УДК 004.891.3

А. Р. Абдулхаков, А. С. Катасёв, А. П. Кирпичников

МЕТОДЫ РЕДУКЦИИ НЕЧЕТКИХ ПРАВИЛ В БАЗАХ ЗНАНИЙ ИНТЕЛЛЕКТУАЛЬНЫХ СИСТЕМ

Ключевые слова: база знаний, редукция, кластеризация, генетический алгоритм, нечеткое правило.

В работе решается задача повышения эффективности использования интеллектуальных систем за счет редукции нечетких правил в базах знаний. Предлагается два метода редукции: на основе алгоритма кластеризации и на основе генетического алгоритма. На примере редукции базы знаний классификации ирисов показывает эффективность и сравнение предложенных подходов.

Keywords: *knowledge base, reduction, clusterization, genetic algorithm, fuzzy rule.*

We solve the problem of efficiency of intelligent systems at the expense of fuzzy rules reduction in the knowledge bases. Proposed two methods for the reduction: based on a clustering algorithm, and based on the genetic algorithm. On the example of the reduction of the knowledge base iris classification shows the effectiveness and comparison of the proposed approach.

Информационные системы используются во всех сферах человеческой деятельности и к настоящему времени накопили в себе достаточно большие объемы данных. В целях эффективного использования накопленного опыта все большую роль играют технологии извлечения знаний из баз данных и интеллектуальные методы их обработки при построении интеллектуальных систем [6,8]. 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 [2]: извлечение у эксперта и использование инструментов интеллектуального анализа данных. Первый подход требует большой аналитической работы эксперта, которому часто бывает невозможно изложить свои знания, опыт и интуицию в рамках 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 интеллектуальных систем.

Очевидно, что использование второго подхода является наиболее предпочтительным, поскольку он позволяет значительно сократить временные издержки, снизить зависимость от экспертов и учитывать полный набор исходных данных. Эксперт может принимать участие на этапе оценки сформированной базы знаний и ее корректировке.

Однако, несмотря на все достоинства подхода, в процессе формирования базы знаний формируется большое количество правил, что в свою очередь усложняет работу эксперта по интерпретации и учитывает ошибки погрешности, которые ухудшают результаты работы интеллектуальной системы. Для решения данной проблемы необходимо производить редукцию правил базы знаний за счет их структурного упорядочивания и минимизации.

Данная задача впервые была сформулирована в [5], как задача таксономии знаний. Однако, ее практические реализации стали появляться лишь в последние несколько лет. Так, в работе [1] 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 [10,11] задача кластеризации знаний в системах искусственного интеллекта решается с применением муравьиных алгоритмов. Метод редукции баз знаний с применением генетических алгоритмов описан в [7].

Однако, несмотря на положительные результаты имеющихся решений, проблема редукции (сокращения числа правил) баз знаний остается актуальной. В данной статье предлагается два метода редукции нечетких правил в базах знаний интеллектуальных систем: на основе таксономии в пространстве знаний с применением методов нечеткой логики и кластерного анализа, а также на основе генетического алгоритма. Для однозначного решения поставленной задачи в качестве модели представления знаний в интеллектуальной системе выбрана модель Такаги-Сугено [12].

Пусть для формирования базы знаний используется нечеткая нейронная сеть ANFIS [4,13]. Процесс формирования может быть реализован в среде моделирования MatLab. 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 (в данном случае 1024 правила). Очевидно, что такое количество правил для решения большинства задач является избыточным. Поэтому требуется редукция автоматически сформированной при помощи нечеткой нейронной сети базы знаний, сокращение количества имеющихся в ней правил. Рассмотрим формальную постановку данной задачи.

Пусть имеется сформированная в процессе обучения сети *ANFIS* база знаний $R=\{R_1, R_2,\ldots, R_N\}$, где R_i (i=1..N) – нечетко-продукционные правила Такаги-Сугено, N – исходное количество правил в базе знаний. Требуется найти минимальный состав

из k правил ($k<N$), при котором эффективность базы знаний (точность решений) будет максимальной.

Для решения данной задачи рассмотрим разработанный метод редукции базы знаний на основе таксономии (кластеризации) нечетких правил. Пусть правила базы знаний представлены в виде модели Такаги-Сугено [3]:

$$If\ x_1 \text{ is } \widetilde{A}_1\ \&\ x_2\ is\ \widetilde{A}_2\ \&\ \dots x_n\ is\ \widetilde{A}_n\ Then\ y = f(x_1,...,x_n),$$

где $x_1,...,x_n$ – входные лингвистические переменные, $\widetilde{A}_1,...,\widetilde{A}_n$ – их нечеткие значения, y – четкая переменная выхода, $f(x_1,...,x_n)$ – вещественная функция от четких аргументов $x_1,...,x_n$.

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

1) существует эффективный способ сравнения нечетких антецедентов;

2) число различных значений консеквентов нечетких правил при любых значениях аргументов конечно и счетно.

Первое условие требует введения метрики расстояний в пространстве знаний, позволяющей определять «близость» двух нечетких правил, а также эффективного способа представления антецедента в формализованном виде, пригодном для использования в алгоритме кластеризации. Второе условие накладывает ограничение на вид функции, требуя дискретности ее значений. В случае решения задачи классификации данное требование легко выполняется. При этом значениями функции являются константы, указывающие на класс объекта.

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

$$If\ x_1 \text{ is } \widetilde{A}_1\ \&\ x_2\ is\ \widetilde{A}_2\ \&\ \dots x_n\ is\ \widetilde{A}_n\ Then\ y = C_i,$$

где C_i – метка некоторого класса.

Разобьем исходную базу знаний на непересекающиеся подмножества правил по признаку метки класса. Тогда кластеризация будет производиться независимо в каждом подмножестве правил путем объединения их антецедентов в кластеры.

Пусть имеем следующее множество правил, соответствующих одному классу решений:

$$If\ x_1 \text{ is } \widetilde{A}_{11}\ \&\ x_2\ is\ \widetilde{A}_{12}\ \&\ \dots x_n\ is\ \widetilde{A}_{1n}\ Then\ y = 1,$$

$$If\ x_1 \text{ is } \widetilde{A}_{21}\ \&\ x_2\ is\ \widetilde{A}_{22}\ \&\ \dots x_n\ is\ \widetilde{A}_{2n}\ Then\ y = 1,$$

$$\dots$$

$$If\ x_1 \text{ is } \widetilde{A}_{m1}\ \&\ x_2\ is\ \widetilde{A}_{m2}\ \&\ \dots x_n\ is\ \widetilde{A}_{mn}\ Then\ y = 1.$$

Представим каждое из правил вектором своих нечетких ограничений $\widetilde{A}_{ij}$. Тогда система правил примет вид:

$$\{(\widetilde{A}_{11}, \widetilde{A}_{12},\dots, \widetilde{A}_{1n}),\dots, (\widetilde{A}_{m1}, \widetilde{A}_{m2},\dots, \widetilde{A}_{mn})\}.$$

Переходя от нечетких множеств $\widetilde{A}_{ij}$ к их четким аналогам x_{ij} (используя процедуру дефаззификации по методу центра тяжести), получим:

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

Систему полученных векторов можно рассматривать, как множество точек в n-мерном Евклидовом пространстве, каждая из которых является результатом формализации антецедентов соответствующего нечеткого правила. Таким образом, таксономия нечетких правил производится путем объединения данных точек в локальные кластеры.

В общем случае, значения входных параметров нечетких правил измерены в разных шкалах, поэтому, прежде чем приступать к кластеризации, необходимо произвести нормировку 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.

В результате получаем множество точек в нормированном n-мерном пространстве, пригодных для кластеризации и поиска оптимального кластерного решения. В качестве алгоритма кластеризации используется алгоритм k-средних.

В результате кластеризации определяются логические центры кластеров, которые путем создания новых функций принадлежности превращаются в правила, из которых состоит новая база знаний. Лучшей будет база знаний с максимальной классифицирующей способностью при минимальном количестве правил.

Для решения задачи редукции нечетких правил также разработан генетический алгоритм, в котором база знаний кодируется хромосомой из N генов (N – количество правил). Обозначим наличие или отсутствие правила, как «0» или «1». Тогда популяция хромосом будет содержать 2^N особей.

Каждая хромосома оценивается мерой ее приспособленности согласно тому, насколько хорошо соответствующее ей решение задачи. Наиболее приспособленные особи получают возможность воспроизводить потомство с помощью перекрестного скрещивания с другими особями популяции. Это приводит к появлению новых особей, которые сочетают в себе некоторые характеристики, наследуемые ими от родителей. Наименее приспособленные особи с меньшей вероятностью смогут воспроизвести потомков, так что те свойства, которыми они обладали, будут постепенно исчезать из популяции в процессе эволюции. Иногда происходят мутации, или спонтанные изменения в генах. Таким образом, из поколения в поколение хорошие характеристики распространяются по всей популяции. Скрещивание наиболее приспособленных особей приводит к тому, что исследуются наиболее перспективные участки пространства поиска. В итоге популяция будет сходиться к оптимальному решению задачи.

Закодируем базу знаний интеллектуальной системы в виде хромосомы H_i:

H_i	1	0	1	1	0	1	0	...	1
	R_1	R_2	R_3	R_4	R_5	R_6	R_7	...	R_N

,

где $H_i = \{h_{ij}\}, h_{ij} = \begin{cases} 0, if\ R_j\ active, \\ 1, if\ R_j\ not\ active. \end{cases}$

Задача редукции сводится к поиску хромосомы с минимальным числом правил, не теряя качества классификации. Лучшей будет та хромосома, которая позволяет достичь максимума оценки классифицирующей способности (не меньше исходной) базы знаний при минимальном числе правил.

В задаче редукции базы знаний критерием оптимальности может служить ошибка обобщения, получаемая интеллектуальной системой при ее работе на тестовой выборке данных:

$$E = \left(1 - \frac{N_{true}}{N}\right) \to \min\ ,$$

где N_{true} – количество правильно классифицированных примеров, N – общее количество примеров.

Количество хромосом в начальной популяции $\mathrm{K_{ch}} = 2 * \mathrm{round}(\sqrt{\mathrm{N}})$. Отбор начальных хромосом производится случайным образом. В процессе работы алгоритма выполняются операторы селекции, скрещивания, мутации и редукции хромосом.

Селекция родительских хромосом выполняется по методу «колеса рулетки» (см. рис.1).

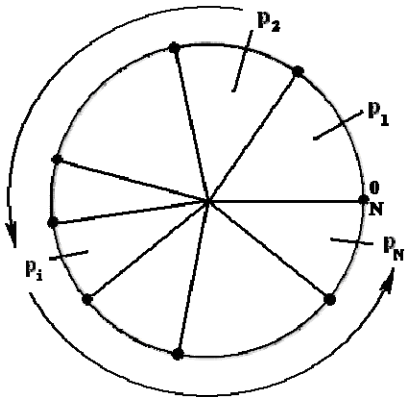

Рис. 1 - Колесо рулетки для селекции хромосом

Вероятность выбора хромосомы определяется следующим образом:

$$p_i = \frac{1 - F(H_i)}{\sum_i (1 - F(H_i))},$$

где $p_i \in [0;1]$ и чем больше значение фитнес-функции $F(H_i)$ для хромосомы, тем больше вероятность выбора хромосомы для скрещивания.

Оператор скрещивания применяется к двум родительским хромосомам, используя одноточечный кроссинговер с вероятностью из интервала [0.5,1] для получения двух дочерних хромосом. Мутация применяется путем случайной инверсии одного из бита дочерней хромосомы с вероятность 0.03. Для получения новой популяции используется оператор редукции, в результате которого происходит удаление двух худших хромосом из текущего хромосомного набора.

Генетический алгоритм выполняется до тех пор, пока в результате проведения вычислений не будут появляться хромосомы с лучшей функцией приспособленности в течение определенного числа поколений. После окончания его работы отбирается одна хромосома с лучшими параметрами фитнес функции, которая и будет определять искомую базу знаний интеллектуальной системы.

Для оценки эффективности разработанных методов редукции нечетких правил спроектируем систему нечеткого логического вывода для задачи классификации ирисов, предложенную Фишером в 1936 году [9]. Данная задача состоит в отнесении ириса к одному из трех классов: Iris Setosa, Iris Versicolor и Iris Virginica.

При классификации используются следующие признаки цветков: x_1 – длина чашелистика; x_2 – ширина чашелистика; x_3 – длина лепестка; x_4 – ширина лепестка. Исходные данные для классификации ирисов записаны в файле iris.dat, входящем в Fuzzy Logic Toolbox. Файл содержит 150 строк, каждая из которых описывает один ирис. Информация о цветке представлена пятеркой чисел – первые четыре числа соответствуют значениям признаков, а пятое – классу ириса.

При построении модели нечеткого логического вывода база данных была разделена на 2 группы: признаки соответствуют классу Iris Virginica и признаки не соответствуют данному классу. Для оценки эффективности построенных моделей использована тестовая выборка, полученная случайным отбором 10 и 20 признаков из соответствующих групп признаков. Таким образом, обучающая выбора состояла из 120 строк, а тестовая из 30.

При построении исходной базы знаний были выбраны треугольная и трапециевидная функции принадлежности с 3 градациями. Таким образом, исходная база знаний содержала 81 правило.

Эффективность системы характеризуется двумя типами ошибок:

1) ошибкой первого рода – ложный пропуск класса Iris Virginica, то есть неверное отнесение цветка к классу Iris Virginica;

2) ошибкой второго рода – ложное срабатывание, т.е. неверное отнесение к классу Iris Virginica.

Пусть N_1 – количество попыток классификации ирисов, n_1 – число ложных пропусков класса Iris Virginica. Тогда, ошибка первого рода:

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

Ошибка второго рода, соответственно:

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

где N_2 – количество попыток классификации ирисов; n_2 – число ложных срабатываний.

В таблицах 1 и 2 представлены типовые результаты тестирования интеллектуальной системы классификации ирисов на исходной и на редуцированной базах знаний.

Как видно из таблиц, подход на основе кластеризации нечетких правил обладает недостатком, связанным с чрезмерно длительной работой алгоритма, в то время как время выполнения генетического алгоритма существенно ниже. Однако, при

этом результаты кластеризации базы знаний на тестовой выборке показали преимущество по сравнению с методом редукции баз знаний с применением генетического алгоритма.

Таблица 1 - Треугольная функция принадлежности

	Исходная база знаний	Кластеризация БЗ	Генетический алгоритм
Ошибки 1-го рода	0%	0%	0%
Ошибки 2-го рода	0%	0%	5%
Время обработки, у.е.	-	58.7340	2.6970
Количество правил	81	8	25
Время выполнения, у.е.	0.0099	0.0078	0.0070

Таблица 2 - Трапециевидная функция принадлежности

	Исходная база знаний	Кластеризация БЗ	Генетический алгоритм
Ошибки 1-го рода	0%	0%	0%
Ошибки 2-го рода	0%	0%	5%
Время обработки, у.е.	-	33.2364	0.3292
Количество правил	81	53	35
Время выполнения, у.е.	0.0168	0.0072	0.0078

Применение редукции нечетких правил в базе знаний интеллектуальной системы:

- уменьшает объем базы знаний;
- повышает ее интерпретируемость;
- уменьшает неопределенность выбора того или иного правила при принятии решения;
- повышает точность и скорость работы интеллектуальной системы.

Таким образом, практическая ценность предложенного подхода к редукции нечетких правил заключается в возможности повышения эффективности использования интеллектуальных систем в любой сфере человеческой деятельности.

Литература

1. Бухнин А.В., Бажанов Ю.С. Оптимизация баз знаний экспертных систем с применением нейронных нечетких сетей // Нейрокомпьютеры: разработка, применение. 2007. №11.
2. Гаврилова Т.А., Хорошевский В.Ф. Базы знаний интеллектуальных систем. – СПб.: Питер, 2001. – 384 с.: ил.
3. Глова В.И., Аникин И.В., Катасёв А.С., Кривилёв М.А., Насыров Р.И. Мягкие вычисления: учебное пособие. Казань: Изд-во Каз. гос. технич. университета им. А.Н. Туполева, 2010. – 206 с.
4. Емалетдинова Л.Ю., Катасёв А.С., Кирпичников А.П. Нейронечеткая модель аппроксимации сложных объектов с дискретным выходом // Вестник Казанского технологического университета. – 2014. – Т. 17, № 1. – С. 295-299.
5. Загоруйко Н.Г. Прикладные методы анализа данных и знаний. – Новоси-бирск: Изд-во Ин-та математики, 1999. – 270 с.
6. Кирпичников А.П., Осипова А.Л., Ризаев И.С. Повышение аналитических возможностей баз данных // Вестник Казан. технол. ун-та. – 2012. – № 3. – С. 157-160.
7. Комарцова Л.Г. Эволюционные методы формирования нечетких баз правил // Open Semantic Technologies for Intelligence Systems, 2011. С.181-184.
8. Титов А.Н., Нуриев Н.К., Тазиева Р.Ф. Оценка параметров вероятностной модели по экспериментальным данным // Вестник Казан. технол. ун-та. – 2013. – № 19. – С. 324-330.
9. Штовба С.Д. Классификация объектов на основе нечеткого логического вывода // Exponenta Pro - Математика в приложениях. – 2004. –№ 1(5). – С. 68-69.
10. Щуревич Е.В. Кластеризация знаний в системах искусственного интеллекта // Информационные технологии. 2009. №2. С. 25-29.
11. 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.
12. 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.
13. 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.

© **А. Р. Абдулхаков** – аспирант кафедры систем информационной безопасности КНИТУ-КАИ, e-mail: aidar_abdulhakov@mail.ru; **А. С. Катасёв** – канд. техн. наук, доц. кафедры систем информационной безопасности КНИТУ-КАИ, e-mail: kat_726@mail.ru; **А. П. Кирпичников** – д-р. физ.-мат. наук, профессор, зав. кафедрой интеллектуальных систем и управления информационными ресурсами КНИТУ, e-mail: kirpichnikov@kstu.ru.

© **A. R. Abdulhakov** – Postgraduate Student of the Department of Information Security Systems, KNRTU named after A.N. Tupolev, e-mail: aidar_abdulhakov@mail.ru; **A. S. Katasev** – PhD, Associate Professor of the Department of Information Security Systems, KNRTU named after A.N. Tupolev, e-mail: kat_726@mail.ru; **A. P. Kirpichnikov** – Dr. Sci, Prof, Head of the Department of Intelligent Systems & Information Systems Control, KNRTU, e-mail: kirpichnikov@kstu.ru.