УДК 004.891.3

КЛАСТЕРНО-ГЕНЕТИЧЕСКИЙ МЕТОД РЕДУКЦИИ БАЗ ЗНАНИЙ ИНТЕЛЛЕКТУАЛЬНЫХ СИСТЕМ

Абдулхаков А.Р., Катасёв А.С.

ФГБОУ ВПО «Казанский национальный исследовательский технический университет им. А.Н. Туполева – КАИ», Казань, e-mail: aidar_abdulhakov@mail.ru, Kat_726@mail.ru

В данной работе решается задача редукции автоматически сформированных баз знаний интеллектуальных систем. Предлагается кластерно-генетический метод редукции, основанный на алгоритме кластеризации нечетких правил и генетическом алгоритме минимизации числа кластеров. Алгоритм кластеризации правил исходной базы знаний формирует промежуточную базу знаний, состоящую из логических центров кластеров. Идентификация значений параметров функций принадлежности в данных центрах производится на основе разработанного метода средних координат. Для получения значимых кластеров предложен генетический алгоритм, на вход которого подается промежуточная база знаний, а на выходе формируется искомая (редуцированная) база знаний. На основе разработанного математического обеспечения реализован программный комплекс, позволяющий редуцировать базы знаний интеллектуальных систем. Для оценки эффективности его работы выполнены исследования по редукции сформированных баз знаний на известном наборе данных. Эксперименты показали, что редуцированные базы знаний обладают более высокой классифицирующей способностью, а также требуют меньше времени на выполнение логического вывода. Это указывает на эффективность редукции нечетких правил в базах знаний интеллектуальных систем.

Ключевые слова: нечеткое правило, база знаний, интеллектуальная система, редукция нечетких правил, кластеризация, генетический алгоритм, идентификация значений параметров функций принадлежности

THE FUZZY RULES CLUSTER-GENETIC REDUCTION METHOD IN INTELLIGENT SYSTEMS KNOWLEDGE BASES

Abdulkhakov A.R., Katasev A.S.

Kazan National Research Technical University n.a. A.N. Tupolev – KAI, Kazan, e-mail: aidar_abdulhakov@mail.ru, Kat_726@mail.ru

In this paper solves the problem of reduction the automatically generated knowledge base intelligent systems. It is proposed to cluster-genetic method of reduction based on fuzzy clustering algorithm rules and genetic algorithms to minimize the number of clusters. Clustering algorithm rules the original knowledge base generates intermediate knowledge base consisting of a logical cluster centers. The parameters identification of membership functions in this centers is based on secondary origin developed method. For meaningful clusters proposed genetic algorithm which is applied to the input intermediate knowledge base and the output generated the desired (reduced) knowledge base. Based on the developed mathematical software implemented software package that allows to reduce the knowledge base of intelligent systems. To assess the effectiveness of its work carried out research on the reduction of formed knowledge bases on the known data set. The experiments have shown that the reduced knowledge base have a higher classifying ability and require less time to perform inferencing. This indicates the effectiveness of fuzzy rules reduction in intelligent systems knowledge bases.

Keywords: fuzzy rule, knowledge base, intelligent system, reduction of fuzzy rules, clustering, genetic algorithm, membership functions parameters identification

В настоящее время интеллектуальные системы, основанные на знаниях, получили широкое распространение в различных прикладных областях для решения таких задач, как прогнозирование, распознавание образов, диагностика, управление и другие [2]. Основным компонентом интеллектуальных систем является база знаний, включающая набор правил принятия решений, выраженных в форме четких или нечетких продукций. Использование нечетко-продукционных правил позволяет решать практические задачи в условиях нечеткости, неопределенности и неполноты исходных данных.

Для формирования правил базы знаний в последнее время все чаще применяют методы автоматического формирования нечетких правил, основанные на методах и алгоритмах интеллектуального анализа данных. Использование такого подхода значительно упрощает и ускоряет процесс разработки интеллектуальной системы. Однако, несмотря на его достоинства, автоматически сформированные базы знаний, как правило, обладают избыточностью, что не позволяет использовать их с максимальной эффективностью при решении практических задач. Для повышения эффективности использования интеллектуальных систем требуется оценка избыточности и редукция их баз знаний [1]. Проведенный анализ показал, что существующие подходы к редукции нечетких правил не лишены недостатков, в частности возможности снижения

точности принимаемых решений на основе редуцированных баз знаний. Это актуализирует необходимость разработки и апробации новых эффективных методов и алгоритмов редукции нечетких правил в базах знаний интеллектуальных систем. Таким образом, для повышения эффективности использования интеллектуальных систем в работе решается задача разработки кластерно-генетического метода редукции нечетких правил, основанного на алгоритме их кластеризации и генетическом алгоритме минимизации числа кластеров.

Кластерно-генетический метод редукции нечетких правил

В основу разработанного метода редукции нечетких правил положены принципы таксономии знаний (кластеризации нечетких правил) [3], а также эволюционного моделирования (генетической оптимизации) [4]. Обобщенная схема кластерно-генетического метода представлена на рис. 1.

Разработанный метод редукции нечетких правил состоит из 2-х этапов:

1) кластеризация (таксономия) нечетких правил в исходной базе знаний с получением промежуточной базы знаний, состоящей из правил, соответствующих центрам кластеров;

2) редукция нечетких правил промежуточной базы знаний на основе генетического алгоритма, позволяющего минимизировать число правил и сформировать искомую базу знаний.

Рассмотрим каждый из этапов более подробно. Пусть имеется исходная база знаний $R = \{R_1, R_2, ..., R_m\}$, где R_i ($i = 1…m$) – нечетко-продукционные правила Такаги – Сугено вида

ЕСЛИ $x_1 = \tilde{A}_{11}$ И $x_2 = \tilde{A}_{12}$ И … $x_n = \tilde{A}_{1n}$ ТО $y = B_1$,

ЕСЛИ $x_1 = \tilde{A}_{21}$ И $x_2 = \tilde{A}_{22}$ И … $x_n = \tilde{A}_{2n}$ ТО $y = B_2$,

…

ЕСЛИ $x_1 = \tilde{A}_{m1}$ И $x_2 = \tilde{A}_{m2}$ И … $x_n = \tilde{A}_{mn}$ ТО $y = B_k$,

где $x_1, ..., x_n$ – входные лингвистические переменные; $\tilde{A}_{11},..., \tilde{A}_{1n}, \tilde{A}_{21},..., \tilde{A}_{mn}$ – их нечеткие значения; y – четкая выходная переменная; $B_1, ..., B_k$ – классы решений.

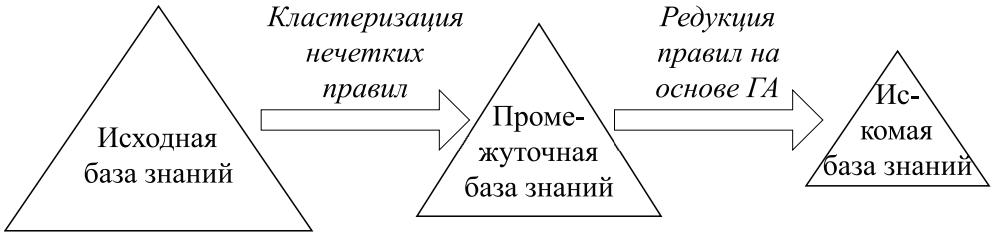

Рис. 1. Схема кластерно-генетического метода редукции нечетких правил

Представим антецеденты нечетких правил в виде вектора их нечетких ограничений. Тогда система правил примет следующий вид:

$$\left(\tilde{A}_{11}, \tilde{A}_{12},..., \tilde{A}_{1n}\right), \left(\tilde{A}_{21}, \tilde{A}_{22},..., \tilde{A}_{2n}\right),..., \left(\tilde{A}_{m1}, \tilde{A}_{m2},..., \tilde{A}_{mn}\right). \quad (1)$$

Перейдем от нечетких множеств $\tilde{A}_{ij}$ ($i = 1...m, j = 1...n$) к их четким аналогам x_{ij}, используя процедуру дефаззификации по методу центра тяжести:

$$x_{\text{цт}} = \frac{\sum_i \left(\mu_{\tilde{A}}(x_i) \cdot x_i\right)}{\sum_i \mu_{\tilde{A}}(x_i)}.$$

После дефаззификации выражение (1) примет вид

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

Таким образом, исходная база знаний представляется точками в n-мерном Евклидовом пространстве, количество координат которых соответствует количеству входных параметров нечетких правил. Следовательно, задача таксономии нечетких правил сводится к задаче кластеризации полученных точек данных.

В общем случае значения входных параметров нечетких правил измерены в разных шкалах, поэтому перед кластеризацией необходимо произвести нормировку дефаззифицированных значений антецедентов, используя метрику вида

$$x'_{ij} = \frac{x_{ij} - \min\limits_{i=1,m}(x_{ij})}{\max\limits_{i=1,m}(x_{ij}) - \min\limits_{i=1,m}(x_{ij})},$$

где x'_{ij} – исходное значение параметра; x'_{ij} – нормированное значение.

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

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

Таксономию знаний необходимо производить независимо для каждого класса решений, поэтому перед ее выполнением необходимо разделить множество точек данных на подмножества по классам решений и в каждом из них кластеризацию производить отдельно.

На рис. 2 представлена блок-схема разработанного алгоритма кластеризации с получением промежуточной базы знаний.

В качестве алгоритма кластеризации точек данных выбран расширенный алгоритм *k*-средних, благодаря его масштабируемости (возможности работы с большими массивами данных) и способности работать с разнотипными параметрами. При выборе лучшего кластерного решения необходимо руководствоваться критерием ошибки обобщения, получаемой интеллектуальной системой при ее работе на тестовой выборке данных:

$$E = \left(1 - \frac{N_{\text{прав}}}{N_{\text{общ}}}\right) \to \min_{k},$$

где $N_{\text{прав}}$ – число правильно классифицированных примеров; $N_{\text{общ}}$ – общее число примеров.

В полученном кластерном решении центры кластеров могут как совпадать с имеющимися правилами в базе знаний (физические центры кластера), так и не совпадать с ними (логические центры кластеров). В последнем случае возникает задача идентификации значений параметров функций принадлежности нового нечеткого правила, соответствующего данной точке. Для ее решения разработан численный метод (метод средних координат). Рассмотрим пример идентификации значений параметров треугольной функции принадлежности.

Рис. 2. Блок-схема алгоритма кластеризации

Треугольная функция принадлежности задается тройкой чисел $\{l, c, r\}$, при этом функция принадлежности определяется по следующей формуле:

$$\mu(x) = \begin{cases} 0, & x \le l; \\ \dfrac{x-l}{c-l}, & l < x \le c; \\ \dfrac{r-x}{r-c}, & c < x \le r; \\ 0, & x > r. \end{cases}$$

Необходимо получить значения параметров функций принадлежности, соответствующих каждой координате из N точек кластера (l_{ij}, c_{ij}, r_{ij}) $i = \overline{1,N}$, $j = \overline{1,n}$ и для каждой j-й координаты логического центра кластера вычислить значения параметров функции принадлежности по следующим формулам:

$$l_{\text{лц}} = \frac{1}{N}\sum_{i=1}^{N} l_{ij};$$

$$c_{\text{лц}} = \frac{1}{N}\sum_{i=1}^{N} c_{ij};$$

$$r_{\text{лц}} = \frac{1}{N}\sum_{i=1}^{N} r_{ij}.$$

Результатом кластеризации нечетких правил в исходной базе знаний является промежуточная база знаний, состоящая из правил, соответствующих центрам кластеров. Для минимизации правил (удаления незначимых центров кластеров) используется редукция нечетких правил на основе генетического алгоритма, позволяющего минимизировать число сформированных кластеров.

Пусть имеется промежуточная база знаний $R = \{R_1, R_2, ..., R_m\}$, содержащая множество нечетких правил R_j, $j = 1...m$, где m – объем базы знаний. Закодируем базу знаний в виде хромосомы

$$H_i = \{h_{ij}\},$$

где $h_{ij} = \begin{cases} 0, & \text{если } R_j \text{ неактивно (исключено из базы знаний),} \\ 1, & \text{если } R_j \text{ активно (включено в базу знаний).} \end{cases}$

Пример кодирования базы знаний:

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

Создание начальной популяции объемом m выполняется следующим образом: в популяцию включается родительская хромосома и набор потомков, полученных в результате случайной мутации ее генов с вероятностью 0,02.

Задача редукции нечетких правил сводится к поиску хромосомы, позволяющей достичь максимума оценки классифицирующей способности базы знаний (не меньше исходной точности классификации) при минимальном числе активных правил. Таким образом, используемая в генетическом алгоритме фитнесс-функция имеет вид

$$F(H_i) = \frac{N_{\text{прав}}}{N_{\text{общ}}} \to \max_{\forall H_i},$$

где $N_{\text{прав}}$ – число правильно распознанных примеров в выборке данных; $N_{\text{общ}}$ – ее объем.

Рассмотрим реализацию генетических операторов.

На этапе селекции производится отбор 2-х родительских хромосом из начального хромосомного набора, используя метод колеса рулетки. В данном методе вероятность выбора хромосомы определяется следующим образом:

$$p_i = \frac{F(H_i)}{\sum_i F(H_i)}.$$

Оператор скрещивания позволяет получать 2-х потомков от родительских хромосом на основе одно- и двухточечного кроссинговера.

Мутация осуществляется путем инверсии с вероятностью 0,02 одного из единичных генов дочерних хромосом.

После определения приспособленности дочерних хромосом выполняется оператор редукции, в результате которого происходит удаление 2-х худших хромосом из текущего хромосомного набора и формируется новая популяция.

Данный алгоритм выполняется до тех пор, пока в результате вычислений не будут появляться хромосомы с лучшей функцией приспособленности в течение определенного числа поколений. После окончания его

работы отбирается одна хромосома с лучшими значениями фитнесс-функции, которая и будет определять искомую базу знаний интеллектуальной системы.

Исходя из условий и ограничений применимости разработанных методов и алгоритмов, предъявляются следующие требования к редуцируемой базе знаний:

1) большая размерность;

2) тип базы знаний: нечетко-продукционная с кусочно-линейными функциями принадлежности нечетких антецедентов;

3) задача, решаемая на правилах базы знаний: задача классификации;

4) известен алгоритм логического вывода на правилах базы знаний;

5) база знаний сформирована автоматически на основе методов и алгоритмов машинного обучения (например, нейронных сетей);

6) доступны все экспериментальные данные, на основе анализа которых сформирована база знаний.

Для оценки эффективности математического обеспечения редукции нечетких правил в базах знаний интеллектуальных систем разработан программный комплекс в среде Matlab.

Проведение исследований на базе программного комплекса

Для проведения исследований на базе разработанного программного комплекса использован набор данных для решения задачи выявления нестандартных транзакций с банковскими картами из общедоступного источника *UCI Machine Learning Repository* [5]. Исходная выборка данных включала 690 записей по 14 входным параметрам и одному выходному с двумя классами решений (стандартные и нестандартные транзакции).

Формирование исходных баз знаний (с использованием треугольных и трапециевидных функций принадлежности) выполнено на основе нечеткой нейронной сети ANFIS. Результаты проведенных исследований показали, что редуцированные базы знаний показывают лучшие свойства классифицирующей способности и скорости логического вывода по сравнению с исходными базами знаний (таблица).

Таким образом, редукция баз знаний интеллектуальных систем на базе разработанного математического обеспечения приводит к устранению избыточности баз знаний, повышению классифицирующей способности и интерпретируемости базы знаний, а также повышению скорости принятия решений.

Заключение

Описанные в работе математическое обеспечение и программный комплекс позволяют редуцировать нечеткие правила в базах знаний интеллектуальных систем. Проверка работы программного комплекса на известных наборах данных показала высокую эффективность разработанного математического обеспечения и практическую пригодность программного комплекса к решению поставленных задач.

Сравнение исходных и редуцированных баз знаний

База знаний	Число нечетких правил		Классифицирующая способность		Время логического вывода, с	
	треуг. ФП	трапец. ФП	треуг. ФП	трапец. ФП	треуг. ФП	трапец. ФП
Исходная	13 608	13 608	0,88	0,91	0,43	0,41
Промежуточная	973 (–93 %)	954 (–93 %)	0,88	0,92 (+1,09 %)	0,08 (–81,3 %)	0,07 (–82,9 %)
Искомая	649 (–95 %)	610 (–96 %)	0,89 (+1,13 %)	0,93 (+2,19 %)	0,07 (–83,7 %)	0,06 (–85,4 %)

Список литературы

1. Абдулхаков А.Р., Катасёв А.С., Кирпичников А.П. Методы редукции нечетких правил в базах знаний интеллектуальных систем // Вестник Казанского технологического университета. – Т. 17, № 23. – 2014. – С. 389–392.

2. Гаврилова Т.А., Хорошевский В.Ф. Базы знаний интеллектуальных систем. – СПб.: Питер, 2001. – 384 с.: ил.

3. Загоруйко Н.Г. Прикладные методы анализа данных и знаний. – Новосибирск: Изд-во Ин-та математики, 1999. – 270 с.

4. Комарцова Л.Г. Эволюционные методы формирования нечетких баз правил // Open Semantic Technologies for Intelligence Systems. – 2011. – С. 181–184.

5. Bache K., Lichman M. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science. 2013.

References

1. Abdulhakov A.R., Katasjov A.S., Kirpichnikov A.P. *Vestnik Kazanskogo tehnologicheskogo universiteta*. 2014, T. 17, no 23, pp. 389–392.

2. Gavrilova T.A., Horoshevskij V.F. *Bazy znanij intellektualnyh sistem*. SPb, Piter, 2001, 384 p.

3. Zagorujko N.G. *Prikladnye metody analiza dannyh i znanij*. Novosibirsk: Izd-vo In-ta matematiki, 1999, 270 p.

4. Komarcova L.G. *Open Semantic Technologies for Intelligence Systems*. 2011, pp. 181–184.

5. Bache K., Lichman M. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science. 2013.

Рецензенты:

Песошин В.А., д.т.н., профессор кафедры компьютерных систем, Казанский национальный исследовательский технический университет им. А.Н. Туполева – КАИ, г. Казань;

Кирпичников А.П., д.ф.-м.н., профессор, заведующий кафедрой интеллектуальных систем и управления информационными ресурсами, Казанский национальный исследовательский технологический университет, г. Казань.