ISSN 2616-5775 Vol. 6, Issue 6: 80-83, DOI: 10.25236/AJCIS.2023.060612

An Improved Ball K-Means Clustering Method Based on SOM

Peilun Han[1], Junxiu An[1,*]

[1]*Chengdu University of Information Technology, Chengdu, China*
[*]*Corresponding author: anjunxiu@cuit.edu.cn*

Abstract: *With the increase of the dimension and quantity of sample data, the calculation cost of K-Means clustering algorithm increases sharply. Therefore, a novel accelerated accurate K-Means clustering algorithm, called "Ball K-Means", has recently been used to reduce the computational cost. Although Ball K-Means reduces the computational cost, both this algorithm and K-Means algorithm lack the global search capability. K-means algorithm may fall into local minima because of its dependence on the initial center. The proper selection of the initial center vector becomes the key to improve the K-means algorithm. Therefore, self-organizing map (SOM) can be used to cluster and determine the clustering range quickly, and then the result can be used as the initial center vector of K-means method. Aiming at the problems that the initial clustering center of Ball K-Means algorithm is randomly selected in the stage of clustering calculation, and the clustering result may fall into a local optimal solution, this article uses SOM network to preliminarily process the data to obtain the initial clustering center of Ball K-Means algorithm, which significantly improves the clustering effect of the algorithm. Taking intrusion detection as an example, the effectiveness and superiority of the algorithm are verified by experiments.*

Keywords: *Data mining, Clustering, Ball K-Means, SOM network*

1. Introduction

At present, clustering has been widely used in customer segmentation, dynamic trend detection, biological data analysis and social network analysis. For various data processing tasks, there are many clustering algorithms, among which K-Means is one of the most classic algorithms [1]. In the early research, the classical K-Means algorithm is the focus of people's attention. The main reason is that the mathematical idea of this algorithm is simple and extensible, and it has linear asymptotic execution time for any variable of the problem [2]. However, the K-Means algorithm also has the following shortcomings: 1) It is difficult to confirm the quantity of K clusters in the data set; 2) It is sensitive to the selection of the initial centroid and lacks the global search ability, which leads to the convergence of the algorithm to the local optimal solution with high probability; 3) With the increase of sample dimension and sample number, the calculation cost increases sharply.

Intrusion detection system consists of software and hardware for intrusion detection [3]. The advantage of K-Means algorithm is that it is simple and fast, and it can deal with large databases effectively, so when it is combined with SOM network, it can not only reduce the quantity of nodes in SOM output layer, but also improve the accuracy of SOM network clustering [4-5]. At this time, a secondary clustering method is formed to meet the requirements of large-scale data detection such as network intrusion detection. Considering the global search ability of the algorithm, this article proposes a new clustering method "SOM-Ball K-Means" combining SOM neural network and Ball K-Means technology. The purpose of SOM-Ball K-Means is to improve the search ability of the algorithm and make the calculation cost of the algorithm as low as possible. In this article, SOM (SOM) network is used to cluster data samples, and the obtained network weight distribution is used as the initial clustering center of Ball K-Means, which effectively enhances the clustering effect and the intrusion detection efficiency of the algorithm.

2. Methodology

2.1 SOM

SOM is a feedforward network for unsupervised competitive learning [6]. Figure 1 shows a two-

ISSN 2616-5775 Vol. 6, Issue 6: 80-83, DOI: 10.25236/AJCIS.2023.060612

dimensional array SOM network model. The output layers are connected laterally, and the adjacent neurons excite each other, while the further neurons inhibit each other, and the further neurons have weak excitation, and finally there is a node or a group of nodes, which reflect the properties of a class of samples.

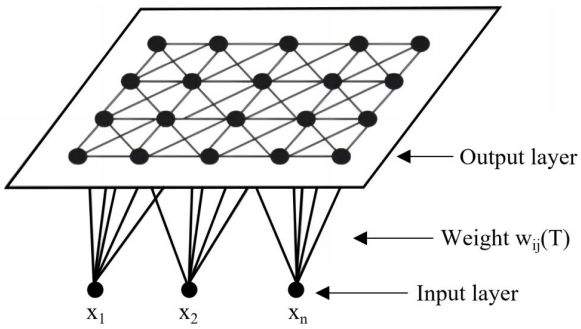

Figure 1: SOM network model

2.2 Ball K-Means clustering algorithm

The "Ball K-Means" algorithm uses the idea of "ball" to describe each cluster, so as to reduce the number of calculations between sample points and cluster centroid. The core idea of this algorithm to reduce the number of calculations is as follows.

1) The definition of cluster centroid and cluster radius in Ball K-Means is as follows:

$$c_i = \frac{1}{|C_i|}\sum_{j=1}^{|C_i|} x_j, r_i = \min_{x \in C_i}(||x - c_i||) \tag{1}$$

In formula (1), the set C_i represents the i cluster, $|C_i|$ represents the sample quantity of the i cluster, c_i represents the centroid of the i cluster C_i, x represents the sample points belonging to the cluster C_i, and r_i represents the radius of the i cluster C_i. The neighbor cluster definition relationship is as follows:

$$\frac{1}{2}||c_i - c_j|| < r_i \tag{2}$$

In formula (2), c_i and c_j represent the centroids of the i and j clusters, respectively. This formula indicates that cluster C_i is a neighbor cluster of cluster C_j (the neighbor clusters are asymmetric). The relationship between clusters and neighboring clusters is shown in Figure 2.

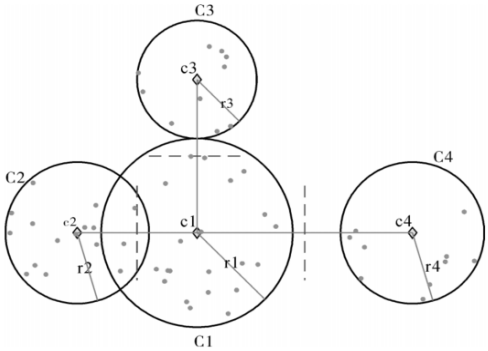

Figure 2: Neighbor cluster relationship of cluster C1

According to the definition of neighbor, the neighbor clusters of cluster C_1 are C_2 and C_3.

2) Each cluster is divided into stable domain and active domain, and the active domain is further divided into multiple ring domains. It is proved mathematically that the sample points in the stable domain will not change, but the samples in the active domain will be redistributed into a neighbor cluster or the original cluster according to the principle of nearest distance, so as to reduce the calculation cost. The specific relationship is shown in Figure 3.

ISSN 2616-5775 Vol. 6, Issue 6: 80-83, DOI: 10.25236/AJCIS.2023.060612

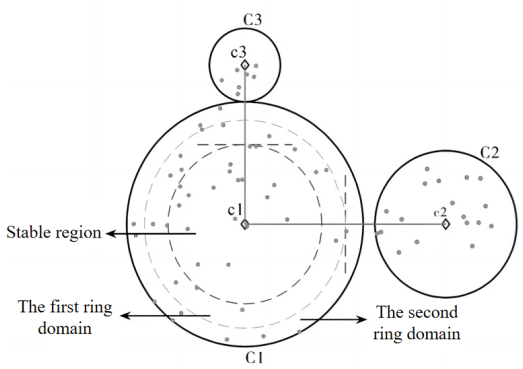

Figure 3: Partition of Cluster C1

According to the definition, the sample points in the first ring domain will be re-divided into clusters C1 and C3, and the sample points in the second ring domain will be re-divided into C1, C2 and C3, while the samples in the stable domain will not be re-divided.

3) Moreover, because when the quantity of clusters is large, it needs high calculation cost to find the neighbor clusters of each cluster. Set C_i and C_j to represent the i and j clusters. If Equation (6) is satisfied, it means that it cannot be a neighbor cluster in the current iteration, so the distance calculation between the centroids of these two clusters can be ignored.

$$dist\left(c_i^{(t-1)}, c_j^{(t-1)}\right) \geq 2r_i^{(t)} + \delta(c_i^{(t)}) + \delta(c_j^{(t)}) \tag{3}$$

In formula (3), $c_i^{(t-1)}$ represents the centroid of the j cluster in the $t-1$ generation, $dist\left(c_i^{(t-1)}, c_j^{(t-1)}\right)$ represents the cluster centroid distance between the i cluster and the j cluster in the $t-1$ generation, and $\delta(c_i^{(t)})$ represents the difference between the $t-1$ generation and the t generation of the i cluster centroid. The purpose is to find out the impossible neighbor relationship.

2.3 Ball K-Means clustering algorithm based on SOM improvement

The classifier based on SOM neural network has good running speed but low working accuracy. Various measures have been taken to improve the classification accuracy, but the test results are still unsatisfactory, and its application is limited by this shortcoming [7]. The classifier based on K-Means algorithm has good classification accuracy, but the running time is too long. Many measures have been taken to meet the real-time requirements of the system, but it still cannot be well satisfied, which limits its application [8]. SOM network can be clustered unsupervised, while K-Means has high accuracy when the number and center of clusters are known. Therefore, combining the two algorithms, SOM network is used for primary clustering, and the number and center of clustering are obtained, and then K-Means algorithm is used for secondary clustering. The advantage of K-Means algorithm is that it is simple and fast, and it can effectively process large databases. Therefore, when it is combined with SOM network, it can not only reduce the quantity of nodes in SOM output layer, but also improve the accuracy of SOM network clustering.

3. Result analysis and discussion

Table 1: Comparison of intrusion detection rates

	Intrusion detection rate		
	Proposed method	Ball K-Means	SOM
DOS	97.187%	75. 145%	75.558%
PROBE	98.194%	95.271%	86.222%
R2L	99.102%	75.242%	66.217%
U2R	98.885%	65.957%	76.522%

In order to verify the performance of the improved Ball K-Means clustering algorithm based on SOM, this article takes intrusion detection as an example and verifies the effectiveness and superiority of the algorithm through experiments. The experiment uses KDDCUP99 standard intrusion detection data set to carry out the experiment. The experimental data includes R2L, DoS, U2R and Probe. In the experiment,

ISSN 2616-5775 Vol. 6, Issue 6: 80-83, DOI: 10.25236/AJCIS.2023.060612

9255 pieces of training data and 17277 pieces of test data are randomly selected.

Table 2: Comparison of intrusion false alarm rates

	Intrusion false alarm rate		
	Proposed method	Ball K-Means	SOM
DOS	1.125%	3.243%	4.274%
PROBE	1.114%	3.838%	4.145%
R2L	1.117%	3.818%	4.227%
U2R	1.147%	3.725%	4.561%

It can be seen from the experimental data that the average detection rate of this algorithm is between ninety-seven percent and 99%, and the average false alarm rate is between 1.12% and 1.15%. Compared with similar algorithms, the detection rate of the new algorithm has been improved in different degrees in the detection stage of several attack types, and the false alarm rate has been reduced to some extent, which can effectively solve the problems existing in the current K-Means algorithm. Whether the initial clustering center vector of K-means method is suitable or not has great influence on its clustering results. And through the data, it can be analyzed that the initial central vector optimized by SOM net can achieve better clustering effect of K-means method, as shown in Table 1 and Table 2.

4. Conclusion

With the rapid growth of IT, the Internet of Everything leads to a huge amount of data information every day, and data mining technology can efficiently discover hidden and valuable knowledge from these data. K-means algorithm is an important data mining algorithm, but the traditional K-means algorithm may fall into local minima because of its dependence on the initial center. In order to verify the performance of the improved Ball K-Means clustering algorithm based on SOM, this paper takes intrusion detection as an example and verifies the effectiveness and superiority of the algorithm through experiments. The results show that, compared with similar algorithms, the detection rate of the new algorithm is improved to some extent in the detection stage of several attack types, and the false positive rate is reduced to some extent. Whether the initial clustering center vector of K-means method is suitable or not has great influence on its clustering results. And through the data, it can be analyzed that the initial center vector optimized by SOM can achieve better clustering effect of K-means method.

References

[1] Brentan B, Meirelles G, Luvizotto E J, et al. Hybrid SOM+ k -Means clustering to improve planning, operation and management in water distribution systems[J]. Environmental Modelling & Software, 2018, 106(AUG.):77-88.

[2] Jiang N, Liu T. An Improved Speech Segmentation and Clustering Algorithm Based on SOM and K –Means [J]. Mathematical Problems in Engineering, 2020, 2020(1):1-19.

[3] Brentan B, Meirelles G, Luvizotto E, et al. Hybrid SOM plus k-Means clustering to improve planning, operation and management in water distribution systems[J]. Environmental modelling & software, 2018(Aug.):106.

[4] Jia Shengsheng, Peng Dunlu. Domain text self-organizing mapping neural network clustering algorithm supported by CNN [J]. Microcomputer System, 2018, 39(6):6.

[5] Wang Shufen, Wang Wei. Multi-dimensional soil data analysis based on self-organizing feature mapping neural network technology [J]. China Agricultural Science and Technology Herald, 2018, 20(4):11.

[6] Zheng Zhong. Clustering analysis of students' physical health data based on self-organizing feature mapping network method [J]. Sichuan Sports Science, 2020, 39(3):4.

[7] Ma Chunlong, Shi Xiaoqing, Xu Weiwei, et al. Correlation analysis of multi-monitoring indicators of contaminated sites based on self-organizing neural network [J]. Hydrogeology Engineering Geology, 2021, 48(3):12.

[8] Zhan Zhongqiang, Yu Jin, Guo Zhi, et al. Study on short-term photovoltaic output prediction based on improved BP neural network with self-organizing mapping [J]. Sichuan Electric Power Technology, 2018, 41(2):6.