Research and Improvement on K-Means Clustering Algorithm

Xue-mei Wang
Department of Computer Science and Technology,
ChengDong College of Northeast Agricultural
University
Harbin,150025 ,China
e-mail:wxm_lw@163.com

Jin-bo Wang
Liaoning Co., Ltd of China Mobile Group
Harbin, 150001,China
e-mail:wangjinbo@139.com

***Abstract*—According to the defects of classical k-means clustering algorithm such as sensitive to the initial clustering center selection, the poor global search ability, falling into the local optimal solution. A differential evolution algorithm which was a kind of a heuristic global optimization algorithm based on population was introduced in this article, then put forward an improved differential evolution algorithm combined with k-means clustering algorithm at the same time. The experiments showed that the method has solved initial centers optimization problem of k-means clustering algorithm well, had a better searching ability, and more effectively improved clustering quality and convergence speed.**

Keywords- differential evolution algorithm; K-means cluster algorithm; Cluster analysis

I. INTRODUCTION

Clustering is process which divided a given data set into several different clusters. In the Clustering algorithm, Cluster was a set of data objects and must meet the requirements as follows: the data objects had the higher similarity in the same cluster, and had the lower similarity in the different clusters. The guiding ideology of clustering was as far as possibly maximizing not only similarity of data objects in the same cluster, but also differetiation of data objects in different clusters.

Clustering analysis is an important researching field which involved in interdisciplinary object about data mining, pattern recognition, machine learning, statistics and so on. Mentioned by J.B.MacQueen, the K-means clustering algorithm was a typical partition algorithm . It was widely used in scientific research and industrial application besause of its advantages such as simplification, rapid convergence and suitable for processing large data sets and so on. In initialization of clustering center, classical K-means clustering algorithm just randomly assigned initial point, and usually found a local optimum clustering results. So, the poor stability affected the accuracy of classification. Therefore, it need a global optimized algorithm which can overcome the defluts of k-means[1].

Differential evolutionary algorithm was suggested by scholars R.S tore and K. Price in 1995. the evolutionary algorithm which was the code based on real number had the characteristics such as simple structure, good robustness and the strong global search ability. The intermediate test individuals were obtained by restructuring differential information from the current population individuals, following by comparing adaptive value between intermediate test individuals and the current evolutionary individuals. Finally, the prior evolutionary individuals which had been selected consisted in the next generational population. Each individual had a parallel evolutional process. During the evolutional process, fitness value of each individual in the population gradually escalated from one generation to the next[2].

Differential evolutionary algorithm which had higher accuracy but slower convergence speed was introduced into k-means clustering algorithm which had faster convergence speed but easily fell into the local optimum. The method could not only avoid the defaults of k-means algorithm and then achieve the higher quality of initial cluster centers, but also accelerate convergence speed of differential evolution algorithm. Based on the principles mentioned above, a new algorithm was suggested in this article. Compared with standard differential evolutionary algorithm, it should probably strength not only global searching ability, but also local searching ability. Finally, experiment the k-means clustering algorithm, the k-means clustering algorithm based on differential evolution and the improved algorithm had been conducted. The experimental results were analyzed and compared.

II. RELEVANT WORK

A. *The classical k-means clustering algorithm*

The K-means clustering algorithm was a partition method based on partition algorithm , which used square sum function of error as clustering criterion function[3].

$$E=\sum_{i=1}^{k}\sum_{j=1}^{m_j}\| x_{ij}-m_i \|^2 \qquad (1)$$

Among them, n_i express sample number of class i; x_{ij} express sample j of class i; m_i express clustering center or mass center of class i.

Essence of K-means clustering algorithm, which is repeated interative operation, is to find out the best clustering center of K numbers, then to make all the sample point distributing to the nearest clustering center, so that to minimize the clustering square error E. The specific operation procedure is as follows:

1) Initialization

Published by Atlantis Press, Paris, France.

K objects which is randomly selected serve as center of initial k clustering collection (m_1, m_2, …, m_k)；

2) Distribution x_i

Searching the nearest clustering center of each sample, and assign x_i to the corresponding clustering center.

3) Modifying clustering center

Calculate the mean value of every new divided clustering center.

$$m_i = \frac{1}{N_i} \sum_{j=1}^{N_j} x_{ij} \qquad (2)$$

4) Calculation of deviation and judgment

If E value converge, then return the best clustering center and terminate the algorithm else return step 2) to search repeatedly.

B. Improving the k-means clustering algorithm based on the differential evolution

The k-means clustering algorithm based on the differential evolution will randomly select the center from data set of clustering to encode, then structure initial population. Perform in the following operation of differential evolution algorithm such as mutation operations, crossover operation, selecting operation in order to obtain optimum individual[4]. The best individual was be decoded and clustered to the best initial clustering center which was abtained.Below are the details about k-means clustering algorithm based on differential evolution[5].

1) Initiation of population

Matrix X (N×d=D) will be generated to store data of current population. N express scale of the population, which represent combinating style of N group K clustering center. d express individual dimension in population. N is usually seted 5-10 times to d. K data samples which were randomly chosed from centralized data are served as an individual of initial population. It were repeated N times, and then structure the initial population.

$X_i(t)=(x_{i1}(t),x_{i2}(t),\ldots,x_{ij}(t),\ldots,x_{iK}(t))$ (i=1,2,… …,N)

The x_{ij} express clustering center j of individual i. Initial value of iterations t is 0.

2) Mutation operation

Mutation operation was based on vector difference of individual in current population. $X_i(t)$ was supposed as individual in current population. Three individuals $X_l(t)$,$X_m(t)$,$X_n(t)$which were randomly selected from current population (l≠m≠n≠i). Vector difference between individual $X_m(t)$ and $X_n(t)$ was supposed as disturbance factor, which was affected by scaling factor, and then plus individual $X_l(t)$. The mutated individual was $U_i(t)=(u_{i1}(t),u_{i2}(t),...,u_{ik}(t))$. $u_{ij}(t+1)=x_{lj}(t)+F(x_{mj}(t)-x_{nj}(t))$.F was Scaling factor.

3) Crossover operation

Cross operation was happen between the mutated individual u_{ij} (t + 1) and the current individual X_i (t) in population, and then generated K dimensional intermediate cross test individual $C_i(t+1)$, which was $C_i(t+1)=$ （$c_{i1}(t+1),c_{i2}(t+1),\ldots,c_{ik}(t+1)$）. The component j of individual was expressed as follows:

$$C_{ij}(t+1)=\begin{cases} u_{ij}(t+1) & if\ rand(0,1) \le CR\ \ or\ \ j = rand(i) \\ x_{ij}(t) & else \end{cases} \quad (3)$$

randj(0,1) was random number which obeys uniform distribution between 0 and 1; Crossover probability CR∈[0,1],it usually ranged from 0.3 to 0.9. Rand(i) was random integer between 0 and D.

4) selecting operation

the current evolutionary individual $X_i(t)$ was compared with the intermediate cross test individual $C_i(t+1)$, and then the best individual was selected into the next generation of population by using greedy algorithm. Fitness value of individual $f(X_i(t))=1/E$。

$$C_{ij}(t+1)=\begin{cases} X_i(t) & if\ \ f(X_i(t)) > f(C_i(t+1)) \\ C_i(t+1) & else \end{cases} \quad (4)$$

5) algorithm termination

Individuals in population X(t+1) was tested. The algorithm terminate if iterations met the condition, or the time of the same optimal results in the running process exceed the fixed value. Otherwise, Iterations plus 1, and then algorithm returned to step 2 for output best clustering center [6-7].

From the analysis mentioned above, differential evolution algorithm, which was applied in optimizing the initial clustering center of k-means clustering algorithm, could remarkably improve the clustering quality, and the design of algorithm structure was simple. Variation operation and cross operation of the algorithm not only ensure the diversity of evolution population, but also strengthen global search ability of algorithm. But the local search ability of the algorithm should be strengthened. Especially, the convergence speed in the later stage of evolution algorithm needed further improve. This article puts forward a kind of new evolution algorithm. The algorithm strengthen the local searching ability under the premise of guaranteeing the global optimization ability.

It was important for population scale N, scale factor F, crossover probability CR to control parameters in differential evolutionary algorithm. In effect, F controled scaling degree of the individual difference information. CR directly affected function of variation individuals to the test individual structure. Therefore, the choice of appropriate and dynamical F and CR , ensurance of the population diversity in the early stage of evolution could enhance the global search ability. With evolution processing, the convergence speed and accuracy should be accelerated to make the algorithm converge to the optimal solution as soon as possible.

Therefore, a kind of linear change strategy was adopted in this article.

$$\mathrm{CR}(t+1) = \mathrm{CR}(t) - \frac{\mathrm{CR}(0) - \mathrm{CR}_{\min}}{\mathrm{G}_{\max}} \qquad (5)$$

$$\mathrm{F}(t+1) = \mathrm{F}(t) + \frac{\mathrm{F}_{\max} - \mathrm{F}(0)}{\mathrm{G}_{\max}} \qquad (6)$$

The CR (0) and F (0) respectively were the mutation probability and scale factor of the initial value, CR_{min} was minimum of mutation probability in evolution process. F_{max} was maximum of mutation probability in evolution process.

G_{max} was the maximum of evolutionary generation. With evolutionary generation increasing，crossover probability reduced linearly and scale factor increased linearly. Self-adjusting of parameters ensured population diversity of differential evolutionary algorithm in the early search stage, enhanced the global search ability and had faster convergence speed in late stage, strengthen local search ability.

III. EXPERIMENTAL ANALYSIS AND RESULTS

IRIS data set, Glass data set and Vowel data set were adopted in this experiment. Table I showed number of data sample, attribute number of data sample and number of category in each data set[8-9]. Three data sets were used to respectively test the k-means clustering algorithm, the k-means clustering algorithm based on differential evolution, k-means clustering algorithm based on the improved differential algorithm which was been proposed in this article.

TABLE I. THE PERFORMANCE CLASSIFICATION COMPARISON OF THE EVERY ALGORITHM IN IRIS DATA SET

	number of data samples	number of data attribute	number of category
IRIS data set	150	4	3
Glass data set	214	9	6
Vowel data set	990	10	11

In the experiment of k-means clustering algorithm based on the differential evolution. Population scale N is seted 10 times to dimension d; Scale factor F=0.6, Crossover probability CR=0.5, the interval of initial population is [0, 8], the biggest iterations t = 200. In the experiment of K-means clustering algorithm based on improved differential evolution. N=10d, The value of Scale factor F and Crossover probability CR was Self-adjusting parameters. the biggest iterations t = 200.

For the IRIS, the results of experiment were showed in table II.

TABLE II. THE PERFORMANCE CLASSIFICATION COMPARISON OF THE EVERY ALGORITHM IN IRIS DATA SET THE PERFORMANCE CLASSIFICATION COMPARISON OF THE EVERY ALGORITHM IN GLASS DATA SET

	k-means clustering algorithm	the k-means clustering algorithm based on differential evolution	the improved algorithm in this paper
Minimum class within distance	90.12354	89.1267	86.8956
Maximum class within the distance	184.5768	99.5783	91.4578
Average class within the distance	118.4832	87.8976	88.5661

For the Glass data set, the results of experiment were showed in table III.

TABLE III. THE PERFORMANCE CLASSIFICATION COMPARISON OF THE EVERY ALGORITHM IN GLASS DATA SET

	k-means clustering algorithm	the k-means clustering algorithm based on differential evolution	the improved algorithm in this paper
Minimum class within distance	18.4178	16.5643	15.3456
Maximum class within the distance	42.3409	26.2567	24.8975
Average class within the distance	30.5543	20.9802	18.6743

For the Vowel data set, the results of experiment were showed in table III..

TABLE IV. THE PERFORMANCE CLASSIFICATION COMPARISON OF THE EVERY ALGORITHM IN VOWEL DATA SET

	k-means clustering algorithm	the k-means clustering algorithm based on differential evolution	the improved algorithm in this paper
Minimum class within distance	5973.5678	5412.3768	5103.6712
Maximum class within the distance	9701.0923	6882.3426	5793.2655
Average class within the distance	7983.1566	6054.3475	5217.5747

The results of table I, table II ,table III. showed that k-means clustering algorithm had higher sensitivity of the random selection of initial clustering center, larger fluctuation of cluster result, and the lower stability. Compared with the k-means clustering algorithm, the value class in k-means clustering algorithm based on differential evolution or in the improved algorithm in this article obviously decreased. The results showed that k-means clustering algorithm based on differential evolution and the improved algorithm in this article obviously strengthened the stability and effectiveness of clustering results, significantly improved the quality of clustering results.

Figuare I showed that convergence speed of the improved algorithm in this article was faster than that in k-means clustering algorithm based on differential evolution. It suggested that the improved scheme which were strengthening the local search ability and using dynamic parameter of improved algorithm was effective. Compared with differential evolution algorithm, the improved algorithm had faster convergence speed, higher accuracy than the original k-means clustering algorithm .It accorded with more important property of the clustering method,that is high similarity in class and lower similarity between class.

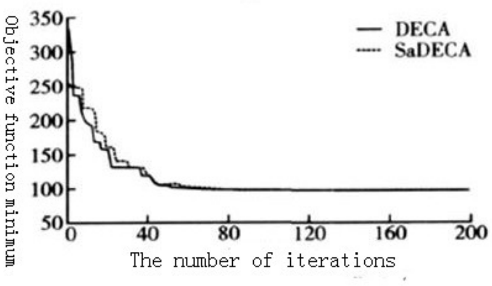

Figure 1. test platform

IV. CONCLUSION

Differential evolutionary algorithm, which was a new evolutionary algorithm, had characteristic such as few control parameters, simple realization, rapid convergence. This paper put forward that the differential evolution algorithm was introduced into the traditional k-means clustering algorithm. The differential evolution algorithm was improved by using dynamic controling parameters. The compared results with improved k-means clustering algorithm showed that optimization ability of the improved differential evolution algorithm to the initial clustering center has been significantly improved. The improved algorithm which had faster global convergence speed, stronger global search ability,thus effectively improve the stability of the clustering results and the clustering quality.

REFERENCES

[1] Qing-hua SU, Zhong-bo HU “Cluster Analysis Based on Differential Evolution Algorithm,” JOURNAL OF WUHAN UNIVERSITY OF TECHNOLOGY, Vol. 32 No. 1 Jan. 2010

[2] Hai-lun WANG, Shi-ming YU, Xiu-lian ZHENG ,“Adaptive Differential Evolution Algorithm and Its Application in Parameter Estimation,”Computer Engineering, Vol.38 No.5 March.2012

[3] Lei XF,Xie KQ,Lin F “An efficient clustering algorithm based on local optimality of K-Means,”Journal of Software,2008,pp.1683-1692

[4] Price K V，Storn R M，Lampinen J A,“Differential evolution a practical approach to global optimization”.New York：pringer，2005.

[5] LI Yinghai, MO Li, ZUO Jian“ Shuffled differential evolution algorithm based on optimal scheduling of cascade hydropower stations,”Computer Engineering and Applications，2012，pp.228-231.

[6] Qi—en YANG, Liang CAI, Yun—an XUE“ A Survey of Differential Evolution Algorithms,” PR&AI, Vol.21 N0.4Aug.2008

[7] Price K, “Differential Evolution vs. the Functions of the 2nd ICEO,” Proc. of the 1997 IEEE International Conference on Evolutionary Computation, Indianapolis, 1997, pp. 153-157.

[8] Hong-Yun MENG, Xiao-Hua ZHANG, San-Yang LIU, “A Differential Evolution Based on Double Populations for Constrained Multi-Objective Optimization Problem,” CHINESE JOURNAL OF COMPUTERS，Vol. 31 No. 2 Feb. 2008

[9] Wang Yue-Xuan, Liu Lian-Chen et al, “Constrained multiob-jective optimization evolutionary algorithm,”Journal of Tsing-hua University (Sci & Tech), 2005, pp. 103-106