A Knowledge Context Fuzzy Clustering Method Based on Genetic Algorithm

Faping Zhang[1], *Li* Li[1,*], and *Cuixiang* Zhou[2]

[1]School of Mechanical Engineering, Beijing Institute of Technology, Beijing, China
[2]No.208 Research Institute of China Ordnance Industries, Beijing, China

Abstract. A fuzzy clustering method based on genetic algorithm is proposed aiming at the problem of automatic clustering of knowledge context. Firstly, the knowledge context model is constructed to determine the similarity measure of knowledge context. Then the initial clustering centers are obtained based on the density peak method. Then the fuzzy C mean clustering result is solved by genetic algorithm, and the clustering of knowledge context is realized. Finally, the knowledge context clustering of an aircraft part design process is taken as an example to illustrate the effectiveness of the algorithm.

1 Introduction

Knowledge plays an increasingly important role as an intangible resource in improving the production efficiency, reducing product costs, especially in knowledge-intensive enterprises.

Knowledge context is the relevant background environment on which knowledge acquisition, understanding, application, evaluation, sharing depend on.

Enterprise knowledge management system, records the user's knowledge behaviour in different knowledge context, and gets the internal relationship between the knowledge context and knowledge through data mining and other ways.

When the business people face a new business context, the system calculates the similarity between the current context and all the knowledge contexts, finds out a set of the most similar knowledge contexts, and then pushes the knowledge associated with the set of knowledge contexts to the users. However, with the accumulation of the number and types of knowledge contexts, the computational similarity of traversal undoubtedly greatly affects the search efficiency. In order to solve this problem, this paper proposes clustering similar knowledge contexts, which only need to be calculated in the class of the current context to get the similarity, so as to improve the computational efficiency and knowledge reuse efficiency.

The main clustering methods include hierarchical clustering [1], density- based [2], grid-based [3] and model-based methods [4], and have been widely used in many fields such as pattern recognition [5], data analysis [6], image processing [7], market research [8] and so on.

Aiming at the clustering problem of knowledge context, this paper proposes a knowledge context fuzzy clustering method based on genetic algorithm. Firstly, the knowledge context model is constructed and the similarity measure method is put forward. Then, the cluster is clustered by the density peak method as the initial solution. Then the genetic fuzzy C-means clustering algorithm is used to realize the clustering of knowledge context. Finally, an example is given to illustrate the effectiveness of the algorithm.

2 Knowledge context model and the similarity measure

2.1. Knowledge context model construction

In the enterprise knowledge management process, the knowledge context not only includes the corresponding business process and task characteristics, but also includes the personality characteristics of users. The knowledge context model is used to describe the context in the form of extensible framework. The enterprise environment is complicated, so it is necessary to consider the characteristics of the task itself, the relationship with the business process, users' cognitive level and knowledge preference.

This section constructs a multi-dimensional multi-attribute context model for the business process of the enterprise in order to describe the context elements of different dimensions more accurately and make the context model more manageable.

The context model is characterized by four main dimensions: person, task, object, and software.

Context={Person, Task, Object, Software}

There are specific portrayals under the four main dimensions:

The person, including personnel roles, personnel skills level, the field described, is

Person = {P-position, P-skills, P-domain}.

* Corresponding author: bitlili@163.com

The task, including the task name, task objectives, tasks required resources, the relationship between tasks and the process, is

Task = {T-name, T-goal, T-resource, T-relation}.

The object, including the object name, object material, and object technical characteristics, is

Object = {O-name, O-material, O-TecFeatures}.

The software, including the software name, software version, software use, is

Software = {S-name, S-version, S-usage}.

After the refinement, the context model is

Context= {P-position, P-skills, P-domain, T-name, T-goal, T-resource, T-relation, O-name, O-material, O-TecFeatures, S-name, S-version, S-usage}.

It should be noted that the enterprise can be constructed according to the actual environment, including but not limited to the above attributes.

2.2 Knowledge context similarity

Clustering is a process that divides a data set into different categories or clusters and makes the properties of objects in the same category the same or similarities, and the properties of the objects in different categories are significantly different. The distance between the knowledge context is measured by the knowledge context similarity, the higher the degree of similarity, the closer the distance is. This section describes the similarity of knowledge context.

Knowledge contexts are portrayed from multiple dimensions, and the attributes of different dimensions may not be the same. The attribute type is divided into four types: numeric attribute, vector type attribute, collection type attribute, text type attribute. Numerical types and vector types can be calculated directly, and the collection type and text type attributes need to be vectorised first.

The vectorization process is:

Step1: the construction of the domain vocabulary. For example, the total number of a set of possible vocabulary is m, then the m words in order to form a vocabulary.

Step2: vectorization.

This type is described by a m-dimensional vector. For a collection type, the {0,1} value indicates whether a vocabulary is present in the set. For text attributes, the TF-IDF value for each word in the text as the value is the vector component.

After the pre-treatment, the n knowledge context models can be expressed by the knowledge context model vector matrix:

$$\sum = \begin{bmatrix} \text{the first attribute} & \overbrace{\text{the second attribute (k-1 dimension vector)}} & & & & \\ x_{11} & x_{12} & \cdots & x_{1k} & \cdots & x_{1h} \\ x_{21} & x_{22} & \cdots & x_{2k} & \cdots & x_{2h} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} & \cdots & x_{nh} \end{bmatrix}$$

For vector type attributes and numeric type attributes, different methods are used for the calculation.

a vector attribute calculation

$$sim_v(v,v') = \frac{\sum_{i=1}^{m}(q_i-\bar{q})(p_i-\bar{p})}{\sqrt{\sum_{i=1}^{m}(q_i-\bar{q})^2}\sqrt{\sum_{i=1}^{m}(p_i-\bar{p})^2}}$$

where v and v 'denote the m-dimensional vector type attributes, the values are (q_1, q_2, q_3, ..., q_m) and (p_1, p_2, p_3, ..., p_m), The similarity of the vector type attribute is in the range of [-1,1]. The larger the value, the higher the similarity.

b numerical attribute calculation

$$sim_n(v,v') = \frac{\min\{n_1,n_2\}}{\max\{n_1,n_2\}}$$

Where v, v 'denote the numeric type attribute, the values are n_1、n_2, n_1、n_2> 0. The value range of the value type attribute is (0,1). The larger the value, the higher the similarity.

The context similarity is the weighted average of the similarity of each attribute:

$$sim(V,V') = \frac{\sum_{i=1}^{s} w_i \cdot sim(v_i,v_i')}{\sum_{i=1}^{s} w_i}$$

Where V, V'denote two contexts, there are s attributes, v_i, v_i' denote the i-th attribute, and w_i the weight of the i-th attribute.

3 Improved fuzzy C - means clustering algorithm for initial value selection

3.1 Fuzzy C-means clustering algorithm

Ruspini[9] introduced the Zadeh fuzzy mathematical theory[10] into clustering analysis, and proposed the concept of fuzzy partitioning. The membership degree $u_{ij} \in \{0,1\}$ in traditional hard clustering is improved to $u_{ij} \in [0,1]$ continuous interval. Later Dunn and Bezdek promoted it to form the widely used fuzzy c-means algorithm (FCM).

The FCM algorithm determines the membership of the sample by minimizing the objective function to further determine the class of the sample.

Assume Knowledge Context Data Set

$$X = \{x_1, x_2, ..., x_k, ..., x_n\} \subset R^h$$

Where n is the number of elements in the data set X, the sample has s attribute value, h attribute dimension (h $\geqslant$s).

The matrix V of the c clusters is:

$$V=\begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1h} \\ v_{21} & v_{22} & \cdots & v_{2h} \\ \vdots & \vdots & & \vdots \\ v_{c1} & v_{c2} & \cdots & v_{ch} \end{bmatrix}$$

Where c is the number of clusters, $c \in [2,n)$. v_{ij} is the value of the j-th attribute dimension of the i-th cluster center.

The objective function is expressed as:

$$V=\begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1h} \\ v_{21} & v_{22} & \cdots & v_{2h} \\ \vdots & \vdots & & \vdots \\ v_{c1} & v_{c2} & \cdots & v_{ch} \end{bmatrix}$$

Where c is the number of clusters, $c \in [2,n)$. the v_{ij} is the value of the j-th attribute dimension of the i-th cluster center.

The objective function is expressed as:

$$J_m(U,V)=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^{m}.d_{ij}^{2} \tag{1}$$

Constraints:

$$\begin{cases} \sum_{j=1}^{c}u_{ij}=1 & 1\le j\le n \\ 0\le u_{ij}\le 1 & 1\le i\le n, 1\le j\le c \\ 0\le \sum_{i=1}^{n}u_{ij}\le n & 1\le i\le c \end{cases}$$

Among them, uij is the membership degree of the i-th sample belongs to the j-th cluster center, and m is the fuzzy index. the larger the m, m≥1, the lower the degree of fuzzy clustering. $U=[u_{ij}]_{n\times c}$ is the membership matrix.

The d_{ij} represents the distance from the i-th sample to the j-th cluster center, where the distance is characterized by similarity (section 2.2).

The Lagrangian multiplier method is used to solve the membership degree. The formula (2) is used to calculate membership, and the clustering center can be updated by the formula (3):

$$u_{ij}=\frac{1}{\sum_{k=1}^{c}\left(\frac{d_{ij}}{d_{ik}}\right)^{\frac{2}{m-1}}} \tag{2}$$

$$v_j=\frac{\sum_{k=1}^{n}\left(u_{kj}\right)^{m}x_k}{\sum_{k=1}^{n}\left(u_{kj}\right)^{m}} \tag{3}$$

In the formula (2) and (3), i = 1, ..., n, j = 1, ..., c.

3.2 Algorithm to improve the initial value of the selected

The FCM algorithm is sensitive to the selection of the initial clustering center. The choice of the initial solution is closely related to the iterative time and the iterative result. If the initial solution is improperly chosen, the iteration time is longer and it is easy to fall into the local optimum. In general, the initial clustering center, that is, the initial solution for the objective function formula (1), is randomly generated within a range of values, which greatly affects the clustering effect.

In this paper, the density peaks clustering proposed by Rodriguez is used in the selection of the initial solution. This method is based on the analysis of the characteristics of the clustering center, and characterizes the data points by local density ρ (x_i) and distance δ (x_i). In general, the clustering center of these two parameters are higher.

Define the local density as:

$$\rho(x_i)=\sum_{j}\chi(d_{ij}-d_c),$$
$$\chi(\Delta d)=\begin{cases}1 & \Delta d<0\\ 0 & \Delta d\ge 0\end{cases} \tag{4}$$

Where ρ(x_i) is the number of data points which the distance with xi is less than d_c, and d_c is the truncated distance, given by the user. The d_{ij} is the distance between the data points x_i and x_j, expressed by similarity (section 2.2).

The data point distance δ (x_i) is the distance from the nearest point of xi whose local density is greater than xi, which is defined as follows:

$$\delta(x_i)=\begin{cases}\min\limits_{j:\rho(x_j)>\rho(x_i)}(d_{ij}), & \rho(x_i)\ne\rho_{\max};\\ \max\limits_{j}(d_{ij}), & \rho(x_i)=\rho_{\max}\end{cases} \tag{5}$$

After the local density and distance of the data points are obtained, the relationship between the two parameters is shown graphically, called the decision graph. It is convenient and quick to determine points with high density and distance with the help of decision graphs. The clustering center determined by this method is used as the initial clustering center of FCM algorithm.

4 Genetic optimization FCM clustering algorithm

Genetic algorithm is a random parallel search algorithm based on natural selection and genetics. It is an efficient method to find the global optimal solution without any initial information. The solution of the problem is treated as a population, and the quality of the solution is getting

better and better through selection, crossover, mutation and other genetic operations.

The process of genetic algorithm optimization as shown below.

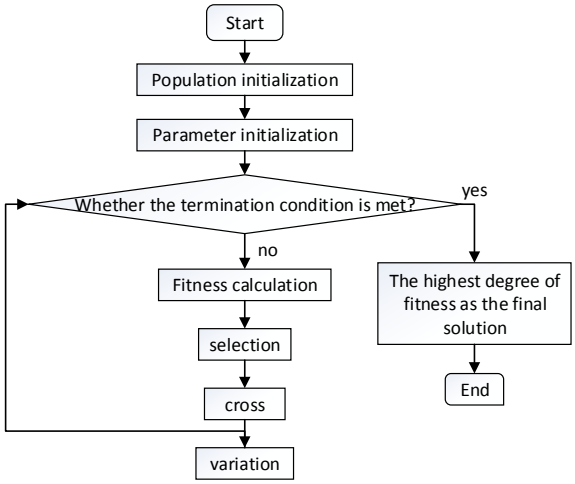

Fig. 1. Genetic algorithm flow chart

The coding, fitness function, population initialization, termination condition, crossover operator need to be determined when the clustering algorithm is optimized by genetic algorithm. The following is a detailed description of the fuzzy clustering problem which is used to identify the knowledge context.

4.1 Encoding

The solution for the fuzzy clustering problem of knowledge context is the matrix of clusters, which is expressed with V. The original data points are numbered in order to facilitate the calculation, and the representation of the data points is numbered instead of vector. If an individual is represented as (3, 6, 8, 10, 18), the meaning is that the cluster center is made up of dots numbered 3, 6, 8, 10, 18. The decimal number is converted to a binary number in actual calculations.

4.2 Population initialization

The density peak method is used to determine the initial population. The density peak method determines the clustering center with a certain subjectivity, and the same decision map may yield different results. Therefore, we can determine the m solution according to the decision graph. As the initial solution set, the number of clusters in each solution should be consistent.

4.3 Fitness calculation

The optimal result of the clustering algorithm corresponds to the minimum value of the objective function, that is, the better the clustering result, the smaller the objective function. In the genetic algorithm, the fitness of the global optimal solution should be the largest. So, the fitness function is defined as:

$$f(v_i) = \frac{1}{J_m(U,V) + \varepsilon_0} \tag{6}$$

ε_0 is a sufficiently small positive number, which avoids the denominator to zero. When the fitness calculation is carried out, the data set X and a set of clustering centers are known. Then, the degree of membership of the data set point to the cluster center is calculated according to formula (2), and then calculated $J_m(U,V)$ according to formula (1). Finally, the degree of fitness is obtained by formula (6).

4.4 Selection

The selection operation makes it possible to select the better individuals in the current population. First, select the individuals with the best fitness in the current population and do not enter the next generation directly. Then, the remaining individuals are selected according to the "proportional selection method". This method is to calculate the individual fitness, the probability that each individual is selected is:

$$P(V_i) = \frac{f(V_i)}{\sum_{j=1}^{N} f(V_j)}$$

i = 1, ..., n. the n is the population size.

4.5 Cross

Cross-operation is the main method of generating new individuals in genetic algorithm. The design of the crossover operator includes how to determine the position of the cross and how to cross it.

Common cross-cutting methods include single-point crossover, two-point crossover, multi-point crossover, uniform crossover and so on. The single-point crossover is used in this case.

The cross formula is as follows:

$$v_A^{t+1} = \alpha v_B^t + (1-\alpha) v_A^t$$
$$v_B^{t+1} = \alpha v_A^t + (1-\alpha) v_B^t$$

The v^t is the value before crossover, and the v^{t+1} is the value after crossover.

4.6 Variation

Variation can maintain group diversity, and it is an auxiliary method that generate a new generation of individual. Set a variation probability P_m, randomly generate a random number r_m, if $r_m < P_m$, for mutation. If the individual adaptation value after mutation is greater than the parent, the individual substitute parent, and become the offspring individual.

4.7 Termination conditions

In this case, the genetic termination condition is set to: the iteration is stopped when the solution of the best fitness is no longer in the solution set, or the iteration number reaches the maximum iteration times L.

5 Example

The knowledge contexts of the development of an aircraft parts are clustered and the number of extraction contexts is 63. This problem is solved by MATLAB iterations. The density peak method is used to calculate the density and distance of each data point to determine the initial solution. The decision chart is shown below:

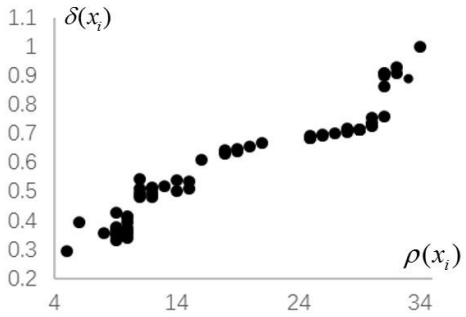

Fig. 2. "$\rho(x_i)$-$\delta(x_i)$"decision chart.

The number of population is 20. We selected 10 groups as the initial values in the upper right area of the decision graph. The other 10 groups are randomly generated to form the initial solution.

The probability of crossing is 0.6, the probability of mutation is 0.01, and the number of iterations is 100. The results of the iteration are as follows:

Table 1. Iterative results.

	value
optimal fitness	0.2932
the algebra of the optimal result	8

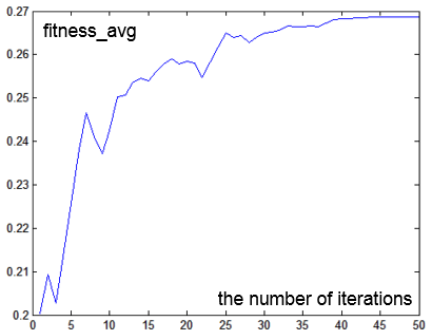

Fig. 3. The process of algorithm convergence.

6 Conclusion

The knowledge context clustering method proposed in this paper is helpful to the construction and reorganization of the knowledge base, which provides a strong support for the knowledge service. In this paper, the density peak method is used to determine the initial solution, then the genetic FCM method is used to solve the problem. In this way, the satisfactory solution can be obtained in a short time.

However, there are still many shortcomings in this paper. The weight of each dimension in knowledge context is not measured by objective data, and the number of clustering centers is determined by subjectivity. In the future, we will use the clustering results of knowledge context to solve the specific problems, and extract the knowledge application cases related to the business process to guide the business practice.

References

1. Murtagh F, Contreras P. Algorithms for hierarchical clustering: an overview[J]. Wiley Interdisciplinary Reviews Data Mining & Know,2012,2(1): 86-97.
2. Mansoori E G. GACH: a grid-based algorithm for hierarchical Clustering of high-dimensional data [J]. Soft Computing,2013, 18 (5): 905-922.
3. Kriegel H P, Kroger P, Sander J, et al. Density based Clustering [J]. Wiley Interdisciplinary Reviews Data Mining & Knowledge Discovery, 2011,1 (3): 231-240.
4. Bouveyron C, Brunet-SA C. Model-based clustering of high-dimensional data: a review [J]. Computational Statistics& Data Analysis, 2013,71 (1): 1-27.
5. Young J A, Fivelman Q L, BLAIR P L, et al. The plasmodium Falciparum sexual development transcriptome: a microarray Analysis using ontology-based pattern identification[J]. Molecular & Biochemical Parasitology, 2005, 143 (1):67-79.
6. Ren T, Zeng W, Wang N, et al. A novel approach for fMRI data analysis based on the combination of sparse approximation and affinity propagation clustering [J]. Magnetic Resonance Imaging, 2014, 32 (6): 736-746.
7. Moulick H N, Ghosh M. Image compression using k means Clustering and nuclear medicine image processing [J]. International Journal of Innovative Research in Computer & Communication Engineering, 2013, 1 (4): 869-877.
8. Grekousis G, Hatzichristos T. Fuzzy clustering analysis in geomarketing research [J]. Environment & Planning B Planning & Design, 2013, 40 (1): 95-116
9. Ruspini E H. A new approach to clustering[J], Information and control, 1969,15(1): 22-32.
10. Zadeh L. Fuzzy sets[J]. Information and Control, 1965, 8(3): 338-353.
11. Rodriguez A, Laio A. Clustering by fast search and find of density peaks [J]. Science, 2014, 344 (6191): 1492-1496.