Clustering Mixed Data Set by Fuzzy Set Partitioning

Nipjyoti Sarma
Assistant Professor
Deptt. Of CSE.
GIMT, Guwahati-17

Arindam Saha
Assistant Professor
Deptt. Of CSE.
GIMT, Guwahati-17

Adarsh Pradhan
Assistant Professor
Deptt. Of CSE.
GIMT, Guwahati-17

ABSTRACT
K mean clustering is a very popular clustering algorithm for clustering numerical data. . 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. In this work an effort have been made to put forward a proposed FCV mean algorithm which is a modified version of the traditional k-mean algorithm and is able to cluster objects having mixed type attributes i.e. numerical and categorical. For categorical data fuzzy set similarity is used and for numerical data differences from maximum dissimilarity is used. Experiment shows that the mixed data are highly clustered with high accuracy compared to other approach in literature.

General Terms
Pattern Recognition

Keywords
fuzzy set, Centroid vector, dissimilarity, categorical, numerical.

1. INTRODUCTION
Clustering a set of data in a dataset into clusters or groups is a basic operation in data mining. Basically the clustering is an unsupervised learning and it partitions a dataset on n data items having m dimensions into k numbers of distinct clusters, such that the data items in a cluster are having similar characteristics and data items of different clusters are of having different characteristics1. For this a similarity measure should be defined which should be used by an efficient algorithm to discover the clusters of most similar data items. The basic clustering approaches are divided into two main categories that are, partitioning clustering and hierarchical clustering. Given a set of objects and a clustering criterion, partitioning clustering obtains a partition of the objects into k(k, is pre-specified) clusters such that each cluster contains at least one object and each object belongs to exactly one cluster. A hierarchical method creates a hierarchical decomposition of the given set of data items. It is a nested sequence of partitions. This can be divided into two clusters namely agglomerative hierarchical clustering and Divisive hierarchical clustering. An agglomerative hierarchical or bottom up clustering starts by placing each object in its own cluster and then merges these atomic clusters into larger clusters, until all objects are in a single cluster[2]. In case of Divisive hierarchical clustering ,also called the top down approach ,the process is reverse that is, it starts
with all the objects in the same cluster. Other categories of clustering algorithm are Spectral clustering algorithm, Grid basd clustering algorithm, Density based clustering algorithm.

For example databases may contain numeric attributes like age, salary and may contain categorical attributes like gender, literate or illiterate(y/n), smoking or non-smoking(y/n) etc. Therefore it is very important to cluster data items having mixed type of attributes. To handle mixed type data items some algorithms like k-mean uses normal numeric distance measure like Euclidian distance to compute similarity between data items having categorical attributes by converting them into numeric integer values. Similarly some algorithms uses only for categorical data although they have been applied to mixed type data by transforming the numerical attributes to categorical ones via discretization. But it is inefficient due to difficulty of assigning proper numeric value to categorical attributes and categorical attribute to numeric value.

2. LITERATURE REVIEW
Yiu Ming Cheung and Hong Jia in [1] discussed a new algorithm which is based on the concept of object-cluster similarity and gives a unified similarity metric which can be simply applied to the data with categorical, numerical, and mixed attributes. The following equation gives the similarity of an object with the cluster

$$\sum_{r=1}^{d_c} \frac{1}{d_c} s(x_{ir(c)}, C_j) + \frac{1}{d_f} s(x_{i(u)}, C_j) = \frac{d_c}{d_f} s(x_{i(c)}, C_j) + \frac{1}{d_f} s(x_{i(u)}, C_j)$$

Further the similarity of categorical and numerical data are given by the following two algorithms respectively:

$$S(x_i^c, C_j) = \sum_{r=1}^{d_c} W_r \, S(x_{ir}^c, C_j)$$

$$S(x_i^u, C_j) = \frac{exp(-0.5 Dis(x_i^u, c_j))}{\sum_{t=1}^{k} exp(-0.5 Dis(x_i^u, c_t))}$$

Where x_i is the data instance and its categorical and numerical parts are x_i^c and x_i^u and cluster c , r is the domain of all possible values of categorical attribute ,d_c is the total no of categorical attribute available in the instances w is the weight of the categorical attribute of that particular data instance.

R. A. Ahmed B. Borah D. K. Bhattacharyya and J K Kalita in [2] proposed a similarity measure between two clusters that enables hierarchical clustering of data with numerical and categorical attributes. This similarity measure is derived from a frequency vector of attribute values in a cluster as shown below:

$$AttV_i(A_j) = \{(d, rf_i(d)) | d \epsilon D_j\}$$

Where D_j is the domain of the j^{th} attribute A_j and the relative frequency $rf_i(d)$ of cluster C_j is defined as:

$$rf_i(d) = \frac{freq(d)}{|C_i|}$$

Where freq(d) is defined as :

$$freq(d) = \sum_{l=1}^{|C_i|} 1 \; Such\ that\ x_{lj} = d$$

LIU Hai-tao, WEI Ru-xiang and JIANG Guo-ping in [3] proposes an iterative method to learn proper ordinal–numerical mappings for ordinal features from a given objects of mixed features, and then the similarity among data with high-dimensional mixed feature values is measured through the fuzzy partition matrix. The authors used the similarity measure S between two fuzzy set A and B as given below:

$$\left(S(A,B) = \frac{|A \cap B|}{|A \cup B|}\right) = \frac{\sum_{j=1}^{N}[U_A(x_j) \cap U_B(x_j)]}{\sum_{j=1}^{N}[U_A(x_j) \cup U_B(x_j)]}$$

Where A∩B is the intersection and AUB is the union of two fuzzy set A and B μ is the membership function and ∨ and ∧ indicates the maximal and minimal of the fuzzy set.

Z. Huang in [4] presented an algorithm which is based on the k-means paradigm by removing the numeric data limitation whilst preserving its efficiency. In this algorithm, objects are clustered against k prototypes. Using this method the intra cluster similarity is maximized by dynamically developing it. For numerical data it is like k means. To create rules for clusters decision tree induction algorithms are used. The method of this algorithm is as given below:

The method of k-prototypes algorithm can be described as follows.

At the first step select k initial prototypes from a data set **X**, one for each cluster. At the second step allocate each object in **X** to a cluster whose prototype is the nearest to it according to Equation given below:

$$d(X_i, Q_l) = \sum_{j=1}^{m_r} \left(x_{ij}^r - q_{lj}^r\right)^2 + \gamma 1 \sum_{j=1}^{m_c} \delta\left(x_{ij}^c, q_{lj}^c\right)$$

Where $X_{ij}{}^r$ and $q_{lj}{}^r$ are values of numeric attributes, whereas $x_{ij}{}^c$ and $q_{lj}{}^c$ are values of categorical attributes for object i and the prototype of cluster l. m_r and m_c are the numbers of numeric and categorical attributes. γ_l is a weight for categorical attributes for cluster

3. FUZZY SET CENTROID VECTOR OF A CLUSTER

Now, we define a Centroid of a cluster C_i having mixed type attribute as based on above definition of fuzzy membership function, termed as fuzzy Centroid vector(FCV_i) of it and is as given below[2]:

$FCV_i = \{V_{i1}, V_{i2}, V_{i3}.....V_{im}\}$

Where *m* is the total number of attributes and V_{ij} is defined as:

$$V_{ij} = \begin{Bmatrix} mean_i(A_j), if\ A_j\ is\ numerical\ attribute \\ \widetilde{A_j}\ , if\ A_j\ is\ categorical\ attribute \end{Bmatrix} -- \quad -- (1)$$

Where for numerical attribute A_j it is the mean of the j^{th} attribute of cluster C_i is

$$\text{Mean}_i(A_j) = \frac{1}{|Ci|} \sum_{k=1}^{|Ci|} p_{kj}$$

And for categorical attribute the A_j, $\widetilde{A}_j$ is calculated as:

If there is an attribute A_i with Domain D_i in a Cluster C_i, then a fuzzy set $\widetilde{A}_i$ in D_i for A_i can be defined as [9]

$\widetilde{A}_i = \{d, \mu_{Ai}(d), | \ d \in D_j\}$----------(2)

Where, $\mu_{Ai}(d)$ is the membership function for the attribute A_i and is defined as

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

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

Here $0 \leq freq(d) \leq |Ci|$ *and hence,* $0 \leq \mu_{Ai}(d) \leq 1$

Hence, the cardinality of the fuzzy set $\widetilde{A}_i$ is defined as

$$|\widetilde{A}_i| = \sum_{d \in Di} \mu_{\widetilde{Ai}}(d)$$

$$Here,\ 0 \leq |\widetilde{A}_i| \leq 1$$

4. CRITERIA FOR SIMILARITY MEASUREMENT OF CATEGORICAL AND NUMERICAL ATTRIBUTES

Then the criteria for similarity measure between, jth categorical attribute of pk and jth categorical attribute of FCVij is calculated by using the following equation.[2][9]

$$Sim_c(p_{kj}, FCV_{ij}) = \frac{(\tilde{p}_{kj} \cap \widetilde{FCV}_{ij})}{(\tilde{p}_{kj} \cup \widetilde{FCV}_{ij})}$$---------- (2)

Where $(\tilde{p}_{kj} \cap \widetilde{FCV}_{ij})$ and $(\tilde{p}_{kj} \cup \widetilde{FCV}_{ij})$ means the set intersection and set union operations between two fuzzy sets $\tilde{p}_{ij}$ and $\widetilde{FCV}_{ij}$.

In the same way for numerical the similarity measure between, j^{th} numerical attribute of p_k and j^{th} numerical attribute of FCV_i is calculated by using the following equation.Where V_{ij} contains the numerical part of FCV_i of cluster C_i.

$Sim_n(p_{kj}, FCV_i) = \sum_{j=1}^{x} \partial(p_{kj}, FCV_{ij})$---------(3)

From equation (2) and (3) ,we can write the total similarity measure for an object p_k having mixed type attribute, with a cluster C_i having centroid FCV_i is[1][3][5]-

$$\boldsymbol{Sim(P_k, FCV_i) = Sim_n(p_k, FCV_i) + Sim_c(p_k, FCV_i)}$$

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. The proposed algorithm of our work is as given below.

Algorithm : The Fuzzy Centroid Vector algorithm for clustering mixed type data :

S1: Normalize all the points in the dataset using normalization approach.

S2: Determine randomly k points from D dataset.

S3: Include the remaining points to nearest cluster fuzzy Centroid vector based on the similarity measure as in computed by the equation (2) and (3)

S4: Update each fuzzy set Centroid vector

S5: Repeat the above two steps

S6: Any improvement occurs in cluster fuzzy set Centroid vector?

S7:If yes again repeat or

S8: If no goto the k clusters and out put.

S9: Final accuracy of the clusters.

5. EXPERIMENT

The performance of the algorithm is compared with the existing counterparts. In the experiments the accuracy is estimated by hwang[2][4] as follows

$$r = \frac{1}{n}\sum_{i=1}^{k} a_i$$

Where a_i is the 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 and k- prototype[4]algorithms.

We investigated the performance of our algorithm on mixed datasets. The statistics of the selected dataset is shown in table 1. [1]

The performance of the algorithm on mixed dataset has been compared with k-means algorithm and k-prototype algorithm. The proposed algorithm is executed 100 times on each data.

The result of k-means algorithm and k-prototype algorithm has been taken from literature[1]. The clustering results are summarized in table 2. Fig.1 and fig.2 shows the accuracy comparison of the dataset in terms of accuracy and execution time and Fig.3 shows the after cluster scatter plot which indicates quality of clustering in the dataset. The dataset is obtained from UCI machine learning repository[10]

Table 1.Statistics of the mixed datasets

Dataset	Instance	Attribute (d_{cat}+d_{nu})	Class	Class Probabilities
Dermatology	366	33+1	6	30.6%,16.67%,19.67%,13.39%,14.21%,5.46%

Table 2.Clustering accuracy of our algorithm on mixed datasets in comparison with k-means and k-prototype

Dataset	k-means	k-prototype	Proposed Algo.
Dermatology	.2994±.9784	.3097±.9745	**.5791**±.5214

The following graph shows the accuracy variations in existing algorithm like K-Means and K-Prototype and the proposed Fuzzy methods on the dermatology dataset. It is a mixed dataset. The graph shows the improvement in accuracy in the proposed method in the sample dataset with random Centroid initialization. The accuracy improves from 29.94%, 30.97% to 57.91%

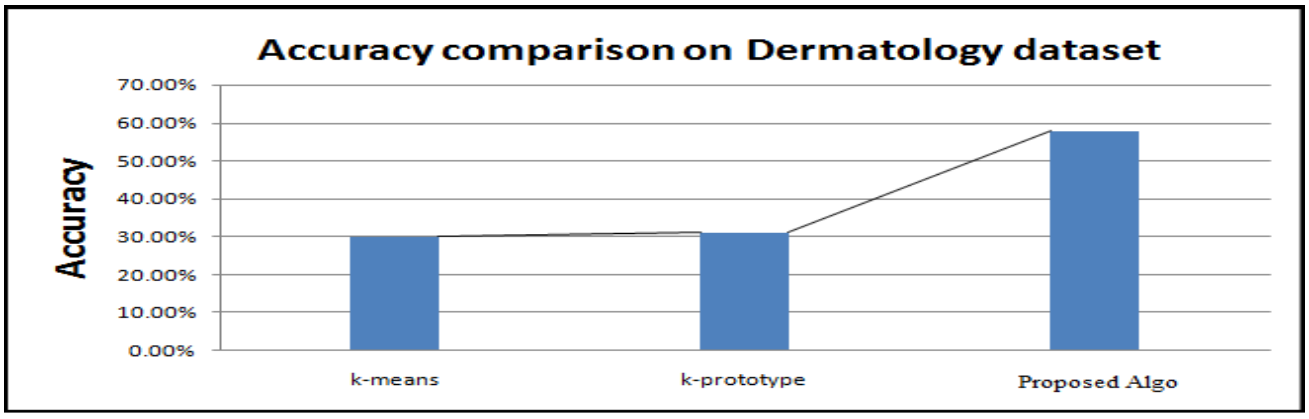

Fig. 1.Accuracy comparison on Dermatology dataset

The following figure shows the time variation between the K-Means and the proposed algorithm on dermatology dataset. It shows the improvement in time from 0.3674to 0.1712 second

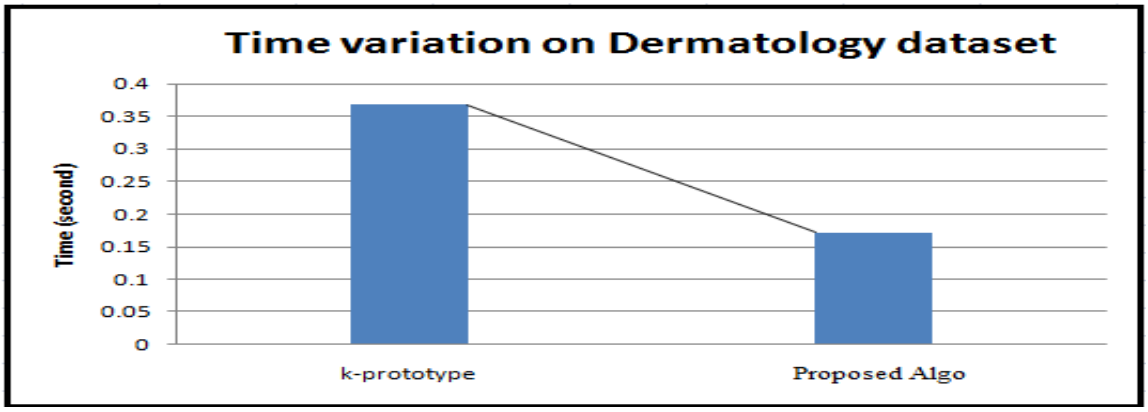

Fig. 2.Time variation on Dermatology dataset

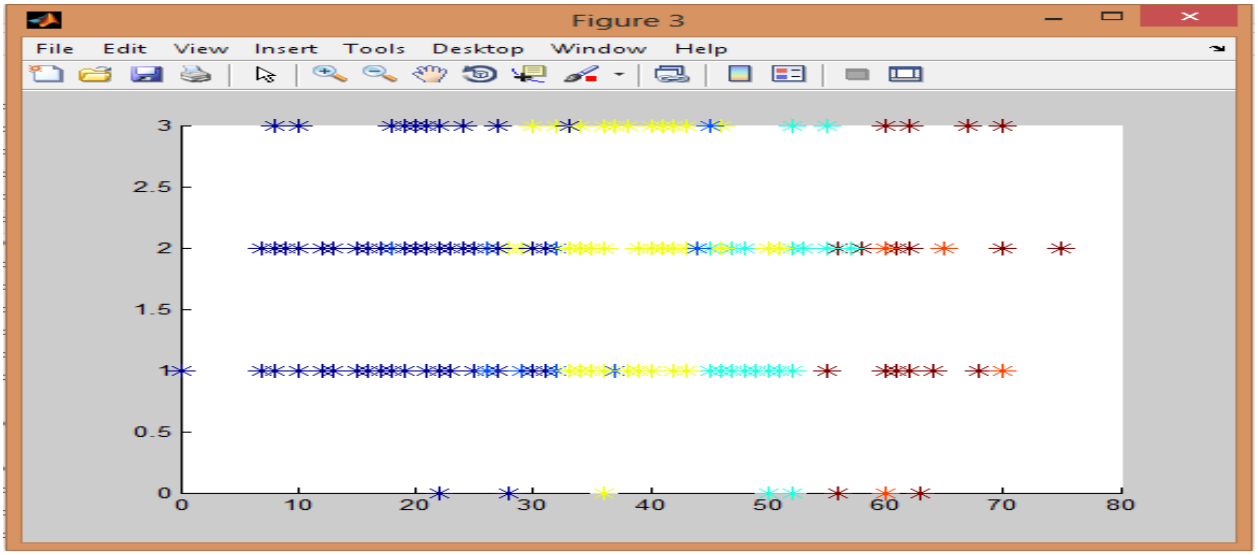

Fig. 3. Scatter plot after clustering for Dermatology dataset

In fig.3, the dermatology dataset has been plotted and the points are grouped in six different colours indicating six clusters.

6. CONCLUSION

In this paper a k means like algorithm is used to cluster mixed type of data in an efficient manner, The experimental evaluation of the proposed algorithm is done on standard mixed dataset obtained from UCI machine learning repository and compared the efficacy with its counterparts like k-mean and k-prototype algorithm. Normal Fuzzy Set approach is used to measure the similarity between the categorical attributes, while for the numeric attributes the distance between the maximum dissimilarity is used. After applying the dataset the algorithm can categorize the numeric and non numeric attributes. The datasets have been pre processed by removing the missing values before applying into the algorithm. This proposed algorithm does not perform well when missing values appears more. The proposed algorithm can also handle only numeric and only categorical features objects and it can also be used with high dimensional data. The algorithm can be applied to any kind of data mining applications where clustering is required. Specially for business data mining, Intrusion Detection System and for e-commerce related datasets that can be clustered using this algorithm.

7. REFERENCES

[1] Yiu Ming Cheung and Hong Jia in "Categorical and numerical data clustering based on a unified similarity metric with out knowing cluster number" in Pattern Recognition 46 (2013) 2228–2238, Elsevier (2013).

[2] R. A. Ahmed B. Borah D. K. Bhattacharyya and J K Kalita in HIMIC : A Hierarchical Mixed Type Data Clustering Algorithm" in"(2005) in http://citeseerx.isi.psu.edu/viewdoc/download?doi=10.1.1.61.6369 & rep=rep1 & type=pdf,

[3] LIU Hai-tao, WEI Ru-xiang and JIANG Guo-ping in Similarity measurement for data with high-dimensional and mixed feature values through fuzzy clustering" in Proceedings of IEEE conference,2012 ,pp-617-621,

[4] Z. Huang in "Clustering large data sets with mixed numeric and categorical values" in Proceedings of the First Pacific-Asia Conference on Knowledge Discovery and Data Mining, , pp. 21–34.1997,

[5] 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,

[6] 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)

[7] 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.

[8] 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,

[9] Fuzzy logic and nural networks by M. Amrithavalli first, second and third chapter Fourth reprint)published by scitech publications(India)pvt. Limited, 2010.

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