора формул, можно вывести больше различных следствий. При поиске ИО интересны наборы формул, из которых можно вывести побольше следствий, но большие ИО малоинтересны, поскольку они практически не отражают специфики ПО, задаваемой примерами от экспертов. Поэтому ИО выбирается таким, чтобы с одной стороны, давало бы сравнительно большое число следствий, а с другой достаточно верно отражало опыт, заложенный в примерах.

Возможны такие ситуации, когда из БЗ можно удалить некоторое правило, ничего при этом не потеряв: все следствия, которые выводились из старой БЗ, будут выводится из новой, и наоборот. Эти ситуации исключаются из рассмотрения в дальнейшем, за счет введения понятие неизбыточной БЗ.

База знаний S называется неизбыточной, если для любого $S'\subseteq S$ множество следствий S' не совпадает с множеством следствий всего S.

БЗ S является неизбыточной, тогда и только тогда, когда $\forall$ $(W\rightarrow x; k)$, $(Q\rightarrow x; h)\in S$ из $W\subseteq Q$ следует, что $h>k$.

Доказательство этого утверждения приведено в [3]. В дальнейшем под ИО будем понимать неизбыточные ИО.

Заметим, выбранные правила вывода позволяют для каждого следствия из БЗ понизить его КУ. Следовательно, знание максимального коэффициента уверенности, с которым данное следствие выводится из БЗ, однозначно определяет множество всех остальных коэффициентов уверенности, с которыми его можно вывести из БЗ. Пользователь, задавая примеры, имеет ввиду именно такие максимальные коэффициенты, т.е. выделить такие ИО, из которых известные примеры нельзя вывести с большим коэффициентом уверенности.

Для заданного набора примеров E индуктивное обобщение S называется характеристическим (ХИО), если для любой формулы $(W\rightarrow x, h)\in E$ и $\forall$ $q>h$ формула $(W\rightarrow x, q)$ не выводится из S.

Таким образом, задача индуктивного построения БЗ сводится к поиску для заданного множества примеров наиболее общего ХИО из множества всех ХИО, где под наиболее общим набором в множестве наборов M понимается такое S, что $\forall$ $T\in M$ выполняется $T\succ S$.

Необходимые и достаточные условия существования ХИО [3]:

Для множества примеров E существует ХИО в том и только в том случае, когда для любых (W, k) и (Q, h) из E из условия $W\subseteq Q$ следует $h\geq k$.

Поскольку при данных правилах вывода продукции с разными целевыми фактами не могут участвовать одновременно в выводе, поэтому при наличии $n>1$ таких фактов задача построения наиболее общего ХИО сводится к n независимым подзадачам с одним очередным целевым фактом.

В [3, 4] было показано, что правила вывода описывает отношение частичного порядка на множестве $P = 2^u\times K$ и отношение предпорядка на множестве 2^P. Здесь через 2^P (2^U) обозначается булеан $P(U)$ [5]. Было показано что, решение задачи построения ХИО свелось нахождению множества $E_{min}(E)$ минимальных элементов относительно этого отношения. Элементы множества $E_{min}(E_i)$ представляют собой в свою очередь - минимальные элементы для продукций $W\in E_i$ с одинаковым коэффициентом уверенности i. А поскольку данное отношение для продукций слоев является предпорядком, то каждой продукции $W\in E_i$ в общем случае соответствует несколько минимальных элементов – $E_{min}(W)=\{m_1,m_2,...,m_r\}$. Поэтому для каждого слоя необходимо выбрать из этих подмножеств минимальное подмножество элементов, которое покрывало все продукции слоя i, это и составит ХИО(E_i). ХИО(E) всего множества продукций представляет собой объединение всех ХИО (E_i).

Следует заметить, что задача выбора минимального подмножества $E_{min}(W)$ элементов слоя E_i из множеств минимальных элементов продукций $W\in E_i$ свелась [4] к задаче покрытия строк столбцами, которая является принципиально трудной [4].

3. Цель и задачи исследования

Исходя из рассмотренных материалов видно, что задача выбора минимального подмножества $E_{min}(W)$ элементов слоя E_i из множеств минимальных элементов $W\in E_i$ свелась [4] к задаче покрытия строк столбцами, которая является принципиально трудной, т.е. переборной. Таким образом, разработка подходов решения задачи сжатия БЗ, в нашем случае обобщения множества нечетких продукций, сложностью меньшей, чем полный перебор является целью данной работы. Для решения данной задачи предлагается эвристический метод построения ХИО. Эвристический метод не дает точного решения, но является одним из подходов к уменьшению объемов продукционной БЗ.

4. Результаты исследования

Задано множество продукций E. Задача состоит в том, чтобы построить множество продукций E' такое, что из E' выводится E и из E' нельзя удалить ни одну продукцию. Для этого выполняются последовательно два алгоритма.

Первый алгоритм производит построение совокупности множеств минимальных продукций для заданного. Он строит для всех продукций из E дизъюнктивных нормальных форм (ДНФ), каждый конъюнкт которой позволяет вывести эту продукцию и не выводит ни одну продукцию с меньшим КУ, а также является минимальным.

Второй алгоритм из каждого множества конъюнктов полученной совокупности выбирает по одному, применяя принцип наибольшей индивидуальности и частоты использования фактов. В результате будет построено ХИО исходного множества E. Во втором алгоритме реализован эвристический метод.

В качестве входной информации ***первого алгоритма*** выступает множество продукций E. Поставим в соответствие множеству E таблицу, строки, которой соответствует продукциям, а столбцы – элементарным фактам. Каждая строка таблицы состоит из 0,1 и имеет свой вес. Наличие 1 в строке означает включение в продукцию соответствующего факта, 0 – отсутствие факта в продукции. В соответствии с правилами вывода, каждая строка таблицы содержит все варианты вывода соответствующей продукции, поэтому каждой строке можно поставить в соответствие булеву функцию. Например, для некоторой продукции (W,0.6) в таблице имеются 1 в столбцах *a, c*, тогда W выводится из *(ac; 0.6)*. Значит, для продукции W можно поставить в соответствие функцию $f(W)= (a \vee c)$.

Представим исходное множество продукций в виде совокупности слоев E_i. С этой целью отсортируем таблицу в порядке возрастания весов строк. Будем обрабатывать слои в порядке возрастания их номеров (весов). Заметим, сложность алгоритма сортировки в лучшем случае будет равняться $O(n \times 2\log 2n)$, в худшем случае – $O(n^2)$, n=|E|.

По строкам слоя E_1 составляется дизъюнкция, содержащая все варианты вывода соответствующих продукций слоя E_1.

Зафиксируем слой E_i, $i \in [2,n]$. Для каждой продукции W слоя E_i построим семейство, элементами которого являются множества $(W \backslash V_j)$, $j<i$. Логически данное выражение описывает продукции (Z,k_i), $Z \subset W$, из которых выводится продукция W с наибольшим КУ k_i и не выводятся ни одна продукция V_j слоев E_j, $j<i$. Каждое множество $W \backslash V_j$ представим в виде дизъюнкции.

Построим формальную конъюнкцию полученных дизъюнкций и преобразуем ее в ДНФ. Преобразование к ДНФ производим с помощью дистрибутивного закона: $(x \vee z)(y \vee z)=xy \vee z$ и закона поглощения $x(x \vee y)=x$. Конъюнктам полученных ДНФ присваивается коэффициент уверенности k_i, соответствующий продукции W или k_1, если рассматривался слой E_1.

Обозначим через $D_i= \{d_{i1}, d_{i2}, .., d_{il}\}$ множество конъюнктивных членов $W \in E_i$.

Второй алгоритм реализует предлагаемый *эвристический подход.*

Исходные данные второго алгоритма – это наборы конъюнктивных членов ДНФ, которые были получены раннее для всех продукций множества E.

В ХИО включаются все продукции из E_1 (из слоя с наименьшим КУ), а также все продукции, для которых получена КНФ, состоящая из конъюнкта ($|D\urcorner|=1$). Соответственно из исходного множества E удаляются эти продукции.

Создается новая таблица, строки которой соответствуют конъюнктам $d_i \in D$ всех продукций, общее количество которых составляет $m = \sum_{i=1}^{|E|} |D_i|$. Столбцы таблицы соответствуют фактам $a \in U$, количество столбцов равно $h=|U|$. На пересечении r-ой строки

и *g*-го столбца стоит 1, если факт *g* ходит в *r*-ый конъюнктивный член, иначе 0. Строки таблицы сгруппированы по слоям, а слои разбиты по зонам. Каждая зона содержит все конъюнктивные члены для одной из продукций $W \in E$.

При выполнении алгоритма таблица уменьшается до тех пор пока не будут учтены все продукции.

Для каждой строки *i* подсчитывается: S_i – сумма значений. Если найдется хотя бы одна строка *i*, для которой $S_i=1$, тогда соответствующая ей продукция включается в ХИО, а остальные строки зоны удаляются из таблицы.

Дальнейшие вычисления выполняются до тех пор, пока не будет получена пустая таблица.

Вычисляется сумму значений каждого столбца (кроме помеченного).

Определяется столбец *g* с max значением S_g. Если их несколько, то берется первый.

Разыскиваются все зоны, в которых имеется хотя бы одна строка, содержащая 1 на пересечении с *g*-ым столбцом. Из найденной зоны удаляются все строки, содержащие 0 в *g*-ом столбце, и если в зоне осталась одна строка, то соответствующая продукция включается в ХИО, а строка удаляется.

Если же в зоне было несколько строк с 1 в *g*-ом столбце, то пересчитываются значения сумм в этих строках по формуле: $S=S-1$, и если для хотя бы в одной строке получится значение $S=0$, то соответствующая продукция включается в ХИО и производится удаление всей зоны из таблицы. После чего, отмечается *g*-ый столбец как просмотренный и производится повтор пересчета по столбцам.

После удаления из таблицы всех строк искомое множество ХИО(Е) построено.

5. Обсуждение результатов исследования

Оценим сложность предложенного алгоритма. Временная сложность предложенного алгоритма определяется количеством просмотров таблицы. Таблица несколько раз просматривается по строкам, и состоит из таких шагов:

- подсчет сумм по строкам (mh)
- выбор столбца с максимальной суммой (mh); за счет еще очередного просмотра (m) таблицы вдоль найденного столбца выполняется удаление из таблицы, по крайней мере, или зоны одной продукций, или одну из строк, уменьшая на 1 суммы по строкам (m).

Заметим, что поскольку при каждом просмотре удаляется из таблицы, по крайней мере, одна строка, то общее число просмотров таблицы не может превосходить m. В результате сложность алгоритма не превосходит величины $mh + m(mh+2m) = mh + m^2(h+2)$ т.е. имеет *полиномиальную сложность*.

6. Выводы

Поскольку проблема обобщения знаний является одной из важных проблем искусственного интеллекта, сложной в математическом аспекте, то решение ее является важным в теоретическом и практическом плане. В работе приведен эвристический метод решения задачи обобщения нечетких продукций, который позволяет заменить полный перебор продукций БЗ на более быструю процедуру, учитывающую частоту появления фактов в продукциях базы знаний. Результаты работы могут быть использованы при построению БЗ ЭС, в целях уменьшения объема БЗ за счет индуктивного обобщения высказываний.

Список использованных источников

1. Шатохина Н. К. Использование вероятностных генетических алгоритмов с адаптивной мутацией для задачи индуктивного обобщения нечеткой базы знаний / Н. К. Шатохина, В. С. Османов // Наукові праці Донецького державного технічного університету. Серія: Інформатика, кібернетика та обчислювальна техніка. – 2016. – Випуск 2 (23). – С.81-86.

2. Османов В. С. Использование генетических алгоритмов для построения индуктивного обобщения нечеткой базы знаний / В.С. Османов // Матеріалы II міжнародної науково-практичної конференції «Сучасна наука: проблеми і перспективи» (ч.2), 15-16 жовтня 2016 р., Київ МЦНД. – 2016. – С.27-29.

3. Шатохина Н. К. Об индуктивном построении базы знаний экспертных систем / Н. К. Шатохина, П. А. Шатохин // Наукові праці Донецького державного технічного університету. Серія: Обчислювальна техніка та автоматика. – 1999. – Випуск 12. – С.158-164.

4. Грунский И. С. Об индуктивном обобщении нечетких заключений / И. С. Грунский, Н. К. Шатохина // Наукові праці Донецького державного технічного університету. Серія: Обчислювальна техніка та автоматика. – 2001. – Випуск 25. – С.154-160.

5. Фоменко Т. Н. Высшая математика. Общая алгебра. Элементы тензорной алгебры / Т. Н. Фоменко. – Москва: Издательство Юрайт, 2018. – 121 с.

References

1. Shatokhina N. K., Osmanow V. S. (2016). The Use of Probabilistic Genetic Algorithms with an Adaptive Mutation for the Problem Of Inductive Generalization of a Fuzzy Knowledge Base. Scientific works of the Donetsk State Technical University. Series: Informatics, Cubernetics and Computing. 2(23). 81-86.

2. Osmanow V. S. (2016). The Use of Genetic Algorithms for Constructing Inductive Generalization of a Fuzzy Knowledge Base. II International Scientific and Practical Conference "Modern Science: Problems and Prospects" (P. 2) October 15-16, 2016 - Kyiv ICSTD. 27-29.

3. Shatokhina N. K., Shatokhin P. A. (1999). On the inductive construction of the knowledge base of expert systems. Scientific works of the Donetsk State Technical University. Series: Computing and Automatics. 12. 158-164.

4. Grunsky I. C., Shatokhina N. K. (2001). On inductive generalization of fuzzy conclusions. Scientific works of the Donetsk State Technical University. Series: Computing and Automatics. 25: 154-160.

5. Fomenko T. N. (2018). Higher Mathematics. General algebra. Elements of a tensor algebra. Moscow: Publishing House Yurayt. 121.

Автори статті (Authors of the article)

Шатохіна Наталія Костянтинівна – к.т.н., доцент кафедри системного аналізу (Shatokhina Nataliia Kostiantynivna – PhD in technic, assistant professor of System Analysis Department). Phone: +380 50 472 5836. E-mail: nkhatokh@gmail.com.

Гололобов Дмитро Олександрович – к.ф.-м.н., доцент кафедри системного анализу (Hololobov Dmytro Oleksandrovych – PhD in physic and mathematic, assistant professor of System Analysis Department). Phone: +380 68 634 7821. E-mail: vobd@ukr.net.

Котомчак Олександр Юрійович – старший викладач кафедри системного анализу (Kotomchak Oleksandr Yuriiovych – senior teacher of System Analysis Department). Phone: +380 67 164 6155. E-mail: katoa@ukr.net.

Сударєва Ольга Валеріївна – ассистент кафедри системного анализу (Sudarieva Olha Valeriivna – assistant of System Analysis Department). Phone: +380 95 851 2683. E-mail: osudareva@gmail.com.