УДК 519.237.8

МОДЕЛЬ И МЕТОД КОНЦЕПТУАЛЬНОЙ КЛАСТЕРИЗАЦИИ ОБЪЕКТОВ, ХАРАКТЕРИЗУЕМЫХ НЕЧЕТКИМИ ПАРАМЕТРАМИ

Назаров А.О.

ФГБОУ ВПО «Казанский национальный исследовательский технический университет им. А.Н. Туполева – КАИ», Казань, e-mail: sas4406@yandex.ru

Разработан новый метод, представляющий собой аналог метода концептуальной кластеризации COBWEB. Данный метод, в отличие от существующих методов кластеризации, позволяет работать с объектами, характеризуемыми нечеткими параметрами и строить модель концептуальной кластеризации для объектов нечеткой природы. Основу метода составляет предложенная в работе модифицированная формула оценки полезности концептуальной кластеризации для объектов, характеризуемых нечеткими параметрами. Проведены численно-параметрические исследования разработанного метода концептуальной кластеризации по отношению к используемым функциям принадлежности. Полученные в работе теоретические результаты были использованы для решения ряда практических задач по кластеризации объектов, характеризуемых нечеткими параметрами, и проведены экспериментальные исследования для сравнительной оценки точности кластеризации. На примере задачи кластеризации животных была сравнена точность разработанного метода кластеризации с другими известными методами.

Ключевые слова: метод кластеризации, нечеткие параметры, функции принадлежности

MODEL AND METHOD OF CONCEPTUAL CLUSTERING OF OBJECTS CHARACTERIZED BY FUZZY PARAMETERS

Nazarov A.O.

A. Tupolev Kazan State Technical University – KAI, Kazan, e-mail: sas4406@yandex.ru

The new method corresponds to the method of conceptual clustering was elaborated. This method, unlike the existing clustering methods, allows you to work with objects, characterized by fuzzy parameters and build a model of conceptual clustering for fuzzy objects of nature. The method is based on the modified mathematical formula which evaluates the usefulness of the conceptual clustering of objects characterized by fuzzy parameters. Numerical-parametric researches of the method of developed a conceptual clustering were conducted with respect to the membership functions. The theoretical results obtained in the research have been used to solve some practical problems of clustering the objects characterized by fuzzy parameters, and conducted experimental study of comparative evaluation of the accuracy of clustering. By the example of clustering of animals the accuracy of this method was compared with other known clustering methods.

Keywords: clustering method, fuzzy parameters, membership function

Задача кластеризации является одной из важнейших задач интеллектуального анализа данных в различных проблемных областях – технических, естественнонаучных, социальных. Кластеризация является примером задачи обучения без учителя и сводится к разбиению исходного множества объектов на подмножества классов таким образом, что элементы одного класса были бы схожи между собой, а элементы различных классов были бы максимально различны [2].

Основной сложностью применения классических методов кластеризации для решения практических задач является то, что многие реальные объекты могут быть описаны исключительно нечеткими параметрами. В связи с этим для кластеризации подобных объектов в последнее время активно развиваются методы нечеткой кластеризации [3]. Исследованиям в данной области посвящены работы известных зарубежных и российских ученых: Bezdek J.C., Pedrycz W., Zadeh L.A., Аверкина А.Н., Батыршина И.З., Вагина В.Н., Вятченина Д.А., Ярушкиной Н.Г. и др.

Целью исследования является разработка модели, метода и комплекса программ концептуальной кластеризации объектов, характеризуемых нечеткими параметрами, на основе классического метода кластеризации COBWEB. Эффективность разрабатываемого метода концептуальной кластеризации определяется его способностью работать с объектами, характеризуемыми параметрами с нечеткими значениями, и достигаемой точностью кластеризации.

Материалы и методы исследования

В настоящее время известно множество алгоритмов нечеткой кластеризации, таких как Fuzzy C-Means, FOPTICS и др. [7]. Данные алгоритмы формируют кластеры, границы которых размыты, а объект может принадлежать более чем одному кластеру с различными степенями принадлежности. Однако следует отметить, что большинство алгоритмов нечеткой кластеризации работают с четкими значениями параметров объектов, формируя кластеры, например, на основе оценки расстояний между объектами и центром кластера. Такой подход не позволяет эффективно осуществлять кластеризацию объектов с нечетко заданными значениями параметров.

В связи с этим актуальной задачей является разработка методов кластерного анализа, способных учитывать нечеткую природу объектов, то есть работать с параметрами, заданными в нечеткой форме в виде функций принадлежности.

Для решения многих практических задач в настоящее время используется концептуальная кластеризация данных, ярким представителем которой является метод COBWEB [6]. Классический вариант реализации метода COBWEB не предполагает работу с параметрами, заданными в нечеткой форме, что актуализирует решение поставленной выше задачи для данного метода.

Для формализации метода кластеризации COBWEB, обозначим через $O = \{O_i\}_{i=\overline{1,r}}$ – множество распознаваемых объектов, характеризуемое бинарными параметрами $A = \{A_j\}_{j=\overline{1,m}}$, принимаемыми одно из возможных значений $V_{ij} \in \{0;1\}$. $\{C_0, C_1, ..., C_n\}$ – множество формируемых кластеров, где n – заранее неизвестно.

Полезность кластеризации в методе COBWEB рассматривается как функция CU, определяющая сходство объектов в рамках одного кластера, и их различие по отношению к объектам из других кластеров. Внутриклассовое сходство определяется условной вероятностью $P(A_j = V_{ij} \mid C_k)$, а межклассовое сходство условной вероятностью $P(C_k \mid A_j = V_{ij})$.

Функция полезности кластеризации определяется согласно [6] в виде

$$CU = \frac{\sum_{k=1}^{n} P(C_k)\left[\sum_j \sum_i P(A_j = V_{ij} \mid C_k)^2 - \sum_j \sum_i P(A_j = V_{ij})^2\right]}{n}, \quad (1)$$

где n – количество кластеров.

Метод COBWEB строит дерево классификации с вероятностными описаниями концептов. Выбор возможного способа кластеризации объектов основан на значениях функции полезности кластеризации (1). При построении дерева классификации используются следующие 4 операции [6]:

– отнесение объекта к наилучшему из существующих кластеров;

– добавление нового кластера, содержащего единственный объект;

– слияние двух существующих кластеров в один новый с добавлением в нее этого объекта;

– разбиение существующего кластера на два и отнесение объекта к лучшему из вновь созданных кластеров.

Предлагается модель концептуальной кластеризации объектов в виде дерева, представленного на рис. 1.

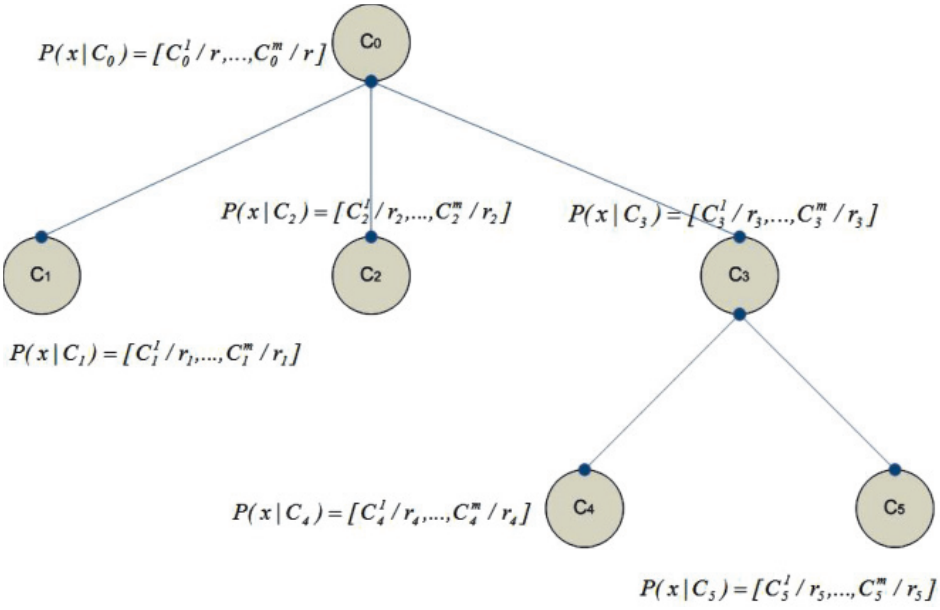

Рис. 1. Модель концептуальной кластеризации

Где C_k^j – количество раз, когда значение параметра $A_j = 1$ для объектов в кластере C_k, r_k – количество объектов в кластере C_k.

Пошаговое описание метода концептуальной кластеризации:

1. Вводится корневой кластер C_0, свойства которого совпадают со свойствами первого объекта $O1 = [V11, ..., V1m]$. Для каждого последующего объекта $Oi = [Vi1, ..., Vim]$ выполняется цикл, реализующий шаги 2–6, в рамках которых выполняются 4 выше представленные операции.

2. Объект Oi добавляется поочередно в кластеры $C1, C2, ..., Ck$. После каждого добавления вычисляется полезность кластеризации $CU1, ..., CUk$.

3. Для объекта Oi создается новый кластер $Ck + 1$, объект помещается в кластер и вычисляется полезность кластеризации $CUk + 1$.

4. Объединяются два кластера с максимальными значением полезности кластеризации из $CU1, ...,$

CUk. Образуется новый кластер, в него добавляется объект *Oi*. Вычисляется полезность кластеризации *CUk* + 2.

5. Объект *Oi* добавляется в кластер с максимальным значением полезности кластеризации из *CU*1 ,..., *CUk*. Образуется новый кластер с двумя кластерами-потомками. Вычисляется полезность кластеризации *CUk* + 3.

6. Выбирается максимальное значение полезности кластеризации среди полезностей *CU*1,..., *CUk*,*CUk* + 1, *CUk* + 2, *CUk* + 3, в соответствии с ним выбирается операция разбиения объектов по кластерам.

В статье предложена модификация метода концептуальной кластеризации, основанная на методе COBWEB, позволяющая работать с объектами, характеризуемыми нечеткими параметрами. Данный метод предполагает реализацию классического метода концептуальной кластеризации COBWEB в следующих условиях:

1. Множество распознаваемых объектов $O = \{O_i\}_{i=\overline{1,r}}$, характеризуется нечеткими параметрами $\tilde{A} = \{\tilde{A}_j\}_{j=\overline{1,m}}$.

2. Значение параметра $\tilde{A}_j$ для объекта O_i определяется в виде функции принадлежности $\mu_{\tilde{A}_{ij}}(x) \in \{0;1\}$.

3. Степень сходства двух функций принадлежности $\mu_{\tilde{A}_{ij}}(x)$ и $\mu_{\tilde{A}_{ij}}(x)$ определяется их наибольшей верхней границей [5] в виде

$$v_{jit} = \sup_{x \in X} \min_{x \in X} \left\{ \mu_{\tilde{A}_{ij}}(x), \mu_{\tilde{A}_{ij}}(x) \right\} \in [0,1], \qquad (2)$$

где $\mu_{\tilde{A}_{ij}}(x)$ – функция принадлежности параметра $\tilde{A}_j$ для объекта O_i, а $\mu_{\tilde{A}_{ij}}(x)$ – функция принадлежности параметра $\tilde{A}_j$ для объекта O_t.

4. Основываясь на формуле полезности кластеризации (1) классического метода COBWEB и условиях 1–3, оценка полезности кластеризации осуществляется по модифицированной формуле (3)

$$CU^* = \frac{\sum_{k=1}^{n} P(C_k) \left[\sum_{j=1}^{m} \sum_{i=1, O_i \in C_k}^{r} \sum_{t=1, O_t \in C_k}^{r} v_{jit} / |C_k| - \sum_{j=1}^{m} \sum_{i=1}^{r} \sum_{t=1}^{r} v_{jit} / r \right]}{n}, \qquad (3)$$

где n – количество кластеров.

Также были проведены численно-параметрические исследования разработанного метода по отношению к используемым функциям принадлежности. Рассмотрено два основных типа функций принадлежности: кусочно-линейные и П-образные[1].

Наибольшая верхняя граница П-образных функций $f_{\mu(x)}(x; a_{\mu(x)}, b_{\mu(x)}, c_{\mu(x)}, d_{\mu(x)})$ и $f_{t(x)}(x; a_{t(x)}, b_{t(x)}, c_{t(x)}, d_{t(x)})$ может быть представлена в виде точки пересечения S-образных функций $f^S_{\mu(x)}(x; c_{\mu(x)}, d_{\mu(x)})$ и $f^S_{t(x)}(x; a_{t(x)}, b_{t(x)})$ [1]. Исходя из способа задания П-образных функций принадлежности, точку пересечения П-образных функций принадлежности можно представить как

$$x = \frac{a_{t(x)} c_{\mu(x)} - b_{t(x)} d_{\mu(x)}}{c_{\mu(x)} - d_{\mu(x)} - b_{t(x)} + a_{t(x)}}, \qquad (4)$$

Исходя из (4) получаем степень сходства функций принадлежности $f_{\mu(x)}(x; a_{\mu(x)}, b_{\mu(x)}, c_{\mu(x)}, d_{\mu(x)})$ и $f_{t(x)}(x; a_{t(x)}, b_{t(x)}, c_{t(x)}, d_{t(x)})$.

Пересечение двух кусочно-линейных функций принадлежности, исходя из способа задания кусочно-линейных функций принадлежности, определяется как

$$x = \frac{b_{t(x)} c_{\mu(x)} + b_{\mu(x)} a_{t(x)}}{c_{\mu(x)} + a_{t(x)}}. \qquad (5)$$

Исходя из (5), получаем степень сходства функций принадлежности.

Выбор типа функции принадлежности основан на практических результатах, полученных при решении задачи кластеризации конкретных объектов. Экспериментальным путем показано, что определение значений параметров объектов в виде кусочно-линейных функций принадлежности позволяет увеличить разделяющую способность кластеров в разработанном методе по сравнению с П-образными функциями. Таким образом, использование кусочно-линейных функций принадлежности для описания нечетких параметров объектов более предпочтительно по сравнению с П-образными функциями.

Результаты исследования и их обсуждение

В качестве примера была решена задача автоматизации формирования пользовательских ролей [4] на действующей информационной системе конкретной организации. Осуществлялась кластеризация 22 пользователей: $O = \{O_i\}_{i=1}^{22}$. Для описания поведения пользователей было выделено 18 параметров $A = \{A_j\}_{j=1}^{18}$. На основании анализа поведения пользователей по выделенным параметрам осуществлялась кластеризация и распределение пользователей по кластерам. В результате кластеризации было выделено 10 кластеров. Данные кластеры содержали пользователей в соответствии с их функциональными обязанностями. Также было выделено два кластера с аномальным поведением пользователей.

Решение данной задачи позволяет, с одной стороны, значительно упростить работу администратора информационной

безопасности по формированию пользовательских ролей в ИС, с другой стороны, позволяет обнаруживать аномальное поведение пользователей в ИС, выявляя недобросовестных сотрудников, использующих информационные ресурсы организации не только для выполнения своих функциональных обязанностей, но и в личных целях.

Вторая практическая задача, решенная с помощью разработанного метода концептуальной кластеризации, стояла в распределении животных по кластерам на основе их параметров, заданных в нечетком виде. Для каждого семейства животных (медвежьи, зайцевые, кошачьи) взята выборка по 7 видов – $O = \{O_i\}_{i=1}^{21}$. Каждое животное было описано 3 параметрами, описанными в нечетком виде (длина тела, вес, скорость) – $A = \{A_j\}_{j=1}^{3}$.

В результате работы разработанного метода концептуальной кластеризации объекты были распределены по 3 кластерам в соответствии с семействами животных. На примере решения данной задачи разработанный метод нечеткой концептуальной кластеризации показал 100 % точность кластеризации.

Для сравнительного анализа данная задача также была решена с помощью известных методов кластеризации EM и g-means. При этом выполнялась дефазификация параметров объектов, заданных в нечетком виде. Результаты сравнительного анализа точности кластеризации методов представлены на рис. 2.

Разработанный метод концептуальной кластеризации *(точность 100%)*

Кластер	Объект (Животное)
C_1	$O_1, O_2, O_3, O_4, O_5, O_6, O_7$
C_2	$O_8, O_9, O_{10}, O_{11}, O_{12}, O_{13}, O_{14}$
C_3	$O_{15}, O_{16}, O_{17}, O_{18}, O_{19}, O_{20}, O_{21}$

Метод кластеризации g-means *(точность 76,1%)*

Кластер	Объект (Животное)
C_1	O_{15}
C_2	$O_{16}, O_{17}, O_{18}, O_{19}, O_{20}, O_{21}$
C_3	$O_8, O_9, O_{10}, O_{11}, O_{12}, O_{13}$
C_4	O_3, O_4, O_5
C_5	O_1, O_2, O_6, O_7
C_6	O_{14}

Метод кластеризации Expectation Maximization *(точность 80,9%)*

Кластер	Объект (Животное)

Рис. 2. Сравнительный анализ методов кластеризации

Таким образом, видим, что точность решения задачи кластеризации методами EM и g-means составила соответственно 89 и 76,1 %, что меньше точности, полученной в результате работы разработанного метода.

Заключение

Разработанный метод концептуальной кластеризации, в отличие от существующих методов кластеризации, позволяет работать с объектами, характеризуемыми нечеткими параметрами, и строить модель концептуальной кластеризации для объектов нечеткой природы. Основу метода составляет предложенная в работе модифицированная формула оценки полезности концептуальной кластеризации для объектов, характеризуемых нечеткими параметрами. Экспериментальным путем показано, что использование кусочно-линейных функций принадлежности для задания значений нечетких параметров объектов позволяет увеличить разделяющую способность кластеров в разработанном методе по сравнению с П-образными функциями. Полученные в работе теоретические результаты были использованы для решения задачи автоматизации формирования пользовательских ролей в корпоративной информационной сети, включающей в себя 22 пользователя, каждый из которых описывался 18 параметрами. Полученные результаты позволили выделить пользо-

вателей, характеризующихся аномальным поведением в компьютерной сети. На примере решения задачи кластеризации разработанный метод показал 100 % точность. На тех же данных точность известных методов кластеризации EM и g-means составила соответственно 89 и 76,1 %.

Список литературы

1. Алексеев А.В. Интерпретация и определение функций принадлежности нечетких множеств // Методы и системы принятия решений: сб. тр. / под ред. А.Н. Борисова. – Рига: РПИ, 1979. – С. 42–50.

2. Методы и модели анализа данных: OLAP и Data Mining. – СПб.: БХВ-Петербург, 2004. – 336 с.

3. Вятченин Д. А. Нечеткие методы автоматической классификации: монография. – Мн.: УП «Технопринт», 2004 – 219 с.

4. Гайдамакин Н.А. Разграничение доступа к информации в компьютерных системах. – Екатеринбург: изд-во Урал. Ун-та, 2003 г. – 328 с.

5. Севастьянов П.В., Венберг А.В. Конструктивная методика сравнения нечетких чисел и ее применение в задачах оптимизации // Информационные сети, системы и технологии: Тр. VII междунар. конф., БГЭУ, 2–4 окт. 2001 г.: В 3-х томах. – Т. 3. – Минск, 2001. – С. 52-57.

6. Fisher D. Knowledge Acquisition Via Incremental Conceptual Clustering, 1987. – P. 142–153.

7. Sato M., Sato Y., and Jain L. Fuzzy Clustering Models and Applications, Physica-Verlag, Heidelberg, 1997. – P. 135–148.

References

1. Alekseev A.V. Interpretation and definition of membership functions of fuzzy sets. Methods and systems of decision-making. Riga, 1979. pp. 42–50.

2. Methods and models for data analysis: OLAP and Data Mining. St. Petersburg, 2004. 336 p.

3. Vyatchenin D.A. Fuzzy automatic classification methods: Monograph. Moscow, 2004. 219 p.

4. Gaydamakin N.A. Concurrent access to information in computer systems. Ekaterinburg, 2003. 328 p.

5. Sevastiyanov P.V., Venberg A.V. Constructive framework for comparing fuzzy numbers and its application in optimization problems. Information Networks, Systems and Technologies. (Proc. VII Intern. conf., BGEU). Minsk, 2001. pp. 52–57.

6. Fisher D. Knowledge Acquisition Via Incremental Conceptual Clustering, 1987. pp. 142–153.

7. Sato M., Sato Y., and Jain L. Fuzzy Clustering Models and Applications, Physica-Verlag, Heidelberg, 1997. pp. 135–148.

Рецензенты:

Райхлин В.А., д.ф.-м.н., профессор кафедры «Компьютерные системы», ФГБОУ ВПО «КНИТУ им. А.Н. Туполева – КАИ», г. Казань;

Захаров В.М., д.т.н., профессор кафедры «Компьютерные системы», ФГБОУ ВПО «КНИТУ им. А.Н. Туполева – КАИ», г. Казань.

Работа поступила в редакцию 15.07.2014.