IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 5, May 2016

Fuzzy Partitioning Based Clustering Approach

Nipjyoti Sarma[1], Adarsh Pradhan[2], Utpal Barman[3]

Assistant Professor, Department of CSE, GIMT, Guwahati, India[1,2,3]

Abstract: Clustering mixed dataset is a very common approach in day to day life. Several algorithms exists for clustering heterogeneous datasets with accuracy. But most of the algorithms are inefficient and unable to cluster only one type of records. Therefore we have designed an algorithm which is efficient as like k means and can also handle mix type and only one type of data in dataset. Our algorithm has been tested on standard dataset to see the performance.

Keywords: Heterogeneous, Fuzzy, Cluster, Centroid, Partition.

I. INTRODUCTION

In data mining k-mean algorithm is the most popular partition based clustering algorithm. It is popular due to its simplicity of understanding and linear algorithmic complexity measure. But it has the serious limitation of clustering numerical only data. Therefore several researchers tried to improve the k mean algorithm to cluster not only numerical but also categorical dataset. Most of the conventional clustering algorithms have been designed and applied to datasets containing only single attribute type (either numerical or categorical). Recently, approaches to clustering for mixed attribute type datasets have emerged, but they are mainly based on transforming attributes to simply utilize conventional algorithms. The problem of such approaches is the possibility of distorted results due to the loss of information because significant portion of attribute values can be removed in the transforming process. This results in a lower accuracy clustering. To address this problem, we propose a clustering framework for mixed attribute type datasets without transforming attributes. The rest of the paper organized as follows. Section 2 gives the literature survey. The proposed algorithm is presented in Section 3 .Section 4 gives the experimental results. At last, section 5 contains the conclusion and future scope.

II. LITERATURE SURVEY

Limin CHEN , Jing YANG and Jianpei ZHANG,in [1] proposed a novel generic clustering approach of maping a large and mixed type datasets that sample number is very large and attribute space tends to be stationary on a quasi-hyper-image to get a small quasi-hyper-image pixel .This approach makes the clustering process very speedy since the quasihyper- image pixel dataset is small . The mapping process only takes O(m) order of complexity and therefore it does not effect the clustering badly. The large dataset can be clustered rapidly on the base of dataset mapping. Author ah stated that the clustering precision only has direct relation with the particle size of quasi-hyper-image. If the particle size is small, the clustering precision is high, but the storage and calculation complexity will increase. If the particle size is large, the storage is small and calculation speed is quick, but the clustering precision is low. The particle size of quasi-hyper-image is selected according to speed and precision. Authors have done experiments on some dataset which are synthetic . The authors also has taken experiments on a huge dataset of 500000 patients record. The result of the experiments showed that this method CQHIM is effective and efficient, and is quite appropriate for solving large mixed type dataset clustering.

Ming-Yi Shih, Jar-Wen Jheng and Lien-Fu Lai in [2] proposed a two step method for clustering mixed Categorical and Numerical Data. They presented a new idea to convert items in categorical attributes into numeric value based on co-occurrence theory. This method explores the relationships among items to define the similarity between pairs of objects. A reasonable numeric values can be given to categorical items according to the relationship among items. Then a two-step k-means clustering method with adding features is proposed. K-means's shortcomings can be improved by applying this proposed method. They have tested the dataset on some benchmark dataset like credit approval dataset, stat log heart disease dataset, contraceptive method choice dataset.

Jongwoo Lim, Jongeun Jun, Seon Ho Kim and Dennis McLeod in [3] proposed a clustering framework that supports clustering of datasets with mixed attribute type (numerical, categorical), while minimizing information loss during clustering. They proposed clustering framework consists of three main steps. In Step 1, used an entropy based similarity measure with only categorical attributes to extract candidate clusters. In Step 2, the extracted candidate cluster numbers K from Step 1 to cluster the dataset using only numerical attributes. In Step 3, a weighting scheme is applied using the degree of balance in number of objects in the clusters. After determining the weights, the final clustering is processed for the mixed attribute type dataset using the extract candidate cluster numbers from Step 1 and the weights. By doing experiments they showed that the candidate cluster number extracted from only categorical attributes can be used as the candidate cluster number for mixed attribute type dataset in the given dataset and the proposed weighting scheme based on the degree of balance of clustering can improve the accuracy of clustering.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 5, May 2016

Chian Hsu and Yan-Ping Huang in [4] proposes a modified adaptive resonance theory network (M-ART) , which can handle mixed dataset directly. They showed the disadvantages of adaptive resonance theory neural network1 can handle only binary data where as adaptive resonance theory neural network 2 can handle general numerical data. If categorical data presents then it should be first converted to numerical and then can perform clustering. Due to this representation clustering is not well and hence the concept of modified adaptive resonance theory network comes with the conceptual hierarchy tree. The experimental results on synthetic data sets showed that the proposed approach can better reveal the similarity structure among data, particularly when categorical attributes are involved and have different degrees of similarity, in which the traditional clustering approaches do not perform well. The experimental results on the real dataset have better performances than other algorithms. The modified algorithm is experimented on standard adult dataset from UCI machine learning .The authors compared the accuracy of their proposed modified algorithm with k prototype (K mean and k mode) and ART2 algorithm and got better result in mixed type of clustering.

III. PROPOSED WORK

A. Clustering Problem

We assume a dataset of objects O of D dimension is defined by a set of attributes {A1, A2, A3, A4 ,.., A_D},We check the similarity between a point R and the cluster Centroid in such a manner so that the similarity between the object and cluster increases and the intra cluster similarity decreases.

B. Similarity Metrics of Nominal data

The similarity metrics between an attribute of an object and the same attribute of a particular cluster is considered as a fuzzy membership value which is basically the fraction between 0 and 1. It is –

$\mu Ai(d) = \frac{freq(d)}{|Ci|}$

Where, freq(d)= $\sum_{l=1}^{|Ci|} 1$ such that $p_{li} = d$

Here $0 \leq freq(d) \leq |Ci|$ and hence, $0 \leq \mu Ai(d) \leq 1$

C. Fuzzy based Centroid vector of a cluster

For mixed data type we define a Centroid of a cluster Ci having mixed type attribute as based on above definition of fuzzy membership function of it and is as given below[2]:

fuzzyVi = {Vi1,Vi2,Vi3.....Vim}

Where m is the total number of attributes and Vij is defined as:

$$FuzzyV_{ij} = \begin{cases} Average_i(A_j), if A_j \text{ is numerical attribute} \\ \tilde{A}_i, if\ A_j \text{ is categorical attribute} \end{cases}$$

Where for numerical attribute Aj is the average of the j^{th} attribute of cluster C_i and it is-

Average (Aj)= $\frac{1}{|Ci|}\sum_{k=1}^{|Ci|} p_{kj}$

And for categorical attribute the Aj, $\tilde{A}_j$ is calculated by using the equation as given in the above section.

Based on the definition above, now the similarity measurement of mixed type object and cluster is given by separating the categorical part and the numerical part.

D. The Proposed Algorithm

The Fuzzy Partitioning based Centroid Vector algorithm for clustering mixed type data is :

Step no1: Input the dataset with all values in between 0 and 1to avoid unusual values.
Step no 2: Select the same no of points as equal to the cluster no.
Step no 3: Calculate the similarity based on the equations given in section above,
Step no 4: Insert all other dataset instance to the cluster with less proximity.
Step no 5: Update each fuzzy set Centroid vector by calculating the average and dissimilarity function.
Step no 6: Repeat the above two steps after every iteration and see any improvement occurs in the cluster fuzzy based Centroid vector or not?
Step no 7:If changes then repeat it again or
Step no 8: If it do not changes, then output the final cluster.
Step no 9: run this for 100 times
Step no 10: Final output clusters

IV. EXPERIMENTAL WORK

The algorithm is executed 100 times to see the performance and compared with the normal k means algorithms. In the experiments the accuracy is estimated as follows

$$ACC = \frac{\sum_{i=1}^{N} \delta(c_i, l\ d(xi,p)}{N}$$

Where C_i is the cluster and the d(xi,pi) is the distance between maximum number of data objects of a cluster i belonging to the same original classes in the test data (correct answer) and n is the number of data objects in the databases. For comparative studies, the result of the algorithm is compared with k-means algorithms.

The performance of our algorithm on mixed datasets is investigated. The statistics of the selected dataset is shown in table 1.Table 2 shThe dataset credit approval is mixed type wherews the accuracy and table 3 shows the exection time details of both the algorithm of Credit Approval Dataset.In table 4 the statistics of the iris dataset is given.It is a numeric type Table 5 and table 6 shows the accuracy and exection time details of Iris dataset. We shall see the result in both the datasets ,Dataset obtained from UCI MLR [12].

The statistics of the dataset is given in table 1 clustering results are summarized in table 2. Fig.1 and fig.2 shows

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 5, May 2016

the accuracy comparison and execution time of the credit Approval dataset where as Fig.3 and Fig.4 shows the comparison in accuracy and execution time for Iris dataset .Fig.5 and Fig.6 shows the after cluster scatter plot which indicates quality of clustering in the dataset. The dataset is obtained from UCI machine learning repository[10]

TABLE1: Statistics of The Mixed Datasets

Dataset	Instance	Attribute(d_{cat}+d_{nu})	Class	Class Probabilities
Credit Approval	653	9+6	2	54.67%,45.33%

Table 2. : Clustering accuracy of K-FCV algorithm on mixed datasets in comparison with k-means

Dataset	k-means	our approach
Credit Approval	.5513±.0016	**.580**7±.8452

Table 3. Average convergence time of our algorithm on mixed datasets in comparison k-means

Dataset	k-means	our approach
Credit Approval	**0.1323**	0.2777

Table 4. Statistics of the numerical datasets

Dataset	Instance	Attribute (d_{nu})	Class	Class Probabilities
Iris	150	4	3	33.33%, 33.33%, 33.33%

Table 5.Clustering accuracy of our algorithm on numerical datasets

Dataset	K-Means	our approach
Iris	88%	88.92% ±20%

Table 6.Average convergence time comparison with K-means

Dataset	K- Means	our approach
Iris	0.0028	**.0010**

Comparision Between K-Mean and our proposed Approach Based On Credit Approval Dataset. The following graph shows the accuracy variations in existing algorithm like K-Means and the proposed methods on the Credit Approval dataset. It is a mixed type dataset. The graph shows that the improvement in accuracy in the proposed method in the sample dataset with random Centroid initialization and higher than that of K-Means algorithm.

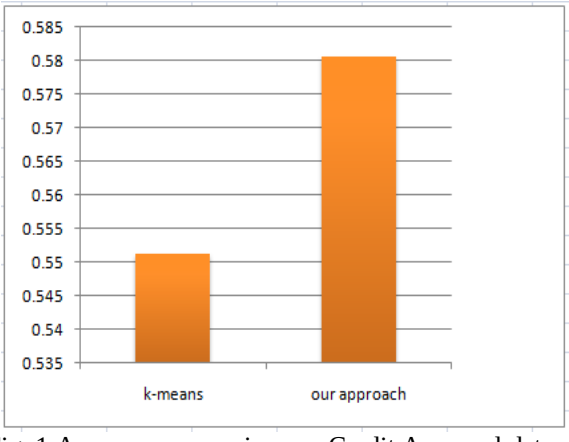

Fig. 1.Accuracy comparison on Credit Approval dataset

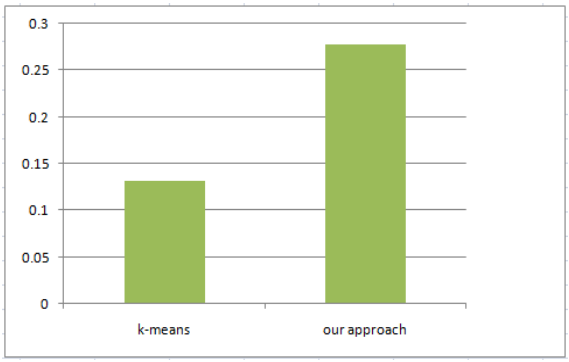

Fig.2. Time variation on Credit Approval dataset

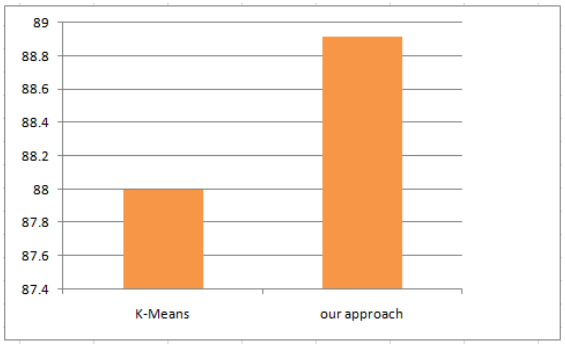

Fig 3: Accuracy comparison on iris dataset

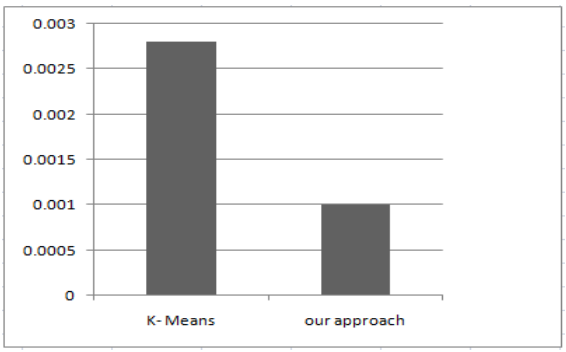

Fig 4: Average convergence time comparison with K-means

Fig.5 shows the plotting of credit approval dataset having two groups in two different colours indicating two clusters.

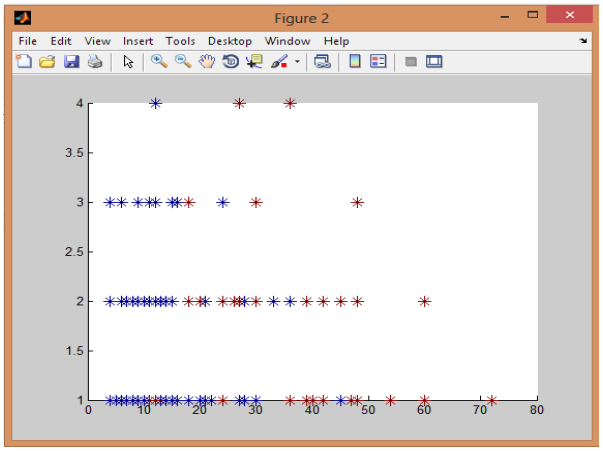

Fig. 5.Scatter plot after clustering for Credit Approval dataset

DOI 10.17148/IJARCCE.2016.5554

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 5, May 2016

Fig.6 shows the plotting of Iris dataset having three groups in three different colours indicating two clusters.

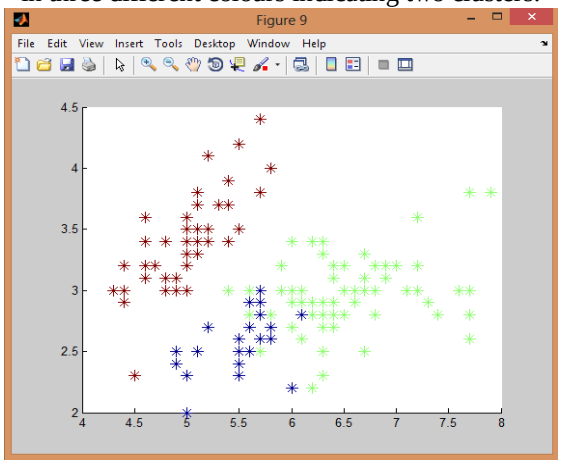

Fig.6. Scatter plot after clustering on iris dataset

V. CONCLUSION

The k-mean algorithm is the most popular partition based clustering algorithm in data mining. Due to its simplicity of understanding and linear algorithmic complexity measure it is very popular in the research community. The limitation of this k means clustering algorithm is that it can only handle numerical data. Therefore several researchers tried to improve the k mean algorithm to cluster not only numerical but also categorical dataset. In this paper we have presented an approach using partition based fuzzy Centroid calculation which we can call as a modified version of the traditional k-mean algorithm. This algorithm is able to cluster objects having mixed type attributes i.e. numerical and nominal. We have done the experiment on credit approval dataset and irisdataset which are taken from UCI machine learing repository[12]The result have been discussed. Due to noise in input dataset the algorithm could not perform well . In future we try to tackle this problem. This algorithm can be applied in e commerce site as well as in fraud detection.

ACKNOWLEDGMENT

Our sincere thanks goes to **K.V Kanimozhi**, HOD dept of CSE, GIMT, Guwahati-17.

REFERENCES

[1]. Limin CHEN , Jing YANG and Jianpei ZHANG, in "An Efficient Clustering Method for Large Mixed Type Dataset" in Journal of Computational Information Systems 8: 22 (2012) 9553–9560,

[2]. Ming-Yi Shih, Jar-Wen Jheng and Lien-Fu Lai in "A Two Step Method for Clustering Mixed Categorical and Numerical data" in Tamkng Journal of Science and Engineering, Vol. 13, No. 1, pp. 11-19(2010)

[3]. Jongwoo Lim, Jongeun Jun, Seon Ho Kim and Dennis McLeod in "A Framework for Clustering Mixed Attribute Type Datasets" in Proceedings of the fourth International Conference on Emerging Databases (EBD),2102

[4]. Chian Hsu and Yan-Ping Huang in "Incremental clustering of mixed data based on distance hierarchy" in Elsevier/Expert Systems with Applications 35, 1177–1185,2008, The authors in this paper

[5]. P. Andritsos, P. Tsaparas, R.J. Miller and K.C. Sevcik in "LIMBO: scalable clustering of categorical data" in Proceedings of the 9th International Conference on Extending Database Technology, pp. 123–146.2004,

[6]. D. Barbara, J. Couto and Y. Li in "COOLCAT: an entropy-based algorithm for categorical clustering" in Proceedings of the 11th ACM Conference on Information and Knowledge Management pp. 582–589,2002. The COOLCAT algorithm,

[7]. Yiu-ming Cheung and Hong Jia in "A Unified Metric for Categorical and Numerical Attributes in Data Clustering" A report from Department of Computer Science, Hong Kong Baptist University, in 2002

[8]. M.J. Zaki and M. Peters, in "CLICK: mining subspace clusters in categorical data via k partite maximal cliques" in Proceedings of the twenty-first International Conference on Data Engineering, pp. 355–356, 2005

[9]. S. Guha, R. Rastogi and K. Shim in " ROCK: a robust clustering algorithm for categorical attributes" in Information Systems 25 (5) 345–366 (2005),IEEE,Digital Library,id754967.

[10]. D.K. Roy and L.K Sharma in "Genetic k-mean clustering algorithm for mixed numeric and categorical dataset" in International Journal of Artificial Intelligence and Applications(IJAIA),Vol-1, No-2,April 2010.

[11]. M. V. Jagannatha Reddy1 and B. Kavitha2 in "Clustering the Mixed Numerical and Categorical Dataset using Similarity Weight and Filter Method" in International Journal of Database Theory and Application Vol. 5, No. 1, March, 2012,

[12]. UCI machine learning Repository- http://archive.ics.uci.edu/ml/.

BIOGRAPHIES

Mr. Nipjyoti Sarma working as an Assistant Professor at GIMT, Guwahati has published many research papers in the areas of Data Mining, Image Processing. His area of interest is Data Mining.

Mr. Adarsh Pradhan working as an Assistant Professor at GIMT, Guwahati has published many research papers in the areas of Machine learning, Image Processing. His area of interest is Machine learning.

Mr. Utpal Barman working as an Assistant Professor at GIMT, Guwahati. His area of interest is Image Processing He has published many research papers in the areas of Machine learning, Image Processing.