International Journal of Advanced Biotechnology and Research (IJBR)
ISSN 0976-2612, Online ISSN 2278–599X,
Vol-7, Issue-2, 2016, pp651-653
http://www.bipublication.com

Review Article

Clustering of Mixed data: A GKMM approach

Abha Sharma[*1] and R. S. Thakur[2]

Maulana Azad National Institute of Technology,
Bhopal, India
[1*]abha_sharma31@yahoo.com
[2]ramthakur2000@yahoo.com

[Received-25/05/2016, Accepted-12/06/2016, Published-30/06/2016]

ABSTRACT-
Clustering is important problem in data mining techniques. *k*-Means algorithm is one of the most capable and easy to employ clustering algorithm but having some difficulties i.e. applicable to numeric data, sensitive to the presence of noise and outliers and have initialisation issues. This paper proposed GKMM algorithm to cluster mixed data where the pre-processed data will be used for the clustering to overcome the limitation of local solution and handle only numeric data issues. This work is based on the concept of utilisation of numeric data based genetics clustering algorithm for mixed data and can be an easier alternative to reduce cost and helpful in optimizing performance. Moreover, decrease obvious sensitivity to the initial guess of the cluster centres.

Keywords: Clustering; Mixed data; *k*-Means algorithm; Genetic Algorithm

I. INTRODUCTION

Cluster analysis is a radiant data mining task which divides a set of data objects into groups say clusters [1]. The motivation behind clustering a set of data is to find its inherent structure and expose that structure as a set of groups [2]. The data objects within each group should exhibit a large degree of similarity while the similarity among different clusters should be minimum. The need for fast and efficient data analysis has driven the research community to continuously develop and improve data clustering methods. Most previous clustering algorithms focus on numerical data whose inherent geometric properties can be exploited naturally to define distance functions between objects. However, many fields from statistics to psychology deal with mixed data, unlike numerical data, mixed data cannot be naturally ordered [3]. An example of categorical attribute is *shapes* having categories *circle, square, rectangle,* etc. Therefore, those clustering algorithms handling numerical data cannot be used to cluster mixed categorical data [4]. Recently, the problem of clustering mixed categorical and numeric data has received much interest [5]. A well-liked robust nature motivated A Genetic Algorithm based clustering technique [6], was basically developed for pure numeric data. In this concept based paper this GA based technique can be used for high dimensional mixed data objects after conversion from mixed to pure numeric data.

The main aim of this paper is to implement the genetic algorithm based *k*-Means clustering for mixed data, which integrates the genetic algorithm and the *k*-Means algorithm in order to find the globally optimal solution of the optimization problem. The content of the paper is as follows. In Section II, the pre-processing is briefly discussed, Section III shows the GA based clustering algorithm, Section IV Flow of improved Genetic algorithm based *k*-Means algorithm for mixed data (GKMM). Finally, a

few concluding observations are given in Section V.

II. Pre-processing

In 2010 Ming-Yi Shih et. al. [7] proposed a method which define the similarity among items of categorical attributes based on the idea of co-occurrence. All categorical values will be converted to numeric according to the similarity to make all attributes contain only numeric value. So that clustering algorithm for numeric data can also be apply inherent structure for further clustering. This algorithm is simple, straightforward and based on the similarity measure among numeric and categorical attributes [8].

III. Genetic Based Clustering

Genetic *k*-means [5] is proposed to handle numerical datasets shown in fig. 1. This technique takes one step *k*-Means algorithm as crossover operator and roulette wheel selection method to transfer number of copies to fit chromosomes or string, solve the problem to initialise cluster centres.

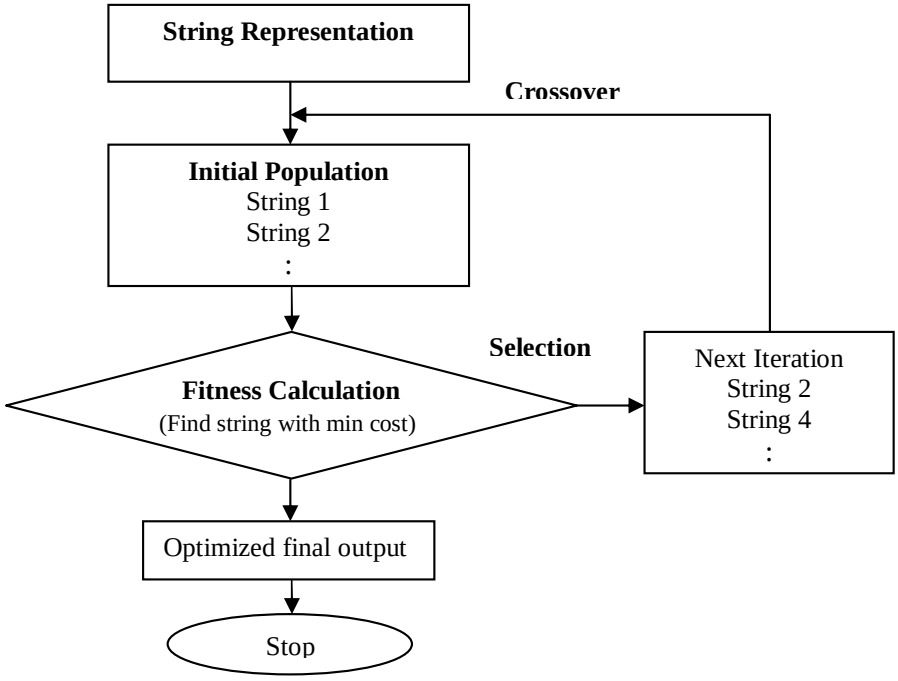

Fig. 1 Flow of Genetic algorithm for clustering

IV. Proposed Algorithm

This proposed algorithm shown in fig. 2 is suitable for mixed data say no limitation for all attributes to be numeric or categorical, where the population is initialized randomly and is evolved on population; the population in the next generation is achieved by applying all the genetic operators. The iteration has been continue until a terminating condition is reached. The genetic operators that are used in GKMM are the distance based fitness function, *K*-means crossover operator and a selection method used is roulette wheel.

String representation

Every string is a series of real numbers like *M*C* where *M* is number of dimension and *K* is number of clusters assumed.

Population initialization

The *K* cluster centres are initialized to P_i randomly chosen points from the population *P*.

Fitness computation

The fitness function is defined as *f*=1/D, thus maximization of the fitness function leads to minimization of D, here D is Euclidian distance measure.

Crossover

Crossover used in this paper is single point crossover with a fixed crossover probability, which swap information between two parent strings to generate two new strings.

Selection

In this paper Roulette wheel selection method is used that implements relative selection strategy,

where number of copies is assigned to a chromosome according to fitness criteria.

Solution of the Empty cluster problem

To the best of our knowledge in almost all the clustering algorithm the problem of creation of empty clusters is the well-known problem in clustering. And the problem becomes big if the optimization techniques are used [4], this paper try to remove the empty cluster issue using following algorithm:

Proposed clustering algorithm for Mixed data

Input :*Mixed dataset(D_1) having M attributes,*

1. *Partition the dataset, $M=m_1$(categorical attributes)+m_2 (Numeric attributes)*
2. *Apply z-score normalization of all the numeric attributes*
3. *Partition the dataset into categorical and numeric attributes*
4. *Find the categorical attribute having highest categories (b_1)*
5. *Find the numerical attribute having lowest standard deviation (b_2)*
6. *Calculate the categorical variables of b_1 with the mean of corresponding values of b_2*
7. *Calculate all the (M_1-b_1) categorical attributes using eq. 1*
8. *Numeric data (D_2)*
9. *Apply genetic k-Means clustering in D_2*
10. *Stop*

Output: *Final Clusters*

Fig. 1 Proposed clustering algorithm for mixed data

V. CONCLUSION

This method has achieved success in initialisation of cluster centres better than traditional methods along with categorical data issues. The method will be highly computational intensive when analysing large datasets. Passing the pre-processed data in genetic k-means remove the limitation of one time initialisation and pre-processing solve the categorical data issue.

ACKNOWLEDGEMENT

The work is supported by research grant from MPCST Bhopal, India under grant no. 1080/CST/R&D/2012 dated 30-06-2012.

REFERENCES

1. MacQueen, J. B., Some Methods for Classification and Analysis of Multivariate Observations, Proce. 5th Berkeley Symposium Mathematical Statistics and Probability, 1, 281-297 (1965)
2. Anderberg M.R., Cluster Analysis for Applications, Academic Press, Inc., New York, 1973.
3. Jain A.K., Murty M.N. , Flynn P.J., Data clustering: a review, ACM ComputingSurveys31,264–323 (1999).
4. Abha Sharma and R. S. Thakur, “A Variant of Genetic Algorithm Based Categorical Data Clustering for Compact Clusters and an Experimental Study on Soybean Data for Local and Global Optimal Solutions” International Journal of Advanced Computer Science and Applications(IJACSA), 7(2), 2016.
5. Maulik U., Bandyopadhyay S., Genetic algorithm-based clustering technique, Pattern Recognition, 33, 1455-1465 (2000)
6. Shih M. Y., Jheng J. W..Lai L.F., A Two-Step Method for Clustering Mixed Categorical and Numeric Data, Tamkang Journal of Science and Engineering, 13, 11-19 (2010)
7. K. Krishna and M. Narasimha Murty, Genetic K-Means Algorithm, IEEE Transactions on systems, Mans and Cybernetics Part B: Cybernetics, 29, 3, 433-439 (1999)
8. Z. Huang. Clustering Large Data Sets with Mixed Numeric and Categorical Values, Knowledge discovery and data mining: techniques and applications. World Scientific.1-14 (1997).