УДК 681.3

Алгоритм и программный комплекс редукции баз знаний мягких экспертных систем

Абдулхаков А. Р.

Казанский национальный исследовательский технический университет им. А.Н. Туполева, КНИТУ-КАИ, ул. Карла Маркса, 10, Казань, 420111, Россия

e-mail: aidar_abdulhakov@mail.ru

Аннотация

Рассматриваются вопросы повышения эффективности использования экспертных систем за счет структурной оптимизации их баз знаний методами нечеткой логики и кластерного анализа. Предлагается алгоритм редукции нечетко-продукционных правил Такаги-Сугено. На примере таксономии базы знаний системы фильтрации нежелательных почтовых сообщений производится анализ эффективности предложенного подхода для оптимизации баз знаний экспертных систем.

Ключевые слова: база знаний, экспертная система, редукция нечетких правил, оптимизация баз знаний экспертных систем

Введение

В настоящее время во многих сферах человеческой деятельности, таких как экономика, медицина, промышленность, для решения сложных практических задач большое распространение получили экспертные системы. Главной функцией

данных систем является поддержка принятия решений, осуществляемая на основе накопленной базы знаний и механизма логического вывода. При этом сам процесс накопления и формализации знаний носит неоднозначный и, как правило, нетривиальный характер.

Существует два основных подхода к получению знаний для экспертных систем [1]: непосредственное извлечение у эксперта или другого источника знаний, а также их формирование с использованием методов и алгоритмов интеллектуального анализа данных. Первый подход требует большой аналитической работы человека-эксперта, которому бывает трудно, а порой и невозможно изложить свои знания, опыт и интуицию в рамках строгих формальных моделей представления знаний. Второй подход к получению знаний привлекает разработчиков и исследователей своей способностью автоматически извлекать знания из данных, производить их оценку и использовать в базах знаний экспертных систем.

Очевидно, что при наличии репрезентативных данных использование второго подхода более предпочтительно, поскольку помимо сокращения времени и упрощения всего процесса получения знаний эксперт может участвовать в процессе оценивания сформированных правил и закономерностей, которые согласуются с его личными знаниями, опытом и интуицией.

Однако при этом, несмотря на все достоинства данного подхода, в процессе формирования знаний часто генерируются сотни и тысячи правил, составляющих базу знаний, что, с одной стороны, усложняет работу эксперта по ее анализу и интерпретации, а, с другой, делает правила базы знаний избыточными и часто

противоречивыми. Все это в последствие может существенно повлиять на точность и скорость принимаемых экспертной системой решений, сводя на нет преимущества процесса формирования знаний.

Для повышения эффективности использования экспертных систем необходимо производить оптимизацию баз знаний за счет структурного упорядочивания и минимизации правил принятия решений без потери их полноты и непротиворечивости.

Данная задача впервые была сформулирована в [2], как задача таксономии знаний. Однако, ее практические реализации стали появляться лишь в последние несколько лет. Так, в работе [3] предложен метод структурно-параметрической оптимизации баз знаний нечетких экспертных систем, основанный на преобразовании базы знаний в нечеткую нейронную сеть и ее параметрической оптимизации с использованием генетического алгоритма. В работах [4,5] задача кластеризации знаний в системах искусственного интеллекта решается с применением муравьиных алгоритмов.

Однако, несмотря на положительные результаты имеющихся решений, проблема редукции (сокращения числа правил) баз знаний остается актуальной и мало изученной. В данной статье для оптимизации баз знаний экспертных систем предлагается подход, основанный на решении задачи таксономии в пространстве знаний с применением методов нечеткой логики и кластерного анализа. Для однозначного решения поставленной задачи в качестве объекта исследования выбрана одна из часто используемых в экспертных системах моделей представления знаний – нечетко-продукционная модель Такаги-Сугено [6].

1 Постановка задачи оптимизации базы знаний

Пусть для формирования базы знаний используется нечеткая нейронная сеть *ANFIS* [7]. Процесс формирования может быть реализован в среде моделирования *MathLab*. При этом требуется обучить сеть, указав число входных параметров сети и их нечетких градаций. Однако уже при 5 входах и 4 градациях обученная сеть формирует большое количество правил, составляющее величину n^P, где P – количество входных переменных модели, а n – число их нечетких градаций (в данном случае 1024 правила). Очевидно, что такое количество правил для решения большинства задач является избыточным. Поэтому требуется оптимизация автоматически сформированной при помощи нечеткой нейронной сети базы знаний, редукция (сокращение количества) имеющихся в ней правил. Рассмотрим формальную постановку данной задачи.

Пусть имеется автоматически сформированная в процессе обучения сети *ANFIS* база знаний $R=\{R_1, R_2,\dots , R_N\}$, где R_i (i=1..N) – нечетко-продукционные правила Такаги-Сугено, N – исходное количество правил в базе знаний. При этом предполагается, что можно сформировать k подмножеств правил ($k<N$), сгруппированных по похожести, т.е. примерно одинаково описывающих закономерности в анализируемых данных.

Обозначим $N_1, N_2, \dots, N_k$ – количество правил в каждом из подмножеств. Тогда, выделяя типичного представителя $R\text{э}_j$ (j=1..k) – эталонное правило в каждом подмножестве, получим новое множество правил $\{R\text{э}_1, R\text{э}_2,\dots , R\text{э}_k\}$, составляющих оптимизированную базу знаний экспертной системы. Остальные правила при этом отбрасываются (см. рис. 1).

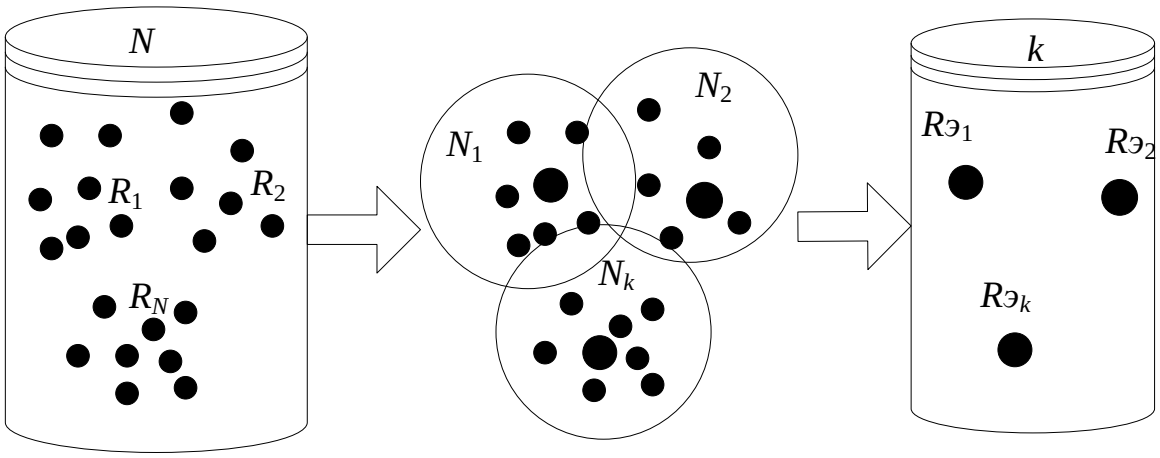

Рисунок 1. Иллюстрация задачи оптимизации базы знаний

Из рисунка видно, что в оптимизированную базу знаний войдут только те правила, которые являются центрами сформированных кластеров. Таким образом, задача редукции нечетких правил сводится к задаче кластеризации, позволяющей объединять схожие правила принятия решений в один кластер. Искомая база знаний формируется из типичных представителей каждого кластера.

Математическая формализация описанной процедуры может быть получена при использовании метрик, позволяющих определять расстояния между антецедентами нечетких правил. В качестве метрики может быть использовано расстояние Евклида. При этом производится вычисление и сравнение расстояний между всеми парами нечетких антецедентов. Дальнейшая процедура основана на решении задачи кластеризации нечетких правил, в результате которой формируется оптимальное кластерное решение.

2 Решение задачи кластеризации нечетких правил

Рассмотрим нечетко-продукционное правило, представленное в виде модели Такаги-Сугено [6]:

$$ЕСЛИ\ x_1\ есть\ A_1\ И\ x_2\ есть\ A_2\ И\ \dots\ x_n\ есть\ A_n\ ТО\ y=f(x_1,\dots,x_n),$$

где $x_1,\ldots,x_n$ – входные лингвистические переменные, $A_1,\ldots,A_n$ – их нечеткие значения, y – четкая переменная выхода, $f(x_1,\ldots,x_n)$ – вещественная функция от четких аргументов $x_1,\ldots,x_n$.

Для кластеризации такого рода нечетких правил необходимо производить оценку «похожести» их антецедентов при одинаковых значениях консеквентов. Данная задача становится выполнимой при одновременном выполнении следующих условий:

1) существует эффективный способ сравнения нечетких антецедентов;

2) число различных значений консеквентов нечетких правил при любых значениях аргументов $x_1,\ldots,x_n$ конечно и счетно.

Первое условие требует не просто введение метрики расстояний в пространстве знаний, позволяющей определять «близость» двух нечетких правил, но и эффективного способа представления антецедента в формализованном виде, пригодном для использования в алгоритме кластеризации.

Второе условие накладывает ограничение на вид функции $f(x_1,\ldots,x_n)$, требуя дискретности ее значений. В случае решения задачи классификации (особенно бинарной) данное требование легко выполняется. При этом значениями данной функции являются константы, указывающие на класс объекта.

С учетом сформулированных условий и введенных ограничений, рассмотрим решение задачи кластеризации правил следующего вида:

$$ЕСЛИ\ x_1\ есть\ A_1\ И\ x_2\ есть\ A_2\ И\ \ldots\ x_n\ есть\ A_n\ ТО\ y=C_i\ ,$$

где C_i – метка некоторого класса.

Рассмотрим способ формирования начальных подмножеств правил. Для этого разобьем исходную базу знаний на непересекающиеся подмножества по значению метки класса консеквента в правилах:

$$R=\{R_1, R_2,\dots, R_N\}=\{Rul_1|y=C_1\}\cup\{Rul_2|y=C_2\}\cup\dots\cup\{Rul_{\mathrm{m}}|y=C_m\},$$

где Rul_l – подмножество правил с консеквентом C_l, l=1..m.

Процесс кластеризации будет производиться независимо в каждом подмножестве правил путем объединения их антецедентов в кластеры.

В качестве примера рассмотрим следующее подмножество правил:

$$Если\ x_1\ есть\ {}^{A_{11}}\ И\ x_2\ есть\ {}^{A_{12}}\ И\dots\ x_{\mathrm{n}}\ есть\ {}^{A_{1\mathrm{n}}}\ То\ y=1$$

$$Если\ x_1\ есть\ {}^{A_{21}}\ И\ x_2\ есть\ {}^{A_{22}}\ И\dots\ x_{\mathrm{n}}\ есть\ {}^{A_{2\mathrm{n}}}\ То\ y=1$$

...

$$Если\ x_1\ есть\ {}^{A_{m1}}\ И\ x_2\ есть\ {}^{A_{m2}}\ И\dots\ x_{\mathrm{n}}\ есть\ {}^{A_{mn}}\ То\ y=1$$

Представим каждое из правил вектором своих нечетких ограничений. Тогда система правил примет вид:

$$\{({}^{A_{11}}, {}^{A_{12}}, \dots, {}^{A_{1\mathrm{n}}}), ({}^{A_{21}}, {}^{A_{22}}, \dots, {}^{A_{2\mathrm{n}}}), \dots, ({}^{A_{m1}}, {}^{A_{m2}}, \dots, {}^{A_{mn}})\}.$$

Перейдя от нечетких множеств ${}^{A_{ij}}$ (i=1..m, j=1..n) к их четким аналогам x_{ij}, используя процедуру дефаззификации по методу центра тяжести, получим:

$$\{(x_{11}, x_{12},\dots, x_{1n}), (x_{21}, x_{22},\dots, x_{2n}),\dots, (x_{m1}, x_{m2},\dots, x_{mn})\}.$$

Полученную систему векторов можно рассматривать, как множество точек в n-мерном Евклидовом пространстве, каждая из которых является результатом формализованного представления антецедента соответствующего нечеткого

правила. Таким образом, таксономия нечетких правил фактически производится путем объединения данных точек в локальные кластеры.

Однако, так как в общем случае значения входных параметров нечетких правил измерены в разных шкалах, то, прежде чем непосредственно приступать к поиску оптимального кластерного решения, необходимо произвести процедуру нормировки дефаззифицированных значений, используя метрику вида:

$$x' = \frac{x - x^{*}}{x^{**} - x^{*}},$$

где x – исходное значение параметра;

x^{*} – минимальное значение;

x^{**} – максимальное значение;

x' – нормированное значение.

Результатом данной процедуры является множество точек в нормированном n-мерном пространстве:

$$\{(x'_{11}, x'_{12}, \ldots, x'_{1n}), (x'_{21}, x'_{22}, \ldots, x'_{2n}), \ldots, (x'_{m1}, x'_{m2}, \ldots, x'_{mn})\}.$$

Таким образом, получили готовые данные, пригодные для кластеризации и поиска оптимального кластерного решения.

Формирование конечных подмножеств правил базы знаний (кластеров) основано на использовании алгоритма k-средних. Важным моментом при этом является выбор оптимального числа кластеров k, на которое разбивается исходное подмножество правил. В задаче оптимизации базы знаний критерием оптимальности может служить ошибка обобщения, получаемая экспертной системой при ее работе на тестовой выборке данных:

¿ ,

где $N_{прав}$ – количество правильно классифицированных примеров, $N_{общ}$ – общее количество примеров.

Минимальное значение ошибки обобщения соответствует оптимальному кластерному решению.

3 Анализ эффективности оптимизации базы знаний

Проведем анализ эффективности предложенного в работе подхода на примере оптимизации базы знаний системы фильтрации нежелательных почтовых сообщений – «спама». В работе [8] предложен подход к решению данной задачи путем формирования базы знаний на основе обучения нечеткой нейронной сети *ANFIS*. Задача решается достаточно эффективно в системе моделирования *MathLab* при условии наличия репрезентативной обучающей выборки, состоящей из значений таких параметров писем, как частота встречаемости слов верхнего регистра, частота встречаемости цифр в письме, количество различных цветов в тексте письма, размер письма в килобайтах.

После подготовки и загрузки обучающих данных генерируется нечеткая нейронная сеть, состоящая из четырех входных нейронов с тремя нечеткими градациями и одного выходного нейрона. Обученная нейронная сеть генерирует базу знаний, состоящую из восьмидесяти одного правила.

Для автоматизации решения задачи кластеризации нечетких правил разработано приложение «Оптимизация баз знаний экспертных систем», интегрированное со средой моделирования *MathLab*. На рисунке 2 представлен пример работы данного приложения.

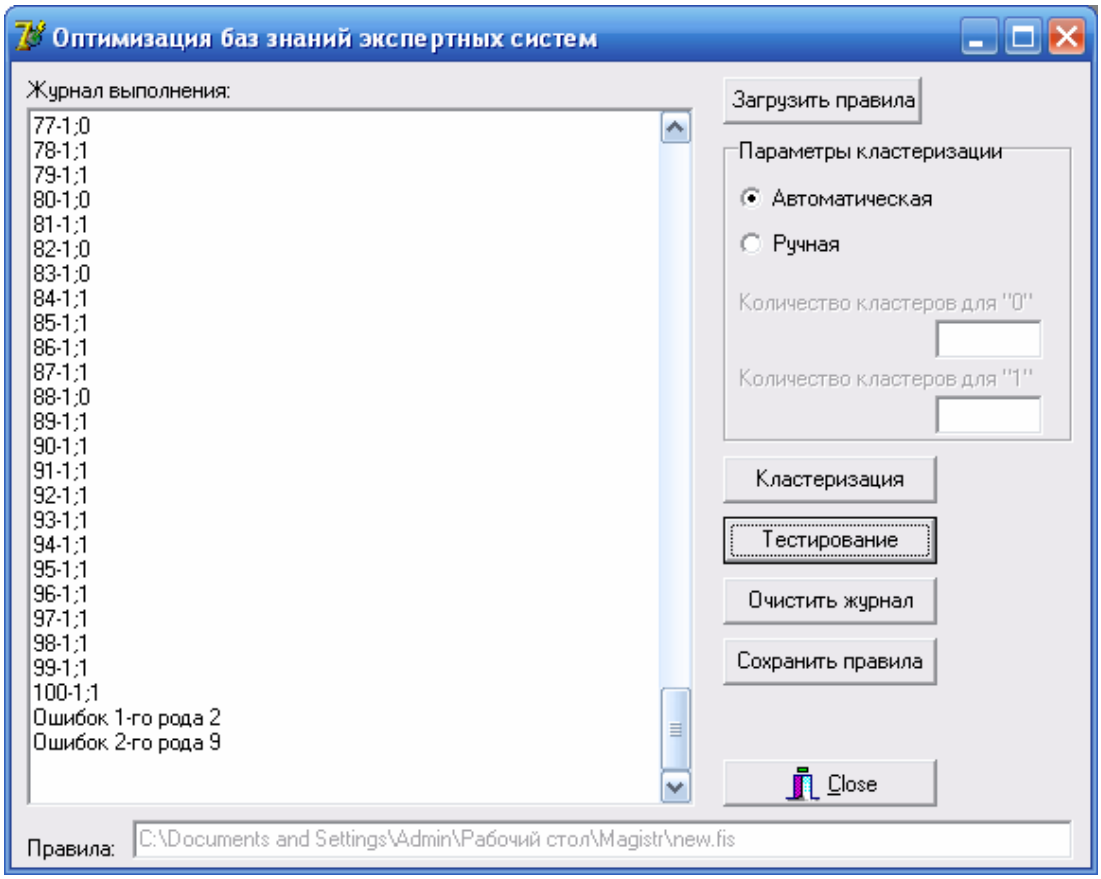

Рисунок 2. Пример работы системы оптимизации базы знаний

Система позволяет загружать автоматически сформированные правила и производить их кластеризацию, используя алгоритм *k*-средних. После поиска оптимального кластерного решения формируется новая база знаний с меньшим количеством правил, представляющих собой центры кластеров.

Для подтверждения априорного предположения о повышении эффективности работы системы «спам»-классификации после оптимизации ее базы знаний произведена серия экспериментов по оценке классифицирующей способности системы на тестовой выборке данных. Для тестирования случайным образом выбирались наборы данных, характеризующие как «спамовые», так и обычные почтовые сообщения. Объем тестовой выборки составил 100 записей (по 50 для каждого класса сообщений).

Эффективность системы характеризуется двумя типами ошибок:

1) ошибкой первого рода – ложный пропуск «спама», то есть неверное отнесение «спамового» письма к «не спаму»;

2) ошибкой второго рода – ложное срабатывание, т.е. неверное отнесение обычного письма к категории «спам».

Пусть N_1 – количество попыток классификации «спамовых» почтовых сообщений, n_1 – число ложных пропусков «спама». Тогда, коэффициент ошибок первого рода рассчитывается по формуле:

$$E_1 = \frac{n_1}{N_1} \times 100 \ .$$

Коэффициент ошибок второго рода, соответственно, рассчитывается по следующей формуле:

$$E_2 = \frac{n_2}{N_2} \times 100 \ ,$$

где N_2– количество попыток классификации обычных сообщений;

n_2– число ложных срабатываний.

В таблице 1 представлены сравнительные результаты тестирования экспертной системы «спам»-классификации на исходной базе знаний, состоящей из 81 правила, и оптимизированной базе знаний, включающей 19 правил.

Таблица 1. Результаты тестирования системы

Критерии эффективности База знаний	Количество правил	Коэффициент ошибок первого рода, %	Коэффициент ошибок второго рода, %
Исходная	81	4	18
Оптимизированная	19	2	12

Из таблицы видно, что точность классификации системы на основе оптимизированной базы знаний улучшилась на 8% по сравнению с исходными результатами. Следовательно, оптимизация базы знаний улучшила обобщающую способность экспертной системы при работе на тестовой выборке данных.

Заключение

Естественно считать, что описанный в данной работе подход к оптимизации баз знаний экспертных систем на основе кластеризации нечетких правил не является универсальным. Требуются дополнительные исследования, позволяющие обобщить предложенную концепцию на множество других моделей представления знаний. Однако полученные результаты уже позволяют утверждать, что применение методики оптимизации придает базе знаний экспертной системы следующие важные качества:

- уменьшает объем базы знаний;
- повышает ее интерпретируемость;

– уменьшает неопределенность выбора того или иного правила при принятии решения;
– повышает точность и скорость работы системы.

Таким образом, практическая ценность предложенного подхода заключается в возможности повышения эффективности использования экспертных систем в любой сфере человеческой деятельности.

Библиография

1.Гаврилова Т.А., Хорошевский В.Ф. Базы знаний интеллектуальных систем. –СПб.: Питер, 2001. – 384 с.
2.Загоруйко Н.Г. Прикладные методы анализа данных и знаний. – Новосибирск: Изд-во Ин-та математики, 1999. – 270 с.
3.Бухнин А.В., Бажанов Ю.С. Оптимизация баз знаний экспертных систем с применением нейронных нечетких сетей // Нейрокомпьютеры: разработка, применение. 2007. №11.
4.Щуревич Е.В., Крючкова Е.Н. Моделирование и анализ знаний в системах искусственного интеллекта // Вестник Алтайского гос. технич. ун-та им. И.И. Ползунова. Барнаул, 2007. №2. С. 173-177.
5.Щуревич Е.В. Кластеризация знаний в системах искусственного интеллекта // Информационные технологии. 2009. №2. С. 25-29.
6.Takagi T., Sugeno M. Fuzzy identification of systems and its application to modeling and control // IEEE Transactions, Systems, Man and Cybernetics, 1985. – V. 15. – pp. 116-132.
7.Jang J.R., Sun C.T. ANFIS: Adaptive-Network-based Fuzzy Inference Systems // IEEE Tranc. on Systems, Man and Cybernetics, 1993. – V. 23. – pp. 665-685.
8.Катасёв А.С., Корнилов Г.С. Адаптивная нейронечеткая модель формирования баз знаний экспертных систем в решении задачи фильтрации «спама» // Инфокоммуникационные технологии глобального информационного общества:

сборник трудов 7-й международной научно-практической конференции. Казань, 2009. С. 507-512.