An Improved Hybrid Clustering Algorithm Based on Particle Swarm Optimization and K-means

Shuying Liu*, Yanfei Zou
School of computing, Xianyang Normal University, Xianyang, 712000, China

*Corresponding author e-mail: lsynate@163.com

Abstract. The most common algorithm in data mining is cluster analysis. Cluster analysis algorithm has been widely used in many fields, especially in data analysis, market research and pattern recognition. K-means clustering algorithm (KMCA), as one of the common algorithms in clustering analysis, occupies an important position in clustering analysis and has relatively more applications. However, in recent years, with the deepening of the application of clustering analysis algorithm, it is found that there are a series of problems in KMCA, which directly affect the accuracy of data analysis results. On this basis, a bionic optimization algorithm, particle swarm optimization algorithm, is generated, which makes up for the shortcomings of KMCA. The aim of this paper is to establish the optimal algorithm for data analysis by means of the research on the improved hybrid clustering algorithm of particle swarm optimization and k-means. The experimental results show that the algorithm can effectively integrate the advantages of particle swarm optimization algorithm and KMCA, which not only improves the speed of the algorithm, but also guarantees the accuracy of the results.

Keywords: Particle Swarm, K-means Clustering, Hybrid Clustering Algorithm, Clustering Analysis

1. Introduction

K-means algorithm was first proposed by MacQueen, which is mainly applied to the problem solving of cluster analysis, and is most widely used in knowledge discovery and data mining. However, with the deepening of related researches, researchers have found that k-means algorithm has many problems. With the continuous development of the society, the data types of the society are increasing, and the data quantity is also showing an explosive growth. Especially under the background of the development of computer information network, the cost of data storage is low, and the data can be generated automatically, which stimulates the further growth of social data.

At present, people not only hope to realize the storage of data, but also hope to realize the effective conversion of data and realize the informatization and knowledge of data [1]. Therefore, data mining must be carried out, and data mining must be carried out by means of cluster analysis. As one of the classical algorithms of clustering analysis, KMCA is obviously unable to adapt to the practical needs of mass data mining, and particle swarm optimization algorithm is generated on this basis [2-3]. The
emergence of this algorithm effectively makes up for the shortcomings of KMCA. However, after consulting relevant data, it is found that particle swarm optimization alone is still unable to realize deep mining of massive data. Based on this situation, an improved hybrid algorithm of particle swarm optimization and k-means clustering is proposed. In order to realize effective data mining, many researchers have conducted relevant researches on cluster analysis [4-5]. The research of clustering analysis covers a wide range of fields. Currently, the main content of clustering research is the study of several effective clustering algorithms capable of mass data operation, among which the KMCA are more in-depth. Through sorting out the research data, the author finds that the current research focuses on the split study of the two algorithms, respectively studying the advantages and disadvantages of each algorithm and putting forward relevant improvement Suggestions. This research is single and cannot realize the complementary advantages between different algorithms [6-7].

From the above aspects, the research still has a lot of room for improvement. In order to promote the deepening of the study, the effects of clustering analysis algorithm to the data information mining, this paper firstly the concept of particle swarm optimization (PSO) with k-means algorithm and the advantages and disadvantages, on the basis of this puts forward the improved particle swarm and KMCA formula of concrete, and with the help of the related experiment verified the advantages of the proposed algorithm [8-9]. The experimental results show that the algorithm can effectively integrate the advantages of PSO algorithm and KMCA, which not only improves the speed of the algorithm, but also guarantees the accuracy of the results. On the one hand, it promotes the continuous improvement of the theory of clustering analysis algorithm, and on the other hand, it lays a certain theoretical foundation for future related researches [10-11].

2. Method

2.1 Particle Swarm Optimization Algorithm and K-means Algorithm are Separately Described
K-means algorithm is the main algorithm used to solve the clustering problem, which has obvious operational advantages, including fast operation speed and simple operation process [12]. The algorithm is widely used in various fields, especially in the image segmentation, feature extraction, data analysis and pattern recognition is more widely applied in the field of this a few, but we must realize the k-means algorithm is also affected by the initial value and prone to the shortcoming of local optimum, to some extent hindered the normal operation and hinder the further expansion of application scope [13-14].

PSO is a new algorithm with short development time, and its essence is a simplified algorithm. This algorithm is built on the basis of the genetic algorithm, which eliminates the operation process of "selection" and "variation" of the original genetic algorithm, and achieves the determination of the global optimal solution with the help of the group optimal solution [15]. This algorithm is a bionic optimization algorithm based on iteration. The algorithm is widely used in various fields because of its fast operation speed, simple operation steps and high accuracy. Through consulting relevant data, it is found that this algorithm has strong capability in global search and local optimization, making up for the shortcomings of k-means algorithm in operation.

2.2 Improved Hybrid Particle Swarm Optimization and KMCA
Through the references, it can be found that if the advantage of global search of improved PSO is combined with the accurate local solution search ability of k-means algorithm, both the operation speed and the operation accuracy of the algorithm can be considered. Therefore, an improved hybrid algorithm of PSO and k-means clustering is proposed. This algorithm is based on a thorough understanding of the merits and demerits of PSO and k-means algorithm. The adjustment of the improved PSO and k-means hybrid clustering algorithm is mainly reflected in the adjustment of adaptive inertial weight, the reproduction and determination of k-means algorithm operation time and the selection of data coding method. Through the adjustment of these three aspects, the advantages of this algorithm are highlighted to the maximum extent. In this algorithm, inertia weight is related to the
The final result of the algorithm and plays an important role in this algorithm. The specific calculation formula for the adjustment of adaptive inertial weight is as follows:

\[ |x_i - c| = \min |x_i - c_i|, k = 1, 2, ..., m \]  

(1)

In this formula, \( x_i \) represents the position of the \( i \)th particle in a spatial dimension; \( c_j \) stands for the output sample. In pso, the adjustment of inertial weight is not carried out in accordance with the convergence of particle swarm. The calculation formula of the adjustment is as follows:

\[ w(t) = w_{in} + \frac{\sigma(1)(w_{max} - w_{min})}{f_{max}(t) + f_{min}(t) - f_{avg}(t)} \]  

(2)

Where, \( w(t) \) represents the inertial weight, \( t \) represents the specific number of iterations, \( W_{max} \) represents the maximum inertial weight, and \( W_{min} \) represents the minimum inertial weight.

3. Algorithm Experiment Process

Firstly, the concrete operation flow of the improved PSO and k-means hybrid clustering algorithm is experimented to ensure the normal calculation of the algorithm. Specifically divided into the following steps: the first step, population initialization; The primary operation experiment of PSO and k-means hybrid clustering algorithm is to initialize the population. In the process of particle initialization, the cluster must be divided first, and then the operation of each cluster center should be carried out. The result of the operation is the position information of the original particle. Then calculate the fitness of the particle and find the optimal position of the individual to process the velocity of the initial particle randomly. Repeating the above steps several times will produce an initial particle swarm of \( N \). The second step is the redetection of particle position. The fitness of the particle is compared with the best position and the fitness of the population particle. If the fitness result is better, the best position of the particle is updated, including local update and global update. Then PSO algorithm is used to adjust the particle velocity and position. The third step is the selection of particles; With the help of PSO algorithm, carry out random value operation to find the mutant particles, and with the help of k-means hybrid clustering algorithm to calculate the mean value of the mutant particles to find the optimal particles. The fourth step is the optimization of selected particles. In this process, the accurate selection of clustering center must be guaranteed.

Secondly, the improved hybrid PSO algorithm and k-means hybrid clustering algorithm are tested. The improved pso and k-means hybrid clustering algorithm are compared with KMCA and PSO algorithm. If the improved algorithm has obvious advantages over the two algorithms, the algorithm is proved to be reliable.

4. Discuss

4.1 Overview of Experimental Results

The above three clustering algorithms are programmed with the help of Matlab language. The actual data of IRIS is taken as the data sample of this experiment, and the transformation of the optimal solution of the above three experimental algorithms in the iterative operation process is specifically analyzed, so as to achieve the detection and comparison of the clustering performance of the three algorithms. Experimental results show that the improved PSO and k-means hybrid clustering algorithm can ensure the smooth operation, and the clustering performance shows obvious advantages compared with the other two algorithms. The specific experimental data are shown in table 1 and figure 1. The data in the chart are the results of the author's experimental arrangement.
Table 1. results of the algorithms running on the same machine

| Algorithm name | Maximum number of iterations | Actual average number of iterations | Minimum fitness | Maximum fitness | Time/s |
|----------------|-------------------------------|-----------------------------------|----------------|----------------|--------|
| K-means        | 1000                          | 30                                | 10.01          | 81.26          | 2.041  |
| PSO            | 1000                          | 462                               | 25.41          | 66.31          | 6.314  |
| Improved algorithm | 1000                          | 181                               | 13.12          | 13.18          | 54.261 |

*Data came from the in-depth analysis of financial data in the experiment

Figure 1. Comparison of clustering performance of IRIS data by three algorithms

4.2 Analysis of Improved Particle Swarm Optimization and K-mean Hybrid Clustering Algorithm
According to the data in table 1 and figure 1, it can be seen that KMCA has a fast convergence speed and a high probability of entering the local optimal solution. The results obtained by PSO algorithm are very accurate. The improved PSO and k-means hybrid clustering algorithm not only ensures the accuracy of the operation results but also improves the overall speed of data operation. Moreover, the results of each operation are similar, and the difference between them and the optimal solution is relatively small, so it is not easy to enter the local optimal solution, and the global optimal solution is obtained. The performance synthesis of the improved PSO and k-means hybrid clustering algorithm is as follows: data clustering performance, high-dimensional data classification performance, and multi-dimensional data clustering performance. For data clustering improved algorithm in this paper, the performance of random mutation probability increase greatly, makes the global search ability has improved, and because the algorithm only considers the variation particle does not consider other particles, so the algorithm time is less, the initial value to minimize the impact on the results, thus promotes data clustering performance boost; Aiming at the classification performance of high-dimensional data, this algorithm improves the searching ability of the global optimal solution under the premise of a certain number of iterations, so that the convergence speed of the algorithm
remains stable. Based on this, the analysis performance of high data has been greatly improved; In view of the clustering performance of diverse local high-order data, the algorithm in this paper considers the standard variance of data, which reduces the computing difficulties caused by data differences, so as to realize the clustering analysis of various data types and improve the clustering performance of diverse local high-order data.

5. Conclusion
On the basis of explaining the concepts and advantages and disadvantages of PSO and k-means algorithm, this paper puts forward an improved hybrid clustering algorithm of PSO and k-means algorithm, analyzes its specific calculation formulas, and verifies the advantages and reliability of this algorithm with the help of relevant experiments. The experimental results show that the advantages of pso and KMCA are effectively integrated, which not only improves the speed of the algorithm, but also guarantees the accuracy of the results. This algorithm can be applied in the practice of cluster analysis.

References
[1] Mohammad reza Bonyadi, Zbigniew Michalewicz. Particle Swarm Optimization for Single Objective Continuous Space Problems: A Review[J]. Evolutionary Computation, 2017, 25(1):1-54.
[2] Yue-Jiao Gong, Member, IEEE. Genetic Learning Particle Swarm Optimization[J]. IEEE Transactions on Cybernetics, 2017, 46(10):2277-2290.
[3] NIU Wenjing, FENG Zhongkai, CHENG Chuntian. Multi-objective quantum-behaved particle swarm optimization for operation of cascade hydropower stations[J]. Journal of Hydroelectric Engineering, 2017, 36(5):47-57.
[4] P. NAMMALVAR, S. RAMKUMAR. Parameter Improved Particle Swarm Optimization Based Direct-Current Vector Control Strategy for Solar PV System[J]. Advances in Electrical & Computer Engineering, 2018, 18(1):105-112.
[5] Shahira Shaaban, Mohamed Farouk, Hesham Ahmed. Semi-supervised Classification: Cluster and Label Approach using Particle Swarm Optimization [J]. International Journal of Computer Applications, 2017, 160(3):39-44.
[6] Maryam Mousavi. Multi-objective AGV scheduling in an FMS using a hybrid of genetic algorithm and particle swarm optimization[J]. Plos One, 2017, 12(3):517-521.
[7] Haibo Yu, Ying Tan, Jianchao Zeng. Surrogate-assisted Hierarchical Particle Swarm Optimization [J]. Information Sciences, 2018,13(15): 454-455.
[8] Li-Der Chou, Hui-Fan Chen, Fan-Hsun Tseng. DPRA: Dynamic Power-Saving Resource Allocation for Cloud Data Center Using Particle Swarm Optimization[J].IEEE Systems Journal, 2018, 12(2):1554-1565.
[9] Xian-Bo Wang, Zhi-Xin Yang, Xiao-An Yan. Novel Particle Swarm Optimization-Based Variational Mode Decomposition Method for the Fault Diagnosis of Complex Rotating Machinery [J]. IEEE/ASME Transactions on Mechatronics, 2017, 37(99):1-3.
[10] Ghulam Abbas, Jason Gu, Umar Farooq. Solution of an Economic Dispatch Problem through Particle Swarm Optimization: A Detailed Survey–Part I[J]. IEEE Access, 2017, 5(1):15105-15141.
[11] WU Yan-wen, WANG Jie, WANG Fei. Collaborative Filtering Recommendation Model Based on Hybrid Particle Swarm Optimization and Genetic Algorithm[J]. Journal of Chinese Computer Systems, 2017,31(17):116-119.
[12] YANG Yongjian, FAN Xiaoguang, GAN Yi. Coverage optimization of sensor network based on improved particle swarm optimization[J]. Systems Engineering & Electronics, 2017, 39(2):310-315.
[13] Liu, Xiaojia, An, Haizhong, Wang, Lijun. An integrated approach to optimize moving average rules in the EUA futures market based on particle swarm optimization and genetic
algorithms [J]. Applied Energy, 2017, 12(8):185.

[14] Ya-Hui Jia, Wei-Neng Chen, Tianlong Gu. A Dynamic Logistic Dispatching System With Set-Based Particle Swarm Optimization[J]. IEEE Transactions on Systems Man & Cybernetics Systems, 2017, 13(99):1-15.

[15] Songyut Phoemphon, Chakchai So-In, Dusit (Tao) Niyato. A Hybrid Model using Fuzzy Logic and an Extreme Learning Machine with Vector Particle Swarm Optimization for Wireless Sensor Network Localization[J]. Applied Soft Computing, 2018, 23(6):65.