Improved K-median Clustering Algorithm Based On Wavelet

Ming-xia Xie
Zhengzhou Institute of Surveying and Mapping
xmx0424@yahoo.cn
China

Clustering is used widely in pattern recognition and data mining, it is a method to self-organize data in compute. There are many clustering algorithms existed. Which one of algorithms is chosen is due to data type, the purpose and application of clustering. On the whole, we can classify the clustering algorithms such as partitioning method, hierarchical method, density-based method, grid-based method and model-based method. During the process of partitioning method, we must know the initial clusters first of all, while it is good or bad will affects the clustering precision directly. The character of hierarchical method and grid-based method is quick processing speed and that of model-based method is robust. Density-based method owns the function that it can acquire clusters from spherical and unbalanced datasets.

All of us are known that the quality of clustering is closely related to the number of clusters of datasets and data quantity. On the one hand, the number of clusters is too large to explain and analyze the result of clustering. Meanwhile, if the number is too small, it will lead to make the information lose and mislead the final decision. On the other hand, the data quantity is too large to execute clustering fast and efficiently.

Therefore, how to process the original data and gain the initial clusters is the key to the research in clustering. To solve this problem, improved k-median clustering algorithm based on wavelet through in-depth study of the partitioning method, hierarchical method and the wavelet and integrating the three methods is put forward that has considered how to select the initial clusters and update the clustering centers.

Firstly, estimate the data quantity. If it is too large that we can carry on the wavelet transformation to the original data in order to compress it and then execute hierarchical clustering to the compressed data, otherwise we can make hierarchical clustering to the original data directly. Secondly, gain the initial clusters and their centers of k-median clustering algorithm according to the result of hierarchical clustering. Finally, do the k-median clustering in terms of the initial clusters and their centers and the result are the final clusters of the original data. During the process of the k-median clustering, the computing and updating of clustering centers is due to the maximum and minimum of elements in each cluster.

While if the value of corresponding element in clustering center is between the minimum and maximum, then don’t change the value of that element, or if it is bigger, then replace it with the maximum and if it is smaller, then replace it with the minimum. We don’t choose the means or medoids of elements in each cluster to computing and updating the clustering centers, because the means may be nonsensical in reality and it is very sensitive to the noise and outlier. Meanwhile, choosing the medoids will make the process of computing and updating the clustering centers become very time-consuming.

In order to assess the effectiveness of proposed techniques, we carried out experiment of image segmentation. Theoretical analysis and experimental results testify that this algorithm outperforms traditional Partitioning methods and hierarchical methods in both precision and consuming time.