УДК 004.02

ИЗВЛЕЧЕНИЕ ЗНАНИЙ ИЗ БАЗ ДАННЫХ С ПОМОЩЬЮ НЕЙРО-НЕЧЕТКОЙ МОДЕЛИ

канд. техн. наук, доц. А.Ф. ОСЬКИН
(Полоцкий государственный университет);
Д.А. ОСЬКИН
(Белорусский государственный экономический университет)

Кратко рассматриваются технологии извлечения знаний и типы извлекаемых знаний. Предлагается метод извлечения знаний, представляемых в виде наборов ассоциативных правил «ЕСЛИ…, ТО…». Метод реализуется с помощью нейро-нечеткого моделирования предметной области. Описываются приемы построения нечеткой логической модели анализируемой предметной области и способы применения нейронных сетей для выделения нечетких правил. Рассмотрено программное обеспечение для построения нечеткой модели. Описывается алгоритм извлечения знаний из анализируемых баз данных.

Ключевые слова: *базы данных, базы знаний, извлечение знаний, нейро-нечеткое моделирование.*

Введение. Современный этап развития вычислительной техники и информационных технологий характеризуется повсеместным использованием баз данных (БД) оперативной информации. Органы государственной власти и управления, промышленные предприятия, организации систем образования и здравоохранения создают и ведут собственные БД на разных уровнях управления. Имеющиеся БД содержат огромные объемы скрытой аналитической информации, извлечение которой позволяет выявлять тенденции, находить новые решения, строить стратегию развития. Таким образом, разработка методов и алгоритмов извлечения знаний из БД является важной и актуальной задачей.

Существует множество определений понятия «извлечение знаний». Однако чаше всего используется определение, сформулированное одним из пионеров этой области знаний Г. Пиатецким-Шапиро [1].

Извлечение знаний из БД (knowledge discovery from databases) или добыча данных (data mining) – это процесс извлечения скрытой, ранее неизвестной и потенциально полезной информации (правил, ограничений, закономерностей) из сырых данных БД.

С помощью технологий data mining можно извлекать важные и полезные знания, зависимости, правила или другую высокоуровневую информацию из подходящих наборов данных в БД. Причем наиболее интересным и перспективным является использование больших БД в качестве источников для генерации и проверки знаний. Извлечение знаний именно из больших БД (big data) является сегодня ключевым направлением обработки данных.

Извлекаемые знания могут иметь различные представления. Традиционно выделяют следующие типы извлекаемых знаний.

Ассоциативные правила [2] – зависимости вида A_1^...^A_m=>C_1^...^C_n, где A_i и C_j для всех $i \in \{1, ...,m\}$ и $j \in \{1,…,n\}$ – набор значений атрибутов из важных множеств БД. Например, на большом множестве данных о покупках в продовольственном магазине может быть найдено такое правило: если покупатель покупает творог, он обычно также покупает и сметану.

Классификация данных [3] – разделение данных на подмножества на основании значений определенных атрибутов. Такой подход, например, может использоваться крупными торговыми компаниями для классификации своих покупателей по их предпочтениям и последующей рассылки им целевой рекламы и каталогов определенной тематики.

Кластеризация данных [4] – это разбиение данных на группы без предварительного задания классифицирующих атрибутов. Такое разбиение базируется на принципе максимальной близости объектов внутри класса и минимальной между классами. Например, кластерный анализ кредитных историй позволяет менеджерам банка выявить кластеры клиентов с различными уровнями риска невозврата кредита, и принять правильное решение при работе с ними.

В настоящей работе мы сосредоточимся на рассмотрении методов и алгоритмов нахождения ассоциативных правил. При этом нами будет использоваться нейро-нечеткая модель извлечения знаний.

Нейро-нечеткая модель предметной области.

Нейро-нечеткая модель предметной области основывается на математической дисциплине, получившей название «Нечеткая логика». Главная идея нечеткой логики заключается в том, что способ рассуж-

дений, опирающийся на естественный язык человеческого общения, не может быть описан с помощью традиционных математических формул. Все, что связано с использованием естественного языка имеет многозначную интерпретацию, в то время как формальному подходу присуща строгая однозначность.

Математический аппарат нечеткой логики использует понятие «лингвистическая переменная», позволяющее перейти от четких высказываний к нечеткому описанию предметной области.

Создатель современной концепции нечеткой логики профессор Л. Заде писал: «В большинстве основных задач, решаемых человеком, не требуется высокая точность. Человеческий мозг использует допустимость такой неточности, кодируя информацию, достаточную для задачи, элементами нечетких множеств, которые лишь приближенно описывают исходные данные… Способность оперировать нечеткими множествами и вытекающая из нее способность оценивать информацию является одним из наиболее ценных качеств человеческого мозга, которое фундаментальным образом отличает человеческий разум от так называемого машинного разума, приписываемого существующим вычислительным машинам» [5].

Сущность концепции нечеткой логики, предложенной Л. Заде, сводится к трем основным положениям:

1) вместо числовых переменных используются нечеткие лингвистические переменные;
2) простые отношения между переменными определяются с помощью нечетких высказываний;
3) сложные отношения описываются нечеткими алгоритмами.

Общая схема нечеткой логической модели представлена на рисунке 1.

Фазификатор преобразует четкие значения входных переменных x в соответствующие лингвистические переменные $A(x)$. Машина нечеткого вывода на основании правил, хранящихся в нечеткой базе знаний, формирует выходную лингвистическую переменную C, которая после дефазификации преобразуется в четкую выходную переменную y. Таким образом, для построения модели необходимо решить следующие задачи:

1. Задать для всех входных переменных семейство функций принадлежности, с помощью которых выполняется переход от четких значений к лингвистическим переменным.
2. Сформировать нечеткую базу знаний.
3. Задать для выходной переменной функции принадлежности, чтобы выполнить обратный перехода от лингвистической переменной к четкому значению.

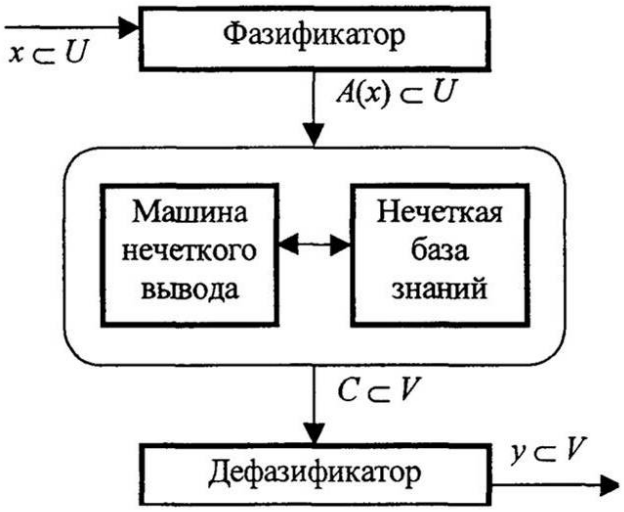


Рисунок 1. – Схема нечеткой логической модели

Нечеткая база знаний представляет собой набор ассоциативных правил вида «ЕСЛИ…, ТО…», извлечение которых из имеющихся БД и является целью нашей работы. Достижение этой цели становится возможным путем изменения точки зрения на построенную нечеткую логическую модель.

Если рассматривать модель как самоорганизующуюся нейронную сеть Кохонена, то посредством обучения сети можно выявить множество правил, образующих нечеткую базу знаний. При этом входные переменные интерпретируются как входы сети Кохонена, а выходным ячейкам сети соответствуют пред-

посылки нечетких правил. В зависимости от того, как далеко выходной нейрон y_i находится от нейрона-победителя для данного входного вектор x, определяется степень $\mu_i(x)$, с которой x удовлетворяет предпосылке соответствующего правила.

Программное обеспечение для построения нейро-нечеткой модели.

Одним из самых популярных приложений для работы с нечеткой логикой является приложение fuzzyTECH [6, 7]. Оно состоит из 4-х модулей визуального редактора, в котором строится схема проектируемой модели, редактора переменных, служащего для задания лингвистических переменных, их термов и функций принадлежности, редактора правил, в котором создается и хранится база знаний системы и модуля анализа, выполняющего отладку и собственно моделирование. Последний модуль снабжен широким набором инструментальных средств для визуализации результатов моделирования, формирования отчетов и т.д.

На рисунке 2 представлен экран редактора переменных с заданными функциями принадлежности для некоторой лингвистической переменной *He*. Эта переменная состоит из трех термов *{low, medium, high}*, каждый из которых задается своей функцией принадлежности.

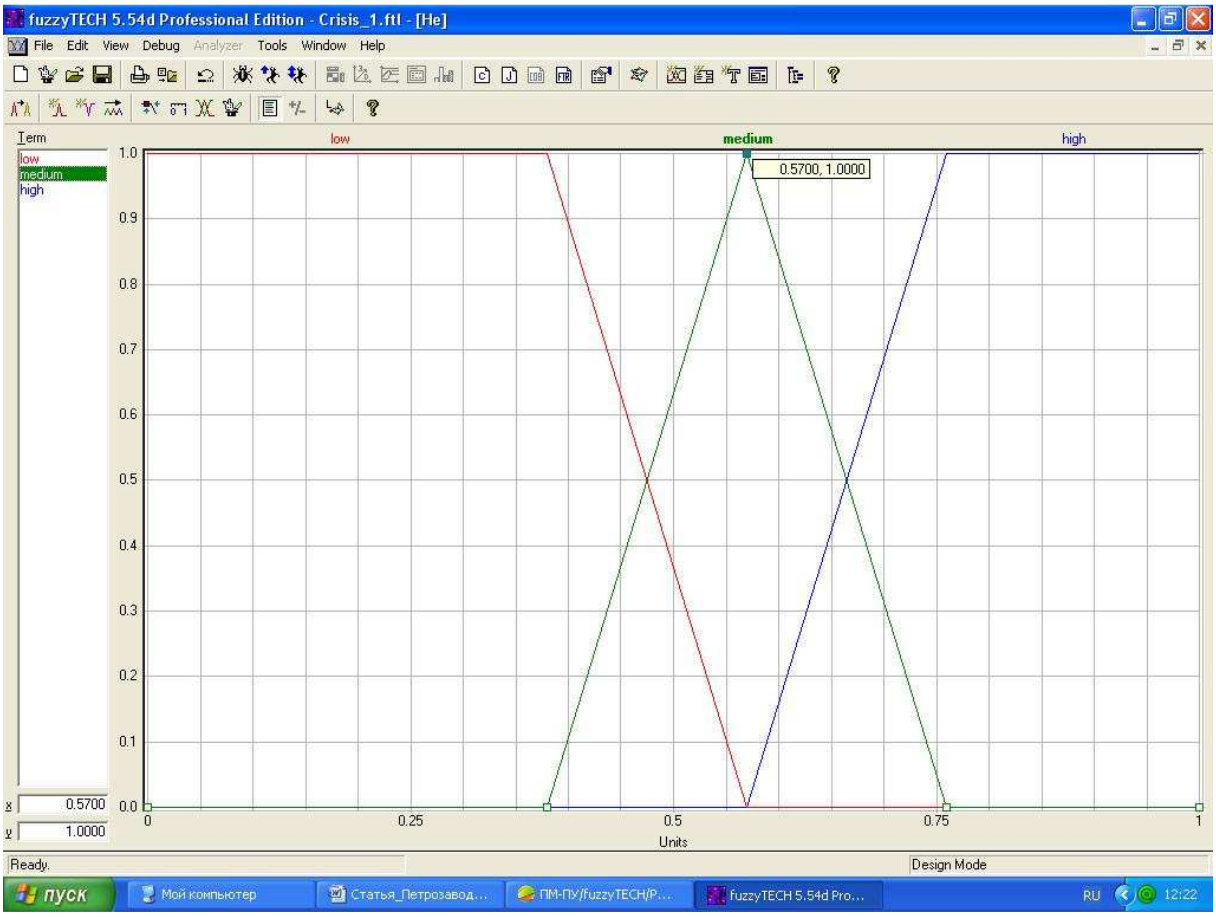
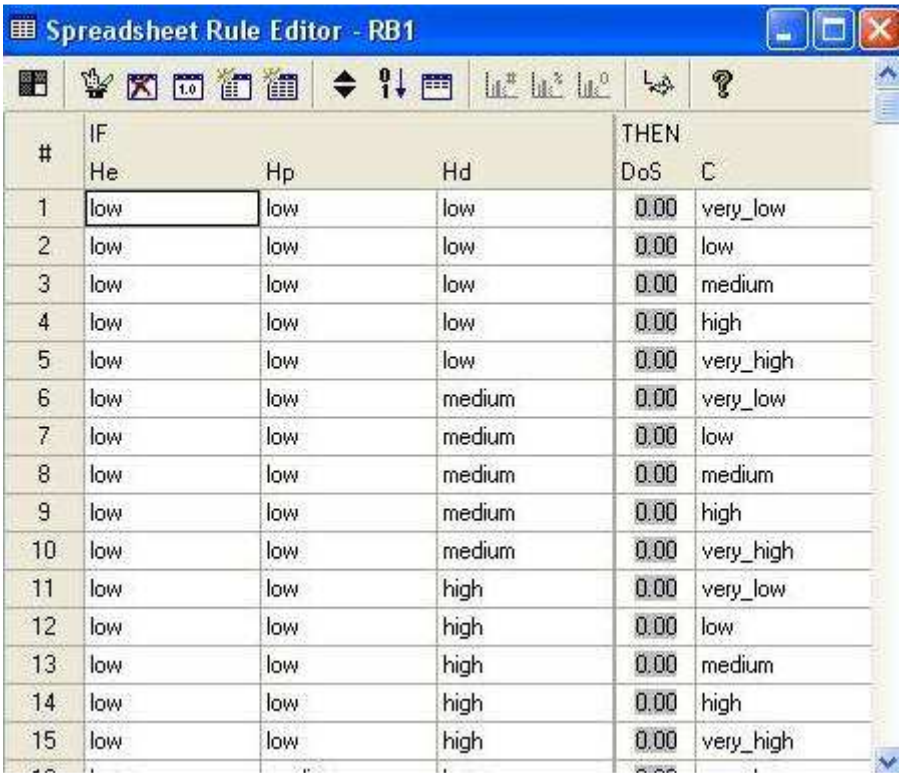


Рисунок 2. – Редактор переменных

Приложение fuzzyTECH снабжено надстройкой NeuroFuzzy, с помощью которого собственно и выполняется извлечение скрытых правил. Извлечение правил осуществляется путем обучения модели на данных из анализируемой БД с использованием нейросетевых технологий. В результате уточняются веса каждого из правил, входящих в базу знаний, и формируется ее окончательная редакция. Функции принадлежности выходной переменной задаются по аналогии со входными переменными. На рисунке 3 показан фрагмент базы знаний системы.

В данном примере на вход модели подается вектор, состоящий из трех компонент, лингвистических переменных: *He*, *Hp*, *Hd*, каждая из которых имеет три терма *{low, medium, high}*. Выходная переменная *C* имеет пять термов *{very low, low, medium, high very high}*. Исходно в системе автоматически генерируется полный набор всех возможных сочетаний значений входных и выходной переменных. После этого выполняется обучение модели, результатом которого является уточнения веса каждого прави-

ла. Правила с нулевым весом исключаются из базы знаний, а наиболее значимые правила получают наибольшие веса. Таким образом формируется база знаний модели и выполняется извлечение знаний из анализируемой БД.



#	IF			THEN	
	He	Hp	Hd	DoS	C
1	low	low	low	0.00	very_low
2	low	low	low	0.00	low
3	low	low	low	0.00	medium
4	low	low	low	0.00	high
5	low	low	low	0.00	very_high
6	low	low	medium	0.00	very_low
7	low	low	medium	0.00	low
8	low	low	medium	0.00	medium
9	low	low	medium	0.00	high
10	low	low	medium	0.00	very_high
11	low	low	high	0.00	very_low
12	low	low	high	0.00	low
13	low	low	high	0.00	medium
14	low	low	high	0.00	high
15	low	low	high	0.00	very_high

Рисунок 3. – Фрагмент базы знаний системы

Алгоритм извлечения знаний из БД.

Перечислим теперь все шаги алгоритма извлечения знаний из БД.

1. Над сырыми данными, хранящимися в различных источниках, относящихся к анализируемой предметной области, выполняются процедуры извлечения, преобразования и очистки данных. После этого очищенные и преобразованные данные загружаются в хранилище данных. Хранилище данных – это БД специального вида, данные в которой подготовлены для дальнейшей обработки и анализа.
2. В зависимости от целей анализа выделяются входные и выходные переменные, определяются соответствующие им лингвистические переменные, задаются термы лингвистических переменных и их функции принадлежности.
3. На основе полученных лингвистических переменных строится нечеткая модель предметной области. Формируется исходная база знаний системы, представляющая собой набор правил для всех возможных сочетаний значений входных и выходных переменных.
4. Выполняется обучение модели на данных из построенного хранилища данных. В результате обучения уточняются веса каждого правила, входящего в базу знаний и формируется ее окончательная редакция.
5. Сформированная база знаний представляет собой конечный продукт анализа, позволяющий судить о скрытых, неявных и неочевидных закономерностях, имеющих место в анализируемой системе.

Выводы.

1. На современном этапе развития вычислительной техники и информационных технологий извлечение знаний из существующих БД является важной и актуальной задачей.

2. Нейро-нечеткая модель предметной области может быть эффективно использована для извлечения неявных, неочевидных и практически полезных знаний из БД.

3. Извлечение знаний производится путем построения нейро-нечеткой модели предметной области с последующим обучением ее на существующих наборах данных из обрабатываемой БД.

4. Результатом обучения является база знаний, представленная в виде ассоциативных правил вида «ЕСЛИ…, ТО…».

ЛИТЕРАТУРА

1. Piatetsky-Shapiro, G. Knowledge Discovery in Databases / G. Piatetsky-Shapiro, W.J. Frawley. – AAAI/MIT Press, 1991.
2. Wittern, I.H. Data Mining: practical machine learning tools and techniques / I.H. Wittern, E. Frank. – 2 ed. – Elsevier, 2016.
3. Aggarwal, C. Data classification: algorithms and applications / C. Aggarwal. – CRC Press, 2014.
4. Флах П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных / П. Флах. – М. : ДМК, 2017.
5. Kasko, B. Neural networks and Fuzzy Systems / B. Kasko. – NY. : Printice – Hall, 1992.
6. fuzzyTECH®INFORM GmbH. – Режим доступа http://www.fuzzytech.com/. – Дата доступа: 13.06.2018.
7. Von Altrock, C. Fuzzy Logic and NeuroFuzzy Application Explained / C. Von Altrock. – Prentice – Hall, 1995.

Поступила 24.09.2018

EXTRACTING KNOWLEDGE FROM DATABASES THROUGH NEURO-FUZZY MODEL

A. OSKIN, D. OSKIN

Briefly discusses the technology of extraction of knowledge and types of extracted knowledge. A method is proposed for extracting knowledge represented as sets of association rules "IF ..., THEN ...". The method is implemented using neural-fuzzy modeling of the subject area. We describe the techniques for constructing a fuzzy logical model of the analyzed domain and how to use neural networks to highlight fuzzy rules. Considered software for constructing a fuzzy model. An algorithm for extracting knowledge from the databases being analyzed is described.

Keywords: *databases, knowledge bases, knowledge extraction, neuro-fuzzy modeling.*