Research on Network Big Data Mining Technology Based on Structured Similarity

Lei Wang

School of Computer Science, Wuhan Donghu University, Wuhan City, Hubei Province, 430212, China
E-mail: wanglei3246@wdu.edu.cn

Abstract: At present, most institutions in my country mainly use clustering algorithms for data mining activities, but low accuracy is one of the common problems they face. The article proposes a network big data mining technology based on structured similarity. It is to convert the target data set in the network system into nearest neighbor network (kNN). It then uses the network big data mining technology based on structured similarity. The target data is clustered to obtain a big data result with a slightly poorer objective function but higher clustering accuracy, so as to meet our needs for accurate data statistical results.

1. Introduction

Toffler put forward the concept of "big data" in his book "The Third Wave" published in 1980. Since the beginning of the new century, with the rapid proliferation of information technologies such as the Internet of Things, cloud computing, and mobile Internet, data types and production scales in various fields of society are increasing at an unprecedented rate. "Big data" once again appears in the public as a brand-new concept. And play an important role in various industries. The latest research by the International Data Corporation (IDC) shows that the global data production volume reached 1.8ZB in 2013, and the total amount of global information doubled every two years. The generation of massive data has profoundly changed the way people describe objective things, and also provided new analysis tools for people to explore the objective laws behind surface phenomena, but the premise is that people can obtain accurate big data statistics in a timely manner. At present, people mainly use partition clustering algorithm, hierarchical clustering algorithm, density-based clustering algorithm and many other clustering analysis methods to collect, organize and classify big data. Some scholars are actively discussing clustering algorithms based on k-nearest neighbors (kNN). But it either regard kNN as a tool to improve the clustering speed, or redesign the clustering algorithm from the perspective of dividing Knn. The angle of clustering clusters the kNN graph, but it is not used as an important technical means to change the accuracy of the clustering results. Low accuracy is still one of the common problems faced by many network data mining methods. Based on this, the article transforms the target data set into a kNN network based on the structured similarity on the basis of the existing big data clustering algorithm, and then clusters it to obtain a more accurate clustering result. This is the network big data mining technology based on structured similarity, and it can also be called the network big data clustering algorithm (SSNCA).

2. Design of network big data mining technology based on structured similarity

Big data mining is to find out all the clusters after subdividing the nodes of the same cluster data
object with higher similarity and the nodes of different cluster data objects with lower similarity, so as to reveal the real cluster structure in the network. The article attempts to construct a new data mining technology from the perspective of network clustering. After in-depth research, it is found that when the target data set exhibits distinct cluster structure characteristics, each data object is in the same cluster as the data with similar characteristics. Based on this, we can learn from the neighborhood topology between data objects Relations to design data aggregation or data mining methods. Newman proposed the Q function in 2004, as shown below, that is, using the difference between the expected number of connections and the actual number of connections in the cluster in a random state to show the advantages and disadvantages of the network cluster structure.

\[ Q = \sum_{i=1}^{k} \left[ \frac{m_s}{m} - \left( \frac{d_s}{2m} \right)^2 \right] \]

Among them, \( k \) represents the number of network clusters, \( m \) represents the total number of network connections, \( d_s \) represents the number of node degrees in the network cluster \( s \), and \( m_s \) represents the number of connections in the network cluster \( s \). At present, most of the network big data clustering algorithms mainly use the Q function to evaluate the accuracy of the clustering results. Although there are certain deviations, they can basically be recognized by the industry. Based on this, the network big data mining technology based on structured similarity proposed in the article also uses the Q function as the objective function.

According to the concept of structural similarity between network nodes proposed by Yuruk in 2008, the network is set to \( N=\{V, E\} \), \( \forall v \in V \), and the structure of node \( v \) can be defined as a set of neighborhoods, expressed by the formula for:

\[ \Gamma(v) = \{ w \in V | (v, w) \in E \} \cup \{ v \} \]

On this basis, assuming \( \forall v, m \in V \), if \( (v, m) \in E \), the following formula can be used to express the structural similarity \( \sigma(v, m) \) of network nodes \( v, m \):

\[ \sigma(v, m) = \frac{|\Gamma(v) \cap \Gamma(m)|}{\sqrt{|\Gamma(v)| |\Gamma(m)|}} \]

Through analysis, it is found that in the process of transforming the target data with cluster structure in the network system into kNN network, if the value of \( k \) is appropriate, the two show the same cluster structure, that is, the nodes in the cluster are more closely connected, and The connections between them are relatively sparse, so the clustering results of the original data are obtained through network clustering. Subsequently, the split-type hierarchical data clustering algorithm based on the structured similarity is used for network clustering, that is, the edge with the smallest structured similarity in the network \( N \) is removed every time, and the calculation is completed. A hierarchical and distinct network clustering result can be obtained. The article then uses the Q function mentioned above to find the optimal division result from the cluster structure formed after the operation is completed, that is, select a division that can maximize the Q function value from the final operation result as the final Therefore, the network big data mining technology or network data clustering algorithm based on structured similarity can be described as:

Input \( N/\)target data set into k nearest neighbor network
Output \( C/\)data clustering results
Procedure Network data clustering algorithm based on structured similarity
begin
step1 The structured similarity results of all edges in the network system \( N \);
step2 delete the edge with the smallest value
step3 Recalculate the structured similarity result after deletion;
step4 Return the clustering result \( C \) that maximizes the Q function value in the hierarchical cluster structure
end
It can be seen that this kind of big data mining method uses structured similarity to measure the similarity between adjacent nodes, and uses the Q function as the objective function to find the network cluster structure by using the "split and recalculate" method. In the process of mining the target data set, it is also necessary to first set the parameter k to construct the k nearest neighbor network. The smaller the value of k, the smaller the scale of the calculated clusters. The larger the value of k, the larger the cluster size is. We need to choose an appropriate value of k according to the actual needs of real big data mining.

3. Network big data mining technology experiment based on structured similarity
We use artificially generated networks and benchmark vector data sets to test the technical means separately, so as to observe the feasibility and accuracy of the technical means from different angles.

3.1 Manually generated network test
It is known that the cluster structure in the random network state is RN (C, s, d, zout). Among them, C represents the number of network clusters, s represents the number of nodes in each cluster; d represents the degree of each node, and zout represents the number of connections between each node in the cluster and other nodes in the cluster. The premise for the random network to be clustered correctly is that it can correctly identify the predefined C network clusters and will not further divide it into more sub-clusters. Accordingly, the article uses this method to evaluate the accuracy of big data mining technology. In order to have a clear understanding of the performance of this technology, we compare its test results with the test results of many network big data mining technologies such as GN algorithm, fast Newman algorithm (FN), clump infiltration algorithm (CPM), and community discovery and extraction method (FEC). A comparison was made. The experimental results are shown in Figure 1.

![Figure 1 Comparison of clustering accuracy of five algorithms for random network](image)

Figure 1 Comparison of clustering accuracy of five algorithms for random network

Among them, the y-axis represents the accuracy of the data mining method, and each point on the curve is the average accuracy obtained by mining fifty random networks using different methods. Comparative analysis found that the accuracy of data mining technology based on structured similarity is significantly higher than other methods, and with the increase of zout, although the network cluster structure becomes more and more fuzzy, the clustering difficulty becomes higher and higher. However, the advantages of data mining technology based on structured similarity are becoming more and more obvious, especially when zout is close to the limit value of 8, the accuracy of other algorithms is already very low, but data mining technology based on structured similarity is still able to correctly divide 91.4% of the network nodes.

3.2 Benchmark vector data set test
We select UCI's image, iris, wine three benchmark vector data sets including pre-labeled class identifiers to determine the performance of data mining technology based on structured similarity. Among them, image has 7 types of outdoor image collections and 210 samples randomly selected...
from them, and each sample shows 19 different attributes; iris has 3 types of iris, each of which has 3 different attributes. It consists of 50 samples; wine includes 3 types of wine made from different plants, each of which is composed of 60 samples with 13 different attributes.

The cluster structure of each data set is basically the same as the cluster structure of the k nearest neighbor network. In addition, a large number of experiments have shown that when the value of k is between the minimum size of the data set and 5, the k nearest neighbor network of a specific data set can show the characteristics of the cluster structure.

As shown in Figure 1, the article uses data mining technology based on structured similarity to compare the test results of the above three vector data sets with the data aggregation accuracy of affine propagation and c-Means when the optimal parameters are selected respectively.

| Table 1 Comparison and analysis of the accuracy of the three algorithms for the benchmark vector data set (%) |
|---------------------------------------------------------------|
| **algorithm**               | **image** | **iris** | **wine** |
| Algorithm based on structured similarity | 80.12     | 98.63     | 97.61    |
| c-Means               | 61.74     | 81.34     | 95.30    |
| AP                  | 65.33     | 92.15     | 93.51    |

It can be seen from the above that the accuracy of data mining technology based on structured similarity in the process of aggregating three benchmark vector data sets is significantly higher than that of c-Means and AP.

After comparative analysis, it is found that the data mining technology based on structured similarity produces more and more small-scale clusters as the value of k decreases, and the AP algorithm shows more and more clusters as the parameter reference value increases. Of small-scale clusters. This shows that both algorithms can freely select parameter values, but the algorithm based on structured similarity will keep the number of clusters within a range close to the true number of clusters regardless of the value of k. Therefore, the value of k The sensitivity is relatively small, that is, it is less affected by external human factors. Compared with the AP algorithm, it is more likely to produce accurate data aggregation results. Based on the structured similarity algorithm and AP algorithm, for three benchmark vector data sets, randomly input the clustering and comparison of parameter changes.

Since both the structured similarity-based algorithm and the AP algorithm have multiple input parameters corresponding to a cluster value, the article selects the calculation result of the optimal objective function value as a representative. After comparative analysis, it is found that the objective function value based on the structured similarity algorithm is slightly better than that of the AP algorithm, which once again shows that the network big data mining technology based on the structured similarity shows high accuracy among many network data aggregation algorithms. Degree, can be applied to the practical operation of network data mining.

4. Conclusion

Generally speaking, on the basis of the existing network data mining algorithms, the article proposes a network data mining technology based on structured similarity, which greatly improves the network big data mining from the perspective of network clustering. Accuracy and efficiency. In the next time, on the one hand, we need to deeply analyze the relationship between the target data set and the k nearest neighbor network, and thus propose a reasonable selection method for the parameter k in a specific data mining process; The network conversion method transforms the target data set into a suitable network form, thereby further improving the accuracy of data mining, and analyzing and demonstrating the advantages of this method compared with other methods.
References

[1] Anand Rajaraman, Internet large-scale data mining and distributed processing[M], Beijing: People's Posts and Telecommunications Press, 2019.12

[2] Lin Qianyu, Analysis of Web Data Mining Technology Model[J], Journal of Jilin Normal University of Engineering and Technology, 2019 (2), P78-81.

[3] Yang Ji, Application and Research of Data Mining Technology in Network Public Security Protection[J], Information and Communication, 2019 (4), P23-25.

[4] Xu Baowen, Research on the Application of Data Mining Technology in Web Prefetching[J], Journal of Computer, 2019 (2), P34-36.

[5] Duan Qi, A network topology inference algorithm based on data aggregation[J], Computer Simulation, 2018 (4), P90-91.