УДК 004.825

ИЗВЛЕЧЕНИЕ ЗНАНИЙ ИЗ БОЛЬШИХ МАССИВОВ ДАННЫХ

© 2014 И.А. Лёзин, Д.Е. Маркелов

Самарский государственный аэрокосмический университет имени академика С.П. Королёва (национальный исследовательский университет), г. Самара

Поступила в редакцию 17.12.2013

Знание – совокупность информации и правил вывода о мире, свойствах объектов, закономерностях процессов и явлений, а также правилах использования их для принятия решений. Главное отличие знаний от данных состоит в их структурности и активности, появление в базе новых фактов или установление новых связей может стать источником изменений в принятии решений.В процессе своей работы как научно-исследовательские отделы, так и коммерческие компании накапливают большой массив фактов, показателей, измерений. Часто эксперт не может охватить весь объем информации. Рассматриваемый подход позволяет провести анализ текущей ситуации, установить взаимосвязи между показателями и сформировать правило влияния факторов друг на друга. Предлагается удобный для практики метод извлечения знаний из больших массивов данных. Метод представляет собой порядок расчетов, позволяющих решить задачу перехода от данных к знаниям. Разработанный метод предназначен для повышения эффективности ведения проектов НИР и ОКР в аэрокосмических приложениях.

Ключевые слова: извлечение знаний, большие массивы данных, нечеткая логика.

Введение

В настоящее время в условиях резкого повышения объема информации является целесообразным разработка методов и инструментов по извлечению знаний из больших массивов данных. При ведении научных исследований, обработке результатов наблюдения или эксперимента обычно приходится сталкиваться с большим числом испытаний. Такие исследования приводят к накоплению массивов числовых и функциональных характеристик значительной размерности, что затрудняет хранение, анализ и интерпретацию полученных результатов. При этом объем выборки может достигать огромных размеров, и оперировать ими становится не очень удобно. Если в условиях конкретной задачи можно исходить из предположения о том, что данная выборка отображает какие-либо функциональные зависимости, пусть даже неизвестные, то в таком случае можно перейти от хранения информации в виде числовых массивов к хранению набора правил, описывающих эти зависимости, то есть к базе знаний.

Проблема извлечения знаний является составной частью инженерии знаний – области информационных технологий, занимающейся решением задачи преобразования знаний в объект обработки на ЭВМ.

Лёзин Илья Александрович, кандидат технических наук, доцент кафедры "Информационных систем и технологий". E-mail: ilyozin@yandex.ru

Маркелов Дмитрий Евгеньевич, студент первого курса магистратуры факультета информатики. E-mail: rtg7@yandex.ru

С другой стороны большинство сложных систем и процессов обладают объективной неопределенностью, что приводит к необходимости использования нечеткой логики при создании моделей таких объектов.

Общая постановка задачи

Преобразование экспериментальной информации в нечеткие базы знаний состоит из двух этапов:

- генерация первоначального набора правил;
- определение оптимальной структуры базы знаний.

Для реализации первого этапа необходимо решить три задачи:

- задать структуру базы нечетких продукционных правил;
- разбить пространство входных и выходных переменных;
- определить метод формирования начальной базы знаний.

Для оптимизации базы знаний необходимо решить следующие задачи:

- определить критерии оптимальности базы знаний;
- сократить число правил;
- провести параметрическую оптимизацию базы знаний на основе экспериментальных данных.

При этом критерием качества извлеченных закономерностей является близость результатов лингвистической аппроксимации и соответствующих экспериментальных данных. Для количественной оценки точности результатов необходимо применить механизм нечеткого вывода, со-

стоящий из следующих этапов:

- фаззификация входных переменных;
- формирование базы нечетких продукционных правил;
- агрегирование предусловий в нечетких продукционных правилах;
- композиция заключений;
- аккумулирование заключений;
- дефаззификация значений.

Исходя из постановки задачи, можно перейти к описанию объекта одного правила базы знаний с nвходами и одним выходом:

$$y = f(x_1, x_2, ..., x_n), \quad (1)$$

для которого известны интервалы изменения входов и выхода:

$$x_i \in [x_i, \overline{x_i}], i = \overline{1,n}, y \in [y, \overline{y}], \quad (2)$$

По имеющейся обучающей выборке из M пар экспериментальных данных входы-выход:

$$\{X_p, y_p\}, \quad (3)$$

где $X_p = \{x_1^p, x_2^p, ..., x_n^p\}$ – входной вектор в р-ой паре, $p = \overline{1,M}$, необходимо синтезировать знания об объекте в виде системы нечетких высказываний вида:

$$\begin{array}{l} ЕСЛИ\ [(x_1 = \alpha_1) и (x_2 = \alpha_2) и ... и (x_n = \alpha_n)] \\ ТО\ y \in d_j = [y_{j-1}, y_j], \end{array} \quad (4)$$

где α_k – некоторый интервал области входных данных x_i,

d_j – интервал выходных данных y.

Рассмотрим основные математические понятия, составляющую базу рассматриваемого метода, а также приведем обоснование их выбора.

Нечеткая логика

Нечеткая логика основана на использовании таких оборотов естественного языка, как “далеко”,“близко”, “холодно”, “горячо”. Классическая или булева логика имеет один существенный недостаток – с еепомощью невозможно описать ассоциативное мышление человека. Классическая логика оперируеттолько двумя понятиями: ИСТИНА и ЛОЖЬ, и исключая любые промежуточные значения, таким образом весь окружающий мир представляется только в черном ибелом цвете, вдобавок исключая из языка любые ответы на вопросы, кроме “да” и “нет”. Термином “лингвистическая переменная” можно связать любую физическую величину, для которой нужно иметь больше значений, чем только “да” и “нет”. Например, можно ввести переменную “возраст” и определить для нее термы “юношеский”, “средний”, “преклонный”.

Лингвистические переменные

Лингвистической называется переменная, принимающая значения из множества слов или словосочетаний некоторого естественного или искусственного языка. Формально, лингвистическая переменная определяется следующим образом.

Лингвистическая переменная задается пятеркой:

$$\langle x, T, U, G, M \rangle, \quad (5)$$

где x – имя переменной,

T – терм-множество, каждый элемент которого (терм) представляется как нечеткое множество на универсальном множестве,

U– универсальное множество,

G – синтаксические правила, часто в виде грамматики, порождающие название термов,

M – семантические правила, задающие функции принадлежности нечетких термов, порожденных синтаксическими правилами G.

С понятием лингвистической переменной тесно связано понятие нечеткого множества.

Нечёткие множества

Пусть E – универсальное множество, x – элемент E,a R – некоторое свойство. Обычное четкое подмножество A универсального множества E, элементы которого удовлетворяют свойству R, определяется как множество упорядоченных пар

$$A = \{mA(x) / x\}, \quad (6)$$

где $mA(x)$ – характеристическая функция, принимающая значение 1, если x удовлетворяет свойству R, и 0 – в противном случае.

Нечеткое подмножество отличается от четкого тем, что для элементов x из E нет однозначного ответа “да-нет” относительно свойства R. В связи с этим, нечеткое подмножество A универсального множества E определяется как множество упорядоченных пар:

$$A = \{mA(x) / x\}, \quad (7)$$

где $mA(x)$ –характеристическая функция принадлежности, принимающая значения в некотором вполнеупорядоченном множестве M (например, $M = [0,1]$).

Коэффициент принадлежности определяется через функцию принадлежности.

Функции принадлежности

Значениями функции принадлежности ($MF(x)$ – membershipfunction) являются рациональные числа из интервала [0,1], где 0 означает отсутствие принадлежности к множеству, а 1 – полную принадлежность.

Степень принадлежности может быть опре-

делена явно функциональной зависимостью $\mu_A(x)$, либо дискретно – путём задания конечной последовательности значений $x \in \{x_n\}$ в виде:

$$A(x) = \left\{ \frac{\mu(x_1)}{x_1}, \frac{\mu(x_2)}{x_2}, \ldots, \frac{\mu(x_n)}{x_n} \right\}, \quad (8)$$

Под физическим смыслом функции принадлежности $\mu_A(x)$ нечеткого множества A понимается вероятность того, что лицо принимающее решение отнесет элемент x к множеству A.

Построение базы знаний

Одним из основных методов представления знаний в экспертных системах являются продукционные правила, позволяющие приблизиться к стилю мышления человека. Любое правило продукций состоит из посылок и заключения. Возможно наличие нескольких посылок в правиле, в этом случае они объединяются посредством логических связок И, ИЛИ.

Нечеткие системы основаны на правилах продукционного типа, в качестве посылки и заключения в правиле используются лингвистические переменные, что позволяет избежать ограничений, присущих классическим продукционным правилам.

База знаний — это совокупность знаний, описанных с использованием выбранной формы их представления

Построение базы знаний состоит из нескольких этапов.

Этап 1. Разбиение пространств входных и выходных переменных.

Каждое значение выборки принадлежит интервалу, который определяется минимальным и максимальным значениями по каждой переменной $x_i \in [x_i^{\min}, x_i^{\max}]$, $y \in [y^{\min}, y^{\max}]$. Необходимо разбить области определений этих переменных на отрезки. Причем число этих отрезков, а также их длина для каждой переменной подбираются индивидуально.

Этап 2. Формирование начальной базы правил.

Подход к формированию начальной базы правил основан на том, что изначально каждому примеру из выборки ставится в соответствие отдельное правило. Для этого для каждого $(x_1^k, x_{2,\ldots,}^k x_n^k, y^k) k = \overline{1,K}$, К – количество экспериментальных пар значений в выборке определяются степени принадлежности заданных значений переменных к соответствующим нечетким множествам. После чего каждому обучающему примеру ставятся в соответствие те нечеткие множества, степени принадлежности к которым у соответствующих значений переменных из этого примера являются максимальными. Сформированное таким образом множество правил и составляет начальную базу правил.

Автоматизация расчетов

Проблема извлечения знаний является составной частью инженерии знаний – области информационных технологий, занимающейся решением задачи преобразования знаний в объект обработки на ЭВМ.

С другой стороны большинство сложных систем и процессов обладают объективной неопределенностью, что приводит к необходимости использования нечеткой логики при создании моделей таких объектов.

Разработанная система позволяет провести анализ текущей ситуации, установить взаимосвязи между показателями и сформировать правило влияния факторов друг на друга.

Для реализации поставленной задачи система должна иметь средства сбора, формализации и дальнейшего использования знаний пользователей для повышения качества выполнения основных функций системы.

Алгоритм генерации базы правил состоит из следующих основных этапов: загрузка обучающей выборки, кластеризация, определение функции принадлежности, удаление противоречивых и дублирующих правил.

Статистическая обработка

Сложность построения базы знаний сильно зависит от объема выборки. Сократить объем выборки можно путем исключения из рассмотрения переменных, которые не вносят существенный вклад в значение результирующей переменной. Таким образом, задачу исключения переменных можно сформулировать как нахождение оценки взаимосвязи переменных, решить которую можно статистическими методами.

Корреляционная зависимость — статистическая взаимосвязь двух или нескольких случайных величин (либо величин, которые можно с некоторой допустимой степенью точности считать таковыми). При этом изменения значений одной или нескольких из этих величин сопутствуют систематическому изменению значений другой или других величин. Для оценки связи переменных между собой необходимо рассчитать коэффициент корреляции.

Коэффициент корреляции рангов, предложенный К. Спирменом, относится к непараметрическим показателям связи между переменными, измеренными в ранговой шкале. При расчете этого

коэффициента не требуется никаких предположений о характере распределений признаков в генеральной совокупности. Этот коэффициент определяет степень тесноты связи порядковых признаков, которые в этом случае представляют собой ранги сравниваемых величин.

При использовании коэффициента ранговой корреляции условно оценивают тесноту связи между признаками, считая значения коэффициента равные 0,3 и менее, показателями слабой тесноты связи; значения более 0,4, но менее 0,7 - показателями умеренной тесноты связи, а значения 0,7 и более - показателями высокой тесноты связи. Таким образом, переменные с низким показателем можно исключить из рассмотрения.

Кластеризация

В результате анализа этапа кластеризации при построении баз знаний было решено, что является целесообразным реализовать его с помощью алгоритма Абе-Лэна. Достоинством данного алгоритма является то, что он рассматривает не только близость значений внутри области переменных, но и связи между входной и результирующей переменными.

Определение функции принадлежности

На этапе определения коэффициента принадлежности было принято допущение, что границы значений функции принадлежности лежат на серединных значениях соседних с рассматриваемым кластером. Точная подстройка вида функции принадлежности осуществляется изменением параметра кривой функции.

Оптимизация базы знаний

Этап оптимизации рассматривает правила во взаимодействии между собой, происходит оптимизация базы знаний в целом. На данном этапе рассматриваются такие качества базы знаний как непротиворечивость и полнота.

Противоречивость

Важнейшим этапом оптимизации базы знаний является исключение противоречивых правил. Исключение проводится на основе подсчета рейтинга правил. Достоинством предложенного решения является то, что поиск групп противоречивых правил осуществляется с помощью синтаксического анализа, что позволяет избавиться от рассмотрения конкретных значений, связей между ними и привязки к конкретной предметной области.

Полнота

Анализ полноты формальной системы в случае нечетких логик интересен возможностью количественно оценивать степень полноты для построенной модели. Степень полноты можно истолковать как оценку качества созданной базы знаний, и в случае низких показателей, служить указанием к ее перепроектированию.

Определение полноты базы знаний заключается в определении некого “предела” совпадения различных характеристик истинности. Для этого определяется супремум всех выводов возможных посылок и инфимум всех возможных заключений по каждому правилу.

Формальная логическая система является полной, если формула (9) верна для любого $X \subseteq F_j$ и формулы $A \in F_j$.

$$C^{syn}(X)(A) = C^{sem}(X)(A). \qquad (9)$$

Описание системы

Реализованная система состоит из 7 основных модулей:

- модуль управления (меню);
- модуль чтение обучающей выборки;
- модуль кластеризации входных данных;
- модуль расчета коэффициента принадлежности;
- модуль генерации базы знаний;
- модуль оптимизации базы знаний;
- модуль нечеткого логического вывода.

На рис. 1 представлен результат работы программы – сформированная база знаний.

После того как база знаний построена, пользователь может ее сохранить или перейти к форме оценке качества построенной базы знаний, представленной на рис. 2.

Моделирование процесса рассуждений эксперта

Результатом работы автоматизированной системы построения и оптимизации баз знаний является набор правил, отражающий закономерности предметной области. Кроме того, система позволяет проанализировать ход построения такого набора, то есть рассмотреть, как бы рассуждал эксперт при построении аналогичной базы знаний.

В процессе построения правил эксперт не оперирует такими понятиями как нечеткое множество, функция принадлежности и интерпретация нечетких операций, однако с их помощью можно объяснить, каким образом была получена та или иная база знаний. Для анализа полученных результатов необходимо рассмотреть, как интерпретируются данные понятия на уровне предметной области.

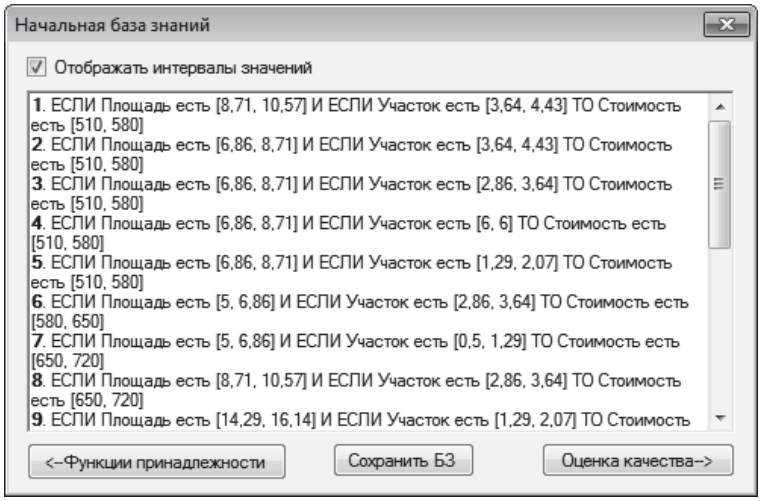

Рис. 1. База правил

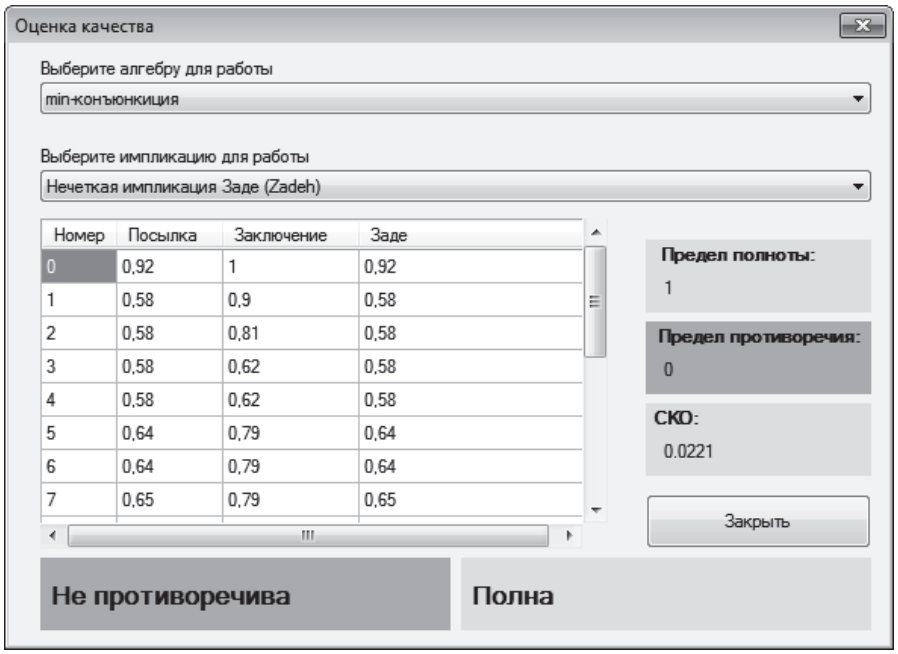

Рис. 2. Оценка качества построения

Определение нечетких множеств

Задание нечетких множеств на области входных данных включает в себя решение о количестве нечетких множеств и определение принадлежности элементов к тому или иному множеству. Таким образом, определение количества нечетких множеств тесно связано с понятием лингвистической переменной. Количество множеств показывает, сколько определений типа “маленькая”, “средняя” или “холодный”, “теплый”, необходимо эксперту для создания набора правил. Набор конкретных членов множества определяет “ширину” распространения лингвистической переменной (понятий типа “теплый”) на числовые данные, то есть интервал каких значений эксперт бы отнес к конкретной переменной (понятию “теплый”).

В разработанной автоматизированной системе данное разбиение реализовано на основе алгоритмов кластеризации, основная цель которых – определение групп значений схожих между собой по некоторому критерию.

Вид функции принадлежности

Под физическим смыслом функции принадлежности $\mu_A(x)$ нечеткого множества A понимается вероятность того, что лицо принимающее решение отнесет элемент x к множеству A.

С другой стороны вид функции принадлежности показывает уверенность, с которой эксперт относит данные к тому или иному множеству.

В рассматриваемой автоматизированной системе вид функции принадлежности подстраивается к предметной области с помощью настройки вида кривой функции принадлежности.

Интерпретация нечетких операций

Интерпретация нечетких операций определяет операции на нечетких множествах. Операции на нечетких множествах используются в проекте на этапе оптимизации базы знаний, а именно на этапе сокращения числа правил и агрегации значений посылки импликации. С точки зрения эксперта вид интерпретации определяет меру оптимизма при определении результата конъюнкции нескольких значений функции принадлежности.

Таким образом, пользователь автоматизированной система построения и оптимизации баз знаний кроме набора продукционных правил может получить качественную характеристику процесса вывода, что в дальнейшем может позволить ему применять полученные знания в качестве эксперта данной предметной области.

Заключение

В данной статье представлена теоретическая основа, позволяющая сформировать решение задачи перехода от данных к знаниям. На основе предложенных алгоритмов была разработана автоматизированная система, реализующей описанные принципы.

СПИСОК ЛИТЕРАТУРЫ

1. *Мурашко А.Г., Шевченко И.В.* Извлечение знаний из баз данных при помощи нейронной сети и нечеткого интерпретатора // Сб. научных трудов/Кременчугский университет экономики, информационных технологий и управления. 2008. Вып. 5. С. 41-44.
2. *Леоненков А.В.* Нечеткое моделирование в среде MATLAB и fuzzyTECH. СПб.: БХВ-Петербург, 2009. 736 с.
3. *Болдырев М.В.* Решение задач с применением нечеткой логики // Энергосбережение, автоматизация в промышленности, интеллектуальные здания и АСУТП. 2010. Вып. 5. С. 5-7.
4. *Заде Л.А.* Понятие лингвистическое переменной и его применение к принятию приближенный решений [пер с англ. под ред. Аверкина А.Н.]. М.: Горячая линия – ТелекоФИЗМАТЛИТ, 2009. 252 с.
5. *Кофман А.* Введение в теорию нечетких множеств [пер с франц.]. М.: Радио и связь, 2007. 432 с.
6. Нейронные сети, генетические алгоритмы и нечеткие системы / *Д. Рутковская, М. Пилинский, Л. Рутковский.* М.: Горячая линия – Телеком, 2006. 452 с.
7. Нечеткие модели и сети / *В.В. Борисов, В.В. Круглов, А.С. Федулов.* М.: Горячая линия – Телеком, 2007. 484 с.
8. *Мандель И.Д.* Кластерный анализ. М.: Финансы и статистика, 2010. 176 с.
9. *Кириллов А.В.* Социально-экономическая статистика: учебное пособие для вузов. Самара.: МИР, 2011.72 с.
10. Математические принципы нечеткой логики / *В.Новак, И.Перфильева, И.Мочкорж* [пер с англ.; под ред. Аверкина А.Н.]. М.: Горячая линия – ТелекоФИЗМАТЛИТ, 2006. 252 с.
11. *Батыршин И.З.* Основные операции нечеткой логики и их обобщение. Казань: Отечество, 2008. 100 с.

KNOWLEDGE EXTRACTION FROM LARGE DATA SETS

© 2014 I.A. Lyozin, D.E. Markelov

Samara State Aerospace University named after Academician S.P. Korolyov
(National Research University)

Knowledge is a collection of information and deduction rules of the world, properties of objects, patterns of processes and events, and also rules of using them to make decisions. The main difference between knowledge and data is their structure and activity. Appearance of new facts in the database, or determination of new connections can be a source of changes in decision-making. During their work many research departments and commercial companies accumulate a large array of facts, figures and measurements. Experts can't analyze all this information. The considered approach allows analysis of current situation, establishes the relationship between indicators and creates rules of influence for each other. Method was created for extracting knowledge from large data sets. A method is a procedure, which can solve the problem of the transition from data to knowledge. Developed method is designed to improve the effectiveness of project management R & D in aerospace applications.

Keywords: knowledge extraction, large data sets, fuzzy logic.

Ilya Lyozin, Ph.D., Assoc. Prof. of Department "Information Systems and Technologies". E-mail: ilyozin@yandex.ru
Dmitry Markelov, First-Year Student of the Faculty of Computer Science Graduate. E-mail: rtg7@yandex.ru