IJIRR

International Journal of Information Research and Review
Vol. 04, Issue, 10, pp.4580-4584, October, 2017

RESEARCH ARTICLE

DESIGN AND IMPLEMENTATION OF THE K-REGRESSION CLUSTERING ALGORITHM

[1,*]Hu Shaolin, [2]Feng Bingqing and [1]Zhang Caixia

[1]Foshan University, Foshan, 528000, China
[2]Key Laboratory of spacecraft fault diagnosis and maintenance, Xi'an, 710043, China

ARTICLE INFO

Article History:

Received 28th July, 2017
Received in revised form
07th August, 2017
Accepted 17th September, 2017
Published online 30th October, 2017

Keywords:

Clustering,
Linear Regression,
K-regression Clustering.

ABSTRACT

Clustering is one of the important technical approaches in data mining, which is widely used in knowledge discovery and machine learning. Some existing clustering methods mainly focus on either the distribution density of elements or the similarity of the morphology formed by elements in the data set, while lacks of the analysis of the intrinsic characteristics of the data. On the basis of the reasonable kernel of K-means clustering algorithm, a new clustering method is proposed in this paper and named as the K-regression clustering to realize the regression clustering. This new clustering method is based on the correlation of different components within the data set. Using this method, the data set is divided into several categories, and the data from the same subset obey a identical linear regression, while there are significant differences between any two different linear regression equations from different data subsets. Simulation results verify the rationality and validity of the K-regression clustering method.

INTRODUCTION

The so-called clustering is essentially a process of dividing a collection of samples, data, or objects into several different subsets with a particular or multiple similar features.In other words, the purpose of clustering is to divide the objects into severaldifferent subsets, and the objects within each subset have the same characteristics as possible, while there are significant differences between any two different subsets.As one of the unsupervised learning methods, clustering analysis has been widely concerned in machine learning theory and applications. And it has formed a series of practical methods, such as the distance-basedclustering (He *et al.,* 2007; Pelleg and Moore, 2000), fuzzy information-basedclustering (Ayed *et al.,* 2014), density-based clustering (Ester *et al.,* 1996; Nagpal *et al.,* 2011), genetic rules-basedclustering (Safaris *et al.,* 2002), neural networks-based clustering (Xu *et al.,* 2015), clustering based on maximum likelihood functions as well as on the maximum posteriori estimation, clustering based on statistical models (Karypis *et al.,* 1999; Chen Lifei *et al.,* 2010; Hai, 2016), grid-based clustering, random sampling-based clustering, and hierarchical clustering of large-scale data from complex structures system (Zhao *et al.,* 2014). Clustering analysis has been widely used in many different fields such as medicine, statistics, biology, management science and so on, which provides an effective technical means to simplify the technical difficulty of object analysis, distinguishing the difference of different objects and mining the knowledge of data sets. Comprehensive review of the existing clustering methods tell us that almost all of the existed methods on clustering data mainly stay in the external correlation between different objects in datacollection, such as the distance between different objects, density distribution, similar shape etc. Obviously, these methods are far from enough to deal with difficult clustering problems in the actual engineering field as well as in daily life, it will encounter a large number of objects whose similarity or differences come from the internal correlation of the structure. For example, the originalregression relationship came from the inter-correlationbetween the heightofchildren and parents. Supposing the acquisition of a large number of sample set is

$$S=\{(h_i^s, h_i^f, h_i^m) \mid i=1,2,3,\cdots,N\}$$

where (h_i^s, h_i^f, h_i^m) isthe ternary array composed by child's height、father's height and mother's height of the i^{th}family.The existing biostatistical studies had revealed that the child's height can be described by the duality linear regression equation of the father's height and the mother's height

$$h_i^s = a + b \cdot h_i^f + c \cdot h_i^m + \varepsilon_i$$

****Corresponding author: HuShaolin,***
Foshan University, Foshan, 528000, China.

However, for different regions, different times or different populations, the coefficients of the regression equations that characterize the height change are usually not the same. Obviously, for the sample data set from different regions, different times, or different populations, whether to use the K-means method, density-based clustering or to use other methods to cluster data set above is defective. A more reasonable clustering approach should be based on the inherent relationship between the heights of the ternary array (h_i^s, h_i^f, h_i^m) to divide the data set. Similar problems will also occur in a wide range of areas such as medical treatment, chemical process, management engineering, disaster analysis and so on. In view of this problem, based on the reasonable kernel ofK-means clustering algorithm, this paper proposes a K-regression clustering method. This method inspects whether the sample data in the subsets is subject to the same regression equation and whether there is a significant difference between the regression equations of different subsets to achieve the regression clustering of the sample data.

Design of K-regression Clustering Algorithm

It is known that $S=\{(X_k \in R^m, y_k \in R) \mid k=1,\cdots,n\}$ is a sample data set with several different structural relationships and $X_k = (x_{k,1},\cdots,x_{k,m})$ is the independent variable, and there may be a linear correlation between the dependent variable **y**with the independent variable. In particular, the sample set S may be consisted of K different parts

$$S=\bigcup_{j=1}^{K} S_j \qquad \cdots\cdots (1)$$

Suppose that the dependent variables **y**and independent variables $X \in R^m$ in the sample subset S_j satisfy the linear regression model

$$y^j = a_{1,j}x_1^j + \cdots + a_{m,j}x_m^j + \varepsilon^j \qquad \cdots\cdots (2)$$

where, the superscript j indicates that the sample is from the set S_j, ε^j represents the zero mean random error component in the j^{th}-regression model.

The basic idea of K-regression clustering is to discriminate each data in sample set S into the most suitablesubset S_j. For each specific subset $S_j = \{(X_{j_i}, y_{j_i}) \mid i=1,\cdots,k_j\}$, the model (2) can be embodied as

$$y_{j_i} = a_{1,j_i}x_1^{j_i} + \cdots + a_{m,j_i}x_m^{j_i} + \varepsilon^j \qquad \cdots\cdots (3)$$

It is not difficult to verify that the least squares unbiased estimation $\hat{\vec{\theta}}_j$ of the model coefficients shown in Eq. (3)and theresidual sum of squaresRsscan be given by Eq. (4)

$$\begin{cases} \hat{a}_j = (X_j^T X_j)^{-1} X_j^T Y_j \\ Rss_j = Y_j^T (I - X_j (X_j^T X_j)^{-1} X_j^T) Y_j \end{cases} \qquad \cdots\cdots (4)$$

where, $\vec{a}_j = \begin{pmatrix} \hat{a}_{1,j_1} \\ \vdots \\ \hat{a}_{m,j_{kj}} \end{pmatrix}$, $X_j = \begin{pmatrix} x_1^{j_1} & \cdots & x_m^{j_1} \\ \vdots & \ddots & \vdots \\ x_1^{j_{kj}} & \cdots & x_m^{j_{kj}} \end{pmatrix}$ and $Y_j = \begin{pmatrix} y_{j_1} \\ \vdots \\ y_{j_{kj}} \end{pmatrix}$。

Based on these series assumptions, the essence of K-regression is to find such a partition approach forthe sample set $S = S_1 \oplus \cdots \oplus S_K$,which is two minimize theresidual sum of squares Rss_j :

$$S_1 \oplus \cdots \oplus S_K = \arg\min_{S_1 \oplus \cdots \oplus S_K} \{\sum_{j=1}^{K} Y_j^T (I - X_j (X_j^T X_j)^{-1} X_j^T) Y_j\} \qquad \cdots\cdots (5)$$

It's difficult to solvethe Eq. (5) because the number of sample points in each subset of the division $S = S_1 \oplus \cdots \oplus S_K$ is unknown and uncertain.And,it is also unknown and uncertain that which subset will be designated for each point in the data set S. Based on the reasonable kernel of K-means clustering algorithm, this paper establishes a method which is simple and easy to implement.

K-regression clustering algorithm

Setting the initial estimation value of K fitting coefficients $\{\hat{\vec{\theta}}_1^{(0)},\cdots,\hat{\vec{\theta}}_K^{(0)}\}$,the absolute residual will be calculatedforeach point $(X_k, y_k) \in S$ away to the K-order regression hyper-planes or the regression lines. Taking the minimum as the label fordata point $(X_k, y_k) \in S$ andcompleting the classification markings for all points in set S, the fitting coefficients are recalculated according to Eq.(3), and the sum of squares of the fitting residuals is calculated according to Eq. (4). Finally, calculating the ratio R of the current squared sum of the fitting residuals to the last fitted residuals RSS,the iteration is calculated until R is close to 1. The K-regression algorithm stated above may changes the mark of the sample pointbased on the absolute value of the fitting residual. Specifically, by each iterative calculation, the absolute value of the difference between the variance measurement of each sample point and the fitting value of the K regression equations is calculated, and the sample pointsare classified into different subsets with the smallest residual absolute value. According to the idea of K-means clustering algorithm,the K-regression clustering calculation is carried out by setting the K fitted initial coefficientsvector $\{\hat{\vec{\theta}}_1^{(0)},\cdots,\hat{\vec{\theta}}_K^{(0)}\}$. The specific calculation process is shown in Fig 1. In the process of actual calculation, the initial estimation value of the coefficient vector can be set withthe scatter plot of the data in combination with the data processing experience. It is also possible to simply divide the sample data into K equivalents, and give the estimated value of the fit coefficient for each set using Eq. (4).

Simulation Results Analysis

Using the linear regression model $y = a + bx + \varepsilon \quad \varepsilon \sim N(0,\sigma^2)$,we set four set of different parameters $(a_1, b_1, \sigma_1) = (24,1.3,1)$, $(a_2, b_2, \sigma_2) = (-6,-1.1,1)$, $(a_3, b_3, \sigma_3) = (12,1.2,1.2)$ and $(a_4, b_4, \sigma_4) = (-12,0.9,1.2)$, and generate four groups data, while each group has 41 simulation data, as shown in Fig. 2.

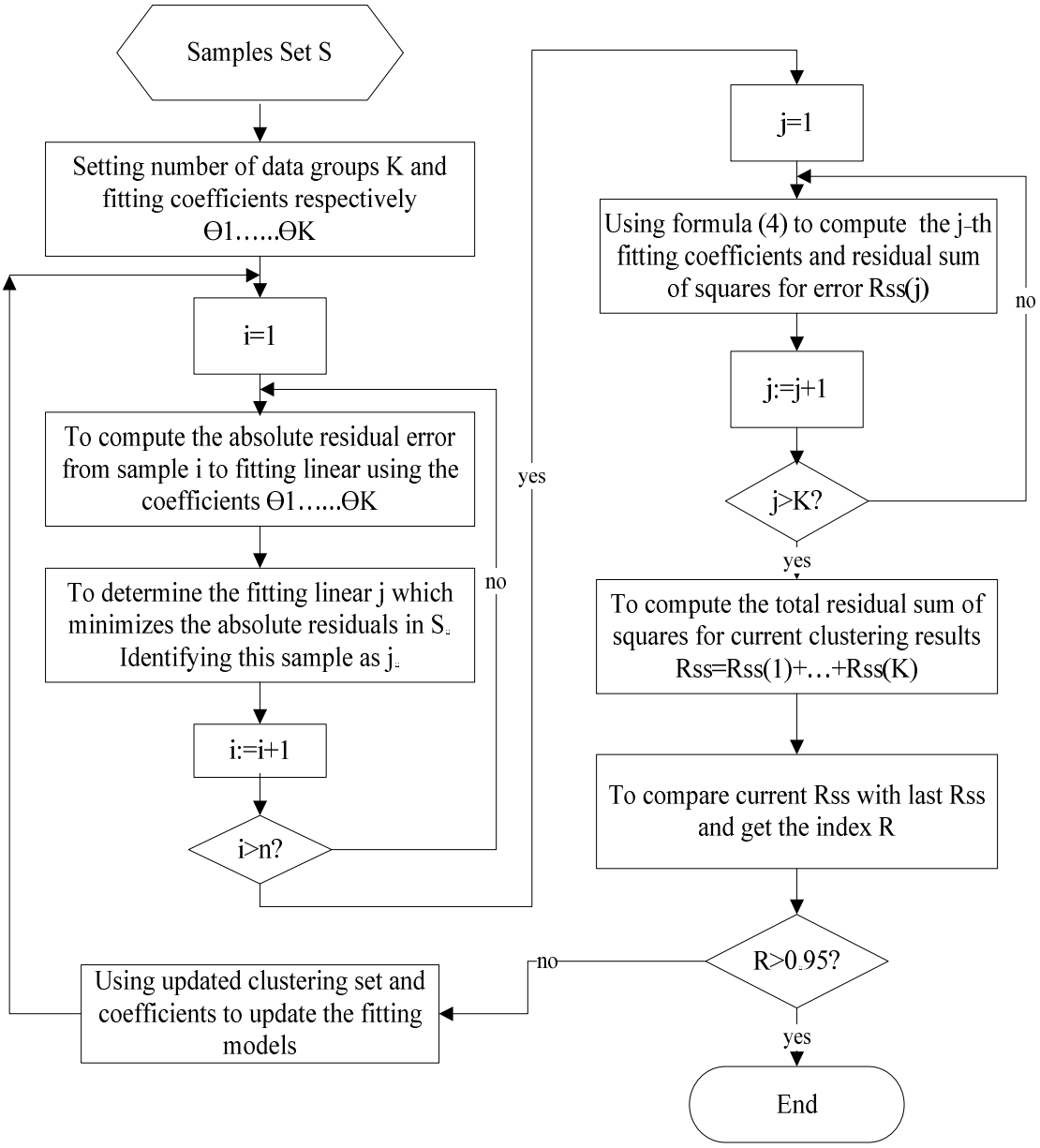

Fig. 1. The process of K-regression clustering

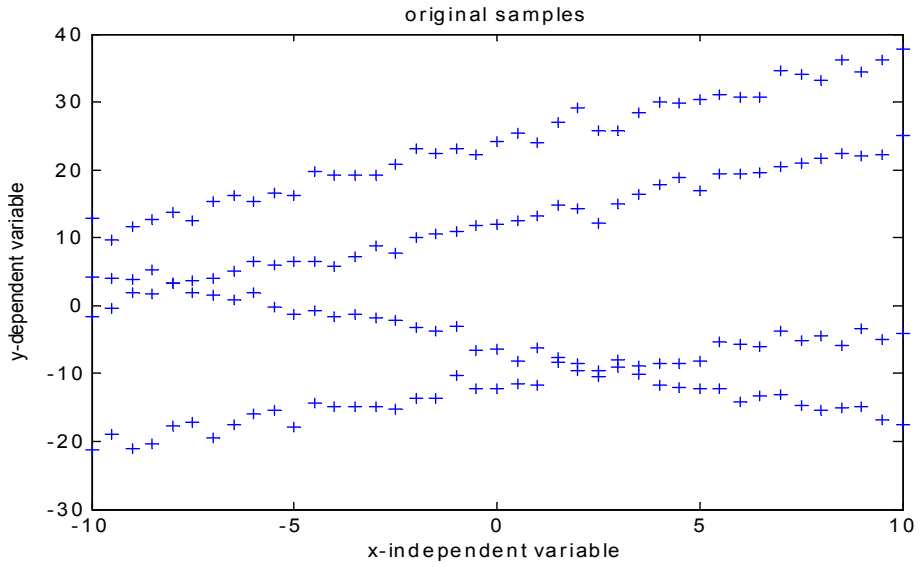

Fig. 2. The scatter plot of the raw data

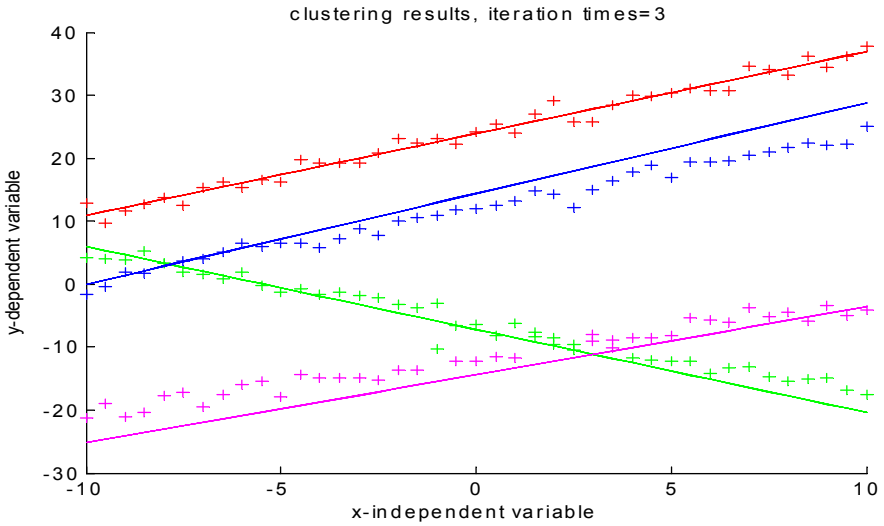

Fig. 3. Regression ClusteringLinear Graph

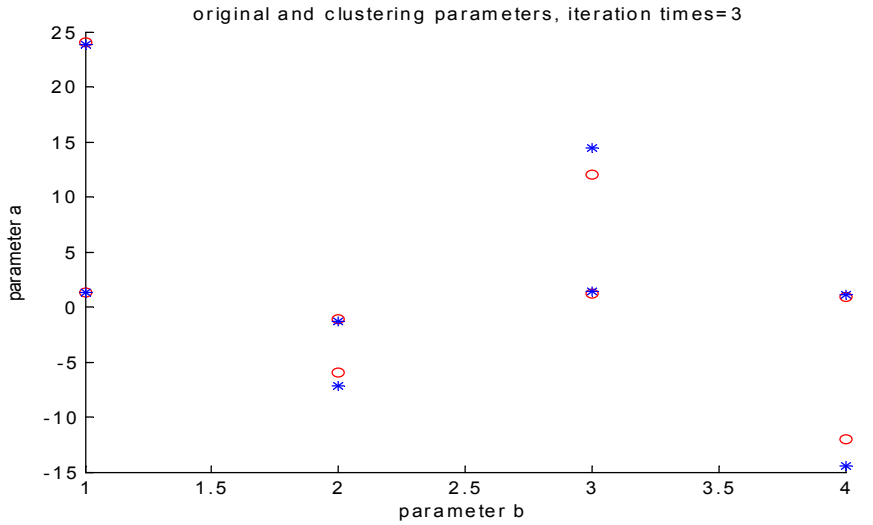

Fig. 4. Results of the Model Coefficients Clustering

As shown in Fig.3, it's the clustering result by using the k-regression clustering algorithm. It can be obviouslyseen from Fig. 3 that the sample data of the four regression models with different coefficients are accurately clustered. In the process of regression clustering, the initial value of the coefficients in theiterative processistheK-regression clustering with 30% offsetof the true coefficients of the model. The model coefficients after clustering better overcome the influence of the initial deviation and regress to the vicinity of the real value, as shown in Fig.4. Comprehensivelyanalyzing Fig.2-4, it can be confirmed that the K-regression clustering algorithm established in this paper can effectively achieve regression clusteringfrom different sample sets with linear dependencies among multivariable variables.Thealgorithm can find out the same or similar regression relationship, and effectively separate the different regression relationship.

Conclusion

Clustering is an important part of in-depth analysis of data sets and accurate discovery intrinsic characteristics of data, which has attracted much attention in the field of data mining and knowledge discovery. Based on the reasonable kernel of K-means clustering algorithm, this paper proposesa new clustering method(K-regression clustering) to find the correlation of different components within the data,which implements data clustering based on regression relation. After clustering, the data in the same class obey the same linear regression, while there are significant differences between different classes in the linear regression equations. In this paper, we use Monte Carlo simulation to verify the rationality and validity of K-regression clustering method. It can be expected that the K-regression clustering method will have practical value in medical, chemical, management engineering and disaster analysis, and so on.

Acknowledgment

The research work in this paper is supported by the National Natural Science Foundation of China (No. 91646108, No. 61473222).

REFERENCES

Ayed, A. B., Halima, M. B., Alimi, M. 2014. Survey on clustering methods: Towards fuzzy clustering for Big Data [6th International Conference of Soft Computing and Pattern Recognition (SoCPaR). IEEE, 331-336.

Chen Lifei, Jiang Qingshan, Wang Shengrui, 2010. Model-based Method for Projective Clustering. *IEEE Transactions onKnowledge and Data Engineering*, (99). http://doi.ieee computersociety.org/10.1109/TKDE.

Ester, M., Kriegel, H.P., Sander, J. et a1. 1996. A Density-based Algorithm for Discovering Clnsters in Large Spatial Databases with Noise.Portland : *Proceedings the 2rd ACM SIGKDD*, pp:226-231.

Hai M. 2016. Survey of Clustering Algorithms for Big Data. *Computer Science,* 43(6A):380-383.

He, L., Wu, L., Cai, Y. 2007. Survey of clustering algorithms in data mining, *Application Research of Computers,* 24(1):10-13.

Karypis, G., Hart, E.H., Kumar, V. C. 1999. A Hierarchical Clustering Algorithm Using Dynamic Modeling. *IEEE Computer, 32*(8):68-75.

Nagpal, P. B., Mann, P. A. 2011. Survey of Density Based Clustering Algorithms. *International Journal of Computer Science and its Applications,* 1 (1): 313-317.

Pelleg, D., Moore, A. 2000. X-means: Extending K-means with efficient Estimation of the Number of the Clusters. Proceedings of the 17th ICML.

Safaris, I., Zalzala, A. M., Trinder, P. W. 2002. A Genetic Rule, based Data Clustering Toolkit. Honolulu: Congress on Evolutionary Computation (CEC).

Xu, R.,Wunsch, D. 2005. Survey of clustering algorithms, Neural Networks. *Journal IEEE Transactions,* 16(3): 645-678.

Zhao, Y., Chen, Y., Liang, Z., et al. 2014. Big Data Processing with Probabilistic Latent Sernantic Analysis on MapReduce, *International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery,* 162-166.

****