Моисеев С. А., аспирант
Воронежский государственный технический университет
Леденева Т. М., д-р техн. наук, проф.
Воронежский государственный университет

ЭВОЛЮЦИОННЫЙ АЛГОРИТМ ДЛЯ ПРОВЕДЕНИЯ НЕЧЕТКОЙ КЛАСТЕРИЗАЦИИ ПРИ ФОРМИРОВАНИИ ТЕРМ-МНОЖЕСТВА ЛИНГВИСТИЧЕСКОЙ ПЕРЕМЕННОЙ

akakay.malahov@mail.ru

В статье рассматривается применение эволюционного алгоритма для проведения нечеткой кластеризации для формирования множества термов лингвистической переменной на примере параметра давления греющего пара процесса вулканизации.

Ключевые слова*: нечеткая кластеризация, индекс Кси-Бени, эволюционный алгоритм, процесс вулканизации.*

В данной статье предлагается метод нечеткой кластеризации с использованием эволюционного алгоритма, а также продемонстрированы результаты его работы при формировании терм-множества лингвистической переменной «Давление греющего пара», которая является одним из параметров технологического процесса вулканизации. Формирование лингвистических шкал является важной задачей при построении экспертных систем. Его можно представить в виде процесса эволюционирования матрицы разбиения, представляющей возможный способ группирования предоставленного набора данных в определенное количество кластеров таким образом, что образцы в одной группе похожи в некотором смысле, а образцы из разных групп в этом же смысле различны.

Существует достаточно большое количество методов кластеризации, среди которых наибольшее распространение имеет метод нечетких с-средних [1].

В данной статье предлагается подход для построения разбиения значений некоторого набора величин на кластеры с применением эволюционного алгоритма со специальными операторами кроссинговера и мутации. Описание принципов работы и основных компонентов эволюционного алгоритма можно найти во многих источниках [2].

Решение поставленной задачи является первым этапом при построении системы нечеткого логического вывода, для чего необходимо последовательно выполнить ряд шагов:

разбиение пространства значений входных и выходных лингвистических переменных на характерные области – термы;

1) первичное формирование базы нечетких правил на основе выделенных термов;

2) редукция количества правил в базе путем исключения малоинформативных, противоречивых и дублирующих правил;

3) присвоение весового коэффициента каждому правилу, показывающего достоверность правила;

4) «тонкая» настройка правил баз данных для получения более точного значений выходной переменной.

Из представленного алгоритма видно, что настройка термов также будет осуществляться на следующих этапах построения нечеткой системы, однако качественное первоначальное формирование терм-множеств позволит упростить общую задачу, повысить скорость и эффективность работы системы.

Эволюционные алгоритмы для проведения нечеткой кластеризации можно разбить на 2 больших группы. К первой относятся алгоритмы, в основе которых лежит метод нечетких с-средних и различные его вариации, при этом применяются техники точной кластеризации, адаптированные для проведения нечетких разбиений [3, 4]. Вторая группа методов использует метод нечетких с-средних для проведения локального поиска на кластерах, обнаруженных эволюционным алгоритмом [5, 6], что приводит к улучшению сходимости. Также за последнее время было представлено множество индексов нечеткой кластеризации, однако их применение в эволюционных вычислениях не раскрыто в достаточной степени. В данной статье представлен эволюционный алгоритм, использующий в качестве критерия нечеткого разбиения индекс Кси-Бени, а также оригинальные операторы мутации и кроссинговера, способствующие повышению качества работы алгоритма.

При разбиении пространства входных и выходных лингвистических переменных на термы предполагается также автоматическое определение оптимального количества термов путем применения генетического алгоритма с переменной длиной хромосом для кодирования различных количеств термов. Данная задача относится к задачам кластеризации и может быть решена также с применением метода нечетких с - средних. Для

измерения значений функции приспособленности хромосом используется индекс Кси-Бени (индекс XB), закодированный в хромосоме. Пусть множество всех возможных матриц разбиения обозначается формулой (1).

$$V = \{U \in \Re^{m\times m}; \sum_{i=1}^{m} v_{ik} = 1, \sum_{k=1}^{n} v_{ik} > 0, v_{ik} \in [0,1]\} \quad (1)$$

Наилучшим разбиением считается такое, которое минимизирует индекс XB, что представлено в (2).

$$V^* \in V; XB(V^*, C^*, X) = \min_{V_i \in U}(XB(V_i, C_i, X)) \quad (2)$$

В формуле (2) V^* - оптимальная матрица разбиения, C^* - центры кластеров, X - вектор входных значений, XB - индекс XB. Количество кластеров и соответствующее нечеткое разбиение данных эволюционирует одновременно.

Используется кодирование вещественными числами, хромосома кодирует множество центров кластеров, количество генов для каждого центра кластера в хромосоме соответствует размерности входного пространства. Таким образом, применяется Питсбургский подход к кодированию пространства решений, когда каждая хромосома представляет собой отдельное решение поставленной задачи. При формировании первоначальной популяции генерируется Q_0 хромосом, число Q_0 задается пользователем. Количество кластеров Q_t в каждой хромосоме, то есть длина каждой отдельной хромосомы, при формировании первоначальной популяции вычисляется по (3).

$$Q_t = INT(RAND(2, M^* + 1)) \quad (3)$$

В (3) M^* представляет оценку верхней границы числа кластеров и используется только для генерации первоначальной популяции. Значение M^* также либо задается пользователем, либо берется как результат работы других более грубых методов нечеткого разбиения. Хромосомы, кодирующие центры кластеров, выбираются случайным образом в соответствии с областью определения каждой переменной.

Функция приспособленности хромосомы показывает степень полезности решения. В данной статье в качестве функции приспособленности применяется индекс XB [7], который представляет собой отношение общего разброса к минимальному расстоянию между кластерами. Значение общего разброса вычисляется по (4), минимальное расстояние между центрами кластеров – по (5), вычисление индекса XB производится по (6).

$$\sigma(V, C, X) = \sum_{i=1}^{m}\sum_{k=1}^{n} v_{ik}^2 D^2(c_i, x_k), \quad (4)$$

$$s(C) = \min_{i\neq j}(\|c_i - c_j\|^2), \quad (5)$$

$$XB(V, C, X) = \frac{\sigma(V, C, X)}{n * s(C)}. \quad (6)$$

В приведенных формулах σ - общий разброс, s - минимальное расстояние между центрами кластеров, XB - индекс XB, n - размерность массива разбиваемых точек данных, m - количество центров кластеров, c - координаты центров кластеров, v_{ik} - элементы матрицы разбиения, D - выбранная мера расстояния между точками данных и центрами кластеров для решаемой задачи.

Когда разбиение компактное и «хорошее», значение σ должно быть малым, в то время как значение s должно быть большим, приводя к уменьшению значения индекса XB. Таким образом, целью является минимизация индекса XB для получения оптимальной матрицы разбиения.

Пусть хромосома кодирует m центров кластеров, которые обозначаются как $c_1, c_2, ..., c_m$. Значения принадлежностей v_{ik}, $i = 1,2,...,m$, $k = 1,2,...,n$ вычисляются по (7)

$$v_{ik} = \left(\sum_{j=1}^{m} \left(\frac{D(c_i, x_k)}{D(c_j, x_k)} \right)^{\frac{2}{W-1}} \right)^{-1} \quad (7)$$

В формуле (7) W - параметр нечеткости, обычно выбирают значение параметра равное 2 и выше, при значении параметра равным 1 получаем точную кластеризацию. Функция приспособленности хромосомы определяется как XB^{-1}, таким образом, максимизация функции приспособленности минимизирует значение индекса XB.

В качестве оператора селекции используется широко распространенный метод колеса рулетки [8].

В данной статье для определения пар родителей, участвующих в скрещивании, применим селективный метод отбора, при котором право участия в формировании потомков приобретают те особи, значение функции приспособленности которых больше, чем среднее значение функции приспособленности по популяции в целом. Это обеспечивает достаточно быструю сходимость алгоритма.

Для выполнения операции кроссинговера применяется следующий подход. Центры кластеров считаются неделимыми, и точки кроссинговера могут располагаться только между двумя центрами кластеров. Оператор кроссинговера применяется случайным образом с вероятностью P_{cross}, при этом необходимо убедиться, что операция кроссинговера произошла и оба потомка кодируют центры как минимум двух новых кластеров.

Каждый локус хромосомы может мутировать с фиксированной вероятностью p_m. Число $\tau \in [0,1]$ генерируется в соответствии с однородным распределением. Изменение значение гена происходит следующим образом: если значение гена до мутации было равно h, после применения оператора мутации оно будет равным $(1 \pm 2*\tau)*h$, когда $h \neq 0$ и $\pm 2*\tau$ если $h = 0$. Знак мутации выбирается с вероятностью $1/2$. В качестве критерия останова работы генетического алгоритма была выбрана генерация максимального количества поколений.

При решении задачи автоматического формирования термов лингвистической переменной количество найденных кластеров будет иметь физический смысл количества термов, сами кластеры, соответственно, представляют термы лингвистической переменной с кусочно-линейными функциями принадлежности. При этом предлагается осуществить переход от кусочно-линейных к треугольным функциям принадлежности без существенной потери точности. Для этого необходимо определить крайнюю левую и правую точки кусочно-линейной функции, соответствующие минимальному значению степени принадлежности и точку с максимальной степенью принадлежности. Полученные три точки характеризуют треугольное нечеткое число. На следующих этапах настройки нечеткой базы правил значения параметров сформированных нечетких чисел будут подвергнуты дополнительной настройке, поэтому процедура перехода от одной функции принадлежности к другой указанным способом представляется весьма разумной.

Рассмотрим применение предлагаемого подхода для разбиения входного пространства лингвистической переменной «Давление греющего пара» на термы. Данный параметр является одним из значимых параметров технологического процесса вулканизации, при этом предлагаемый метод, а также результаты подобного разбиения планируется использовать в дальнейшем для синтеза и анализа деятельности системы диагностирования аварийных ситуаций процесса вулканизации. График зависимости изменения значений параметра от времени представлен на рис. 1. На данном рисунке отображены графики как при нормальном состоянии процесса вулканизации для нескольких процессов, так и при аварийной работе вулканизаторов, совмещение графиков сделано с целью максимально полно представить возможные значения параметра для формирования адекватного терм – множества. Фактически, в данном случае при формировании терм - множества лингвистической переменной нас интересует проекция найденных нечетких кластеров на ось ординат, зависимость от времени для решения подобной задачи носит условный характер.

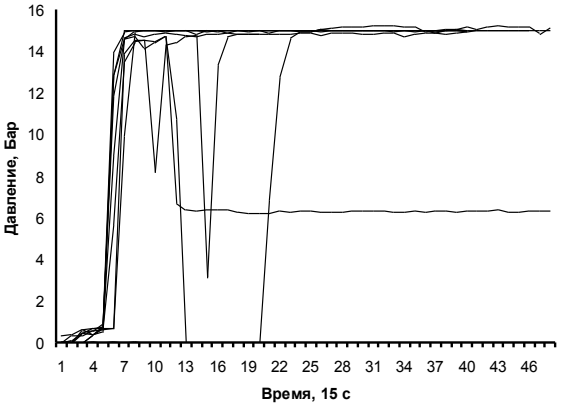

Рисунок 1. График значений давления греющего пара в зависимости от времени

Значения параметров эволюционного алгоритма приведены в табл. 1.

Таблица 1

Значения параметров эволюционного алгоритма

Тип популяции	Вектор действительных чисел
Первоначальное количество кластеров	7
Размер популяции	20
Тип селекции	Колесо рулетки
Количество элитных особей	2
Максимальное количество поколений	10000
Вероятность мутации	0,05
Вероятность кроссинговера	0,95

В результате работы алгоритма было получено нечеткое разбиение пространства значений лингвистической переменной на 4 кластера, координаты центров нечетких кластеров следующие: (0,97;3), (7,7;16), (12,1;11), (15,3; 28). Графически матрица нечеткого разбиения представлена на рис. 2 - 5, при этом по оси абцисс представлены точки данных задачи, по оси ординат – степени принадлежности данного значения к найденным кластерам.

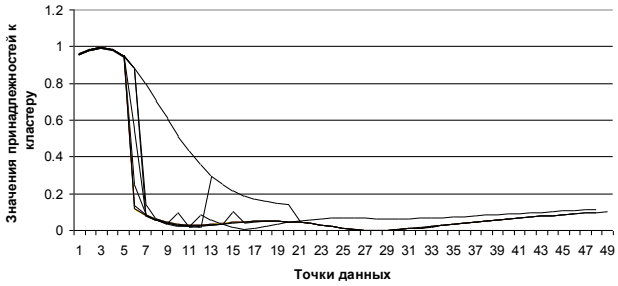

Рисунок 2. График значений принадлежности точек данных к первому кластеру

На основе полученных графиков проведем формирование терм – множества лингвистической переменной в треугольной форме. При этом возможны различные подходы к формированию результирующего терма – среднее, максимальное, минимальное значение принадлежности для точек данных. В работе используем

аппроксимация среднего значения принадлежностей для всех наборов данных. Результаты приведены на рис. 6. Последним этапом формирования адекватного терм – множества является интерпретация полученных результатов компетентным лицом – экспертом. Как видно из рис. 6, термы 2 и 3 располагаются очень близко и фактически перекрывают друг друга, поэтому экспертом может быть принято решение объединения этих термов в один без существенной потери точности.

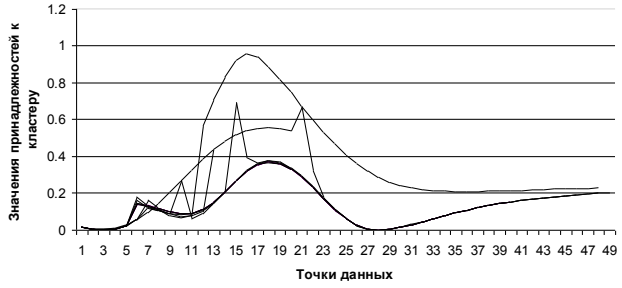

Рисунок 3. График значений принадлежности точек данных к второму кластеру

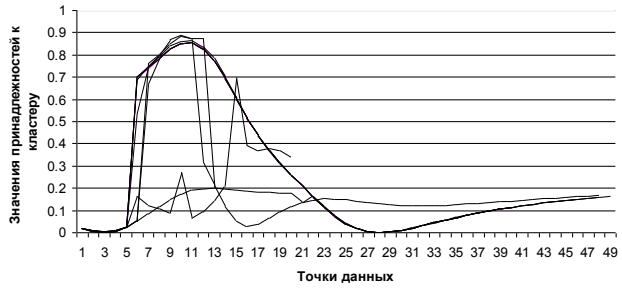

Рисунок 4. График значений принадлежности точек данных к третьему кластеру

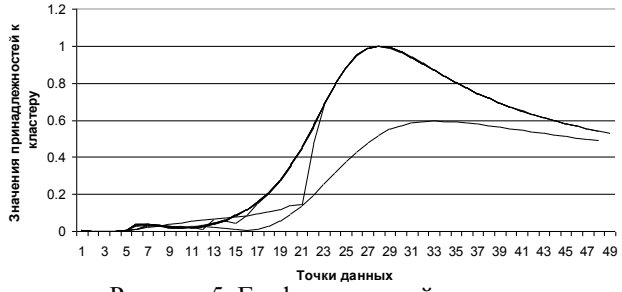

Рисунок 5. График значений принадлежности точек данных к четвертому кластеру

Как можно увидеть, применение разработанного эволюционного алгоритма позволило получить достаточно качественное разбиение пространства значений входной лингвистической переменной на нечеткие термы, обладающие приемлемым уровнем интерпретабельности экспертом.

В дальнейшем необходимо протестировать эффективность предлагаемого метода на других тестовых задачах, в том числе большей размерности, и сравнить результаты работы метода с другими известными методами кластеризации и формирования терм-множеств. Также необходимо более подробно рассмотреть проблему взаимосвязи точности и интерпретабельности результатов нечеткой кластеризации.

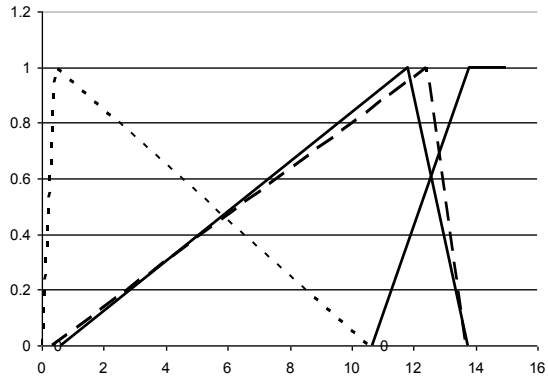

Рисунок 6. Полученные термы лингвистической переменной «Давление греющего пара»

Данный метод может использоваться для первоначального формирования термов лингвистических переменных, представленных временными рядами, при проектировании и настройке нечетких систем поддержки принятия решений. Планируется использовать разработанный метод для построения терм – множеств лингвистических переменных для идентификации нечеткой системы диагностики и управления процессом вулканизации.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК.

1. Dunn J.C. A Fuzzy Relative of the ISODATA Process and its Use in Detecting Compact Well Separated Clusters/ J.C. Dunn// J.Cyber. 1974. – C. 32-57

2. Емельянов В.В. Теория и практика эволюционного моделирования/В.В. Емельянов, В.В. Курейчик, В.М. Курейчик//М.: ФИЗМАТЛИТ. - 2003. – 432 с.

3. Egan M.A. Comparative Study of a Genetic Fuzzy C-Means Algorithm and a Validity Guided Fuzzy C-Means Algorithm for Locating Clusters in Noisy Data/M.A. Egan, M. Krishnamoorthy, K. Rajan//*In Proc. IEEE World Congress on Computational Intelligence*. 1998. - C. 440-445.

4. Hall L.O. Clustering with a Genetically Optimized Approach/L.O. Hall, I. B. Özyurt, J. C. Bezdek//*IEEE Trans. On Evolutionary Computation*. 1999. - №. 3. - C. 103-112.

5. Alves V.S. A Fuzzy Variant of an Evolutionary Algorithm for Clustering/V.S. Alves, R. J. G. B. Campello, E. R. Hruschka//*In Proc. IEEE Int. Conference on Fuzzy Systems*. 2007 - C. 375-380.

6. Hruschka E.R. Evolutionary Search for Optimal Fuzzy C-Means Clustering/E.R. Hruschka, R. J. G. B Campello, L. N. de Castro//*In Proc. Int. Conference on Fuzzy Systems*. 2004. - C. 685-690.

7. Xie X.L. A validity Measure for Fuzzy Clustering/X.L.Xie, G.A. Beni//IEEE Trans. on PAMI. 1999. - vol. 3. - №. 8. – C. 841-846.

8. Рутковская Д. Нейронные сети, генетические алгоритмы и нечеткие системы/Д. Рутковская, М. Пилиньский, Л. Рутковский.//М.: Горячая линия – Телеком. 2006. – 452с.