The Application of Big Data Mining Prediction Based on Improved K-Means Algorithm

Yuchen Qiao, Yunlu Li*, Xiaotian Lv
School of Automation
Wuhan University of Technology
Wuhan, China
luluyunli080533@whut.edu.cn

***Abstract*—In order to solve the problem of low efficiency of K-Means algorithm in processing the data mining prediction problem of big data with more attributes, an annual income prediction method of residents based on improved K-Means algorithm is proposed. The improved K-Means algorithm combines the principal component analysis method with the traditional K-Means algorithm. After reducing the dimensionality of various data attributes, the data are classified with K-Means algorithm. The research makes use of 1994 U.S. census database and conducts a contrastive analysis of the two algorithms. The results show that the prediction accuracy has been significantly improved by 13.3313%, from 53.1016% to 66.4329%. It is clear the improved algorithm can effectively improve the accuracy of clustering and annual income prediction.**

Keywords—K-Means algorithm; Various attributes; Data mining; Principal component analysis; Annual income forecast

I. INTRODUCTION

This is the era of big data, and how to get the law hidden behind a large amount of data through data mining has become a research hotspot. With electronic data as the basis, big data record, adjust, compare and process all the factors involved in social production, including people, events, things and environment, so as to improve production efficiency [1]. In order to make it better applied to reality and improve people's living standards, we need to carry out specific analysis of big data and propose relevant solutions to the problems arising in the process to dig out the rules hidden behind big data.

At present, the mainstream data mining algorithms include C4.5 decision tree algorithm, CART decision tree algorithm, KNN (K Nearest Neighbors) algorithm, Naive Bayes algorithm, PageRank algorithm, K-Means algorithm, etc. The K-Means algorithm is the most classical and widely used clustering method.

K-Means algorithm is a classical unsupervised learning clustering algorithm. Researchers have proposed many improved algorithms based on the traditional K-Means algorithm, including Canopy-K-Means algorithm [2], MinMax K-Means algorithm [3], KMOR algorithm [4], etc. The clustering effect is significantly improved compared with the traditional K-Means algorithm. Ref. [5] proposed a phased clustering method to improve the K-Means algorithm, but its time complexity is high and changes unsupervised clustering into supervised clustering. Ref. [6] optimizes the algorithm by optimizing the initial center point. Ref. [7] combines Canopy [8] algorithm with K-Means algorithm to solve the selection problem of initial center, but this method needs the initial parameters of the manually-selected Canopy algorithm. However, there has been no solution provided to improve the effect of the traditional K-Means algorithm when it is used to deal with the big data with many attributes.

In order to accurately predict the annual income level of residents, first, the K-Means algorithm is improve and the big data with multi-attribute are processed. All the attributes which might influence the residents’ annual income are reduced in dimensionality with principal component analysis (PCA), and then the K-Means algorithm is used to analyze the clustering and establish residents' annual income prediction model. Second, based on the improved K-Means algorithm, we analyze the relationship between the residents' annual income and their education level, marital status, occupation and other attributes, on the basis of which the prediction model for the residents' annual income forecast is established. The improved K-Means algorithm optimizes the method of repeated clustering used by the traditional K-Means algorithm when processing the data with various attributes. Hence, it effectively improves the efficiency and accuracy in predicting the annual income of residents.

II. EXPERIMENTAL DATA AND THEORY

A. Experimental data

In this research, the 1994 U.S. census database is taken as the research object, and samples with incomplete information in the original database are removed. Finally, the relationship between the 14 characteristic variables including age, education level, marital status, occupation, race and national origin and the annual income of the citizen is studied.

In this paper, the annual income is divided by 50K dollars, more than 50K (>50K) and less than 50K (≤50K) dollars, and it is taken as the final output variable. 14 characteristic variables of the data set are classified by both the traditional K-Means algorithm and the improved K-Means algorithm; and the classification results are compared to draw the final conclusion.

This paper is supported by National Students’ Platform for Innovation and Entrepreneurship Training Program in 2019(ID:201911012)

978-1-7281-3936-4/19/$31.00 ©2019 IEEE

B. Related theoretical research

1) Traditional K-Means algorithm

The traditional K-Means algorithm is the unsupervised learning clustering algorithm, that is, clustering analysis is carried out on the unmarked sample set. The algorithm idea is to divide the sample set into K clusters according to the distance between samples, so that the sample distance within the cluster is as small as possible and the distance between the clusters is as large as possible.

Suppose the clusters are divided into C1, C2,…, Ck, the algorithm's goal is to minimize the square error E [9], namely:

$$E=\sum_{i=1}^{k}\sum_{x\in Ci}\left\|x-\mu i\right\|_2^2 \tag{1}$$

Among them, μi is the mean vector of Ci：

$$\mu i=\frac{1}{|Ci|}\sum_{x\in Ci}x \tag{2}$$

Generally, the smaller the E value is, the higher the similarity between samples in the cluster will be.

The flow of traditional K-Means algorithm is as follows:

a) K samples are randomly selected from the data set D({X1、X2、…、Xm}) as the initial k centroid vectors({μ1、μ2、…、μk}). The number of clusters is k, the maximum number of iterations is N, and the output is clustering and dividing cluster C({Y1、Y2、…、Yk}).

b) For n=1,2,... ,N, initialize Ct=$\varnothing$*， t=1,2...k.*

c) For i=1,2... M, calculate the distance between sample xi and each centroid vector μj(1≦j≦k), the formula is as follows:

$$dij=\|xi-\mu j\|_2^2 \tag{3}$$

d) Determine the cluster marker λi of xi with the smallest distance.

e) At this time, the sample is updated and divided into corresponding clusters: Cλi= Cλi ∪{Xi}.

f) For j=1， 2， …， k， For all sample points in Cj, use formula (2) to recalculate the new center of mass.

g) If all k centroid vectors do not change, then the output clustering is divided into cluster C=={Y1、Y2、…、Yk}.

2) Principal component analysis algorithm

Principal component analysis uses the idea of dimensionality reduction, which is a method to recombine many original indicators with certain correlation into a new set of a few unrelated comprehensive indicators [10]. That is:

$$Fp=a_{1i}*Z_{Xp}+a_{2i}*Z_{Xp}+......+a_{pi}*Z_{Xp}\,,$$

$$i=1,...,p \tag{4}$$

The a1i,…,api(i=1,…,p)of the eigenvalues of the covariance matrix Σ, eigenvectors corresponding Z_{X1} 、Z_{X2} …、Z_{Xp} is the standardized value of the original variable. We plan to apply the principal component analysis method to the dimensionality reduction of the 14 input variables affecting residents' annual income, so as to reduce the running complexity of the subsequent K-Means algorithm. The calculation steps of PCA are as follows:

a) Use SPSS software to standardize the raw data.

b) Determine the correlation between indicators.

c) Determine the number of principal components m.

d) The new expression of principal component Fi is calculated, as shown in formula (4).

III. Construction and Application of the Annual Income Prediction Model of Residents Based on the Improved K-Means Algorithm

It is stated clearly in the second part of this paper that the traditional K-Means algorithm has been thoroughly studied. The algorithm is easy to implement and widely adopted. The clustering effect of it is good for the large quantity of data sets, but there still exists some disadvantages. When it is used to process the data sets with many attributes, the efficiency is low because it cannot distinguish the useful attributes from the useless ones and has to classify all the attributes of the samples. Moreover, the interference of those useless attributes leads to a low accuracy in the final clustering.

A. Results of principal component analysis

The original data included 14 input variables that are likely to influence the income of residents, including age, workclass, Fnwgt, education, education-num, marital status, occupation, relationship, race, sex, capital-gain, capital-loss, hours-per-week and native country. We applied python software to conduct principal component analysis of the 14 variables. The eigenvalues and variance contribution rates of the correlation coefficient matrix R in the sample data set are shown in table 1.

In practical applications, in order to make full use of the original information, the corresponding components whose cumulative contribution rate is above 85% are generally

Table 1 the eigenvalues and variance contribution rates of R

The principal components	The eigenvalue	Variance contribution rate	Cumulative contribution rate
G1	11154539444	0.9946	0.9945
G2	59329346.63	0.0052	0.9997
G3	164893.1852	$1.47e^{-05}$	0.9998
G4	2.5989182.3609	$1.63e^{-08}$	0.9999
G5	138.2463	$1.23e^{-08}$	0.9999

selected as the retention components [11]. As can be seen from table 1, the contribution rate of the first five principal components has reached 99.99%, indicating that the first five principal components basically contain all the information of the features. Therefore, we use the first five principal components and the original data to form a new sample set, and we name the new variables as G1, G2, G3, G4 and G5.

B. Model establishment and application of K-Means algorithm based on principal component analysis

In view of the shortcomings of the traditional K-Means algorithm, this paper proposes an improved algorithm. Firstly, principal component analysis is carried out on 14 influencing factors of residents' annual income to reduce the dimensions. Then, the K-Means algorithm in clustering method is used to find the relationship between influencing factors and residents' annual income.

The process of establishing the model of K-Means decision tree algorithm based on principal component analysis is as follows:

a) Digitize the sample data of adult data set.

b) Conduct dimensionality reduction with the main analytic hierarchy process for the digitized data.

c) Take the 30162 sets of data after dimension reduction as the training set, the principal component variable as the input variable, the annual income of residents as the target variable, and use Python software to conduct K-Means algorithm modeling.

d) The remaining 15060 groups of data in the adult dataset are used as the test set to verify the accuracy of the prediction model.

IV. Comparison between the improved K-Means algorithm prediction model and the K-Means algorithm prediction model

We use the traditional K-Means algorithm to build the prediction model of residents' annual income level, and compare it with the model established above, so as to get the conclusion intuitively. Experiments are carried out under the same conditions to obtain the final results, and the running results of the two models are compared, as shown in table 2.

As can be seen from table 2, when the improved K-Means algorithm is used to predict the annual income level of residents, the accuracy of the model is increased by 13.3313%, from 53.1016% to 66.4329%, while it only takes 2.8579 more seconds to finish the operation. The model accuracy of the improved algorithm increases greatly while the running time remains basically unchanged. Therefore, the model established on the improved K-Means algorithm has a better clustering effect on big data with more attributes.

Table 2.Comparison of predicted results

Dataset name	The dimension	Algorithm	The prediction accuracy of residents' annual income level (%)	Run time (s)
adult	14	Traditional K-Means algorithm	53.1016	14.4924
adult	14	Improved K-Means algorithm	66.4329	17.3503

V. Conclusion

In the era of big data, how to find hidden rules from massive data and solve problems has always been a hot research direction. Among the many algorithms of data mining, K-Means algorithm is widely used. However, when solving the problem of data set with more attributes, there still exist the problems of low operation efficiency and low clustering accuracy. In such problems, there are many factors affecting the final clustering accuracy, and the relationship between each influencing factor and the target variable is complex, and there are useless attribute variables, which greatly hinder the establishment of the prediction model.

To solve the above problems, this research uses the PCA to reduce the dimension of the 14 input variables and screen out the useless variables. Therefore, the complexity of the data is greatly reduced. Then the K-Means algorithm is used to analyze the 45,222 sets of resident information from the 1994 U.S. census database. Finally, the prediction model for the residents' income level is set up. Through experimental comparison, the prediction accuracy of the prediction model established by the improved algorithm has increased significantly from 53.1016% to 66.4329%, and the running time remained basically unchanged, which indicates that the model has certain improvement.

There is still room for improvement even though the effect of the newly-established model has already been proved good. It is advisable to combine PCA and K-Means algorithm to analyze the big data with many attributes. In this way, data mining algorithms are more extensively adopted and it provides a new thought in solving the problem of big data by means of data mining algorithms.

Acknowledgment

This paper is supported by National Students’ Platform for Innovation and Entrepreneurship Training Program in 2019(ID:201911012)

References

[1] M. Zhou, Q. Tian, “Application status and prospect of big data,” J. Information and computers (theory), 2019(03):39-41, in press.

[2] D. Mao, “Improved algorithm for Canopy-Kmeans based on MapReduce,” J. Computer Engineering and Application, 2012, 48(27):22-26+68, in press.

[3] G. Tzortzis, A. Likas, G. Tzortzis, “The MinMax Kmeans clustering algorithm,” J. Pattern Recognition, 2014,47(7):2505-2516, in press.

[4] G. Gan, KP. Ng, “K-Means clustering with outlier removal,” J. Pattern Recognition Letters, 2017(90):8-14, in press.

[5] J. Zhang, X. Wu, J. Jiang, “Semi-supervised phase clustering of complex distributed data,” J. Journal of Frontiers of Computer Science ＆ Technology, 2016,10(7):1003-1009, in press.

[6] Z. Wang, G. Liu, E. Chen, “A K-Means algorithm based on optimized initial center points,” J. Pattern Recogniton and Artificial Intelligence, 2009, 22(02):299-304, in press.

[7] R. Qiu, “Canopy for K-Menas in multi-core,” J. Microcomputer Information, 2012(09):486-487, in press.

[8] R.M. Esteves, C. Rong, “Using mahout for clustering wikipedia's latest articles: a comparison between K-Means and Fuzzy C-means in the cloud”. P. Cloud Computing Technology and Science (CloudCom), 2011 IEEE Third International Conference, 2011.

[9] J. Zuo, Z. Chen, “Anomaly detection algorithm based on improved K-Means clustering,” J. Computer science, 2016,43(08):258-261, in press.

[10] S. Zhang, C. Zhang, Q. Meng, “Comprehensive evaluation of regional technological innovation capacity in China based on principal component analysis,” J. Economic journal, 2013(Z2):90-91, in press.

[11] Y. Tian, W. Ma, Y. Liu, Z. Xiao, G. Chen, “Study on prediction of well flooding effect based on PCA-FNN,” J. Journal of Xi 'an Petroleum University (natural science edition), 29(03):83-86+10-11, in press.