# Ensemble Fuzzy Clustering for Mixed Numeric and Categorical Data

J.Suguna
M.Sc., P.B.D.C.A., M.Phil., Ph.D., M.Ed.,
Associate Professor,
Department of Computer Science,
Vellalar College for Women,
Erode-12

M. Arul Selvi
M.C. A., M.Phil.,
Research Scholar,
Vellalar College for Women,
Erode-12

## ABSTRACT

In data mining, clustering is one of the major tasks and aims at grouping the data objects into meaningful classes (clusters) such that the similarity of objects within clusters is maximized, and the similarity of objects between clusters is minimized. The dataset sometimes may be in mixed nature that is it may consist of both numeric and categorical type of data. Naturally these two types of data may differ in their characteristics. Due to the differences in their characteristics, in order to group these types of mixed data, it is better to use the ensemble clustering method which uses split and merge approach to solve this problem. In this paper, the original mixed dataset is splitted into numeric dataset and categorical dataset and clustered using both traditional clustering algorithms (K-Means and K-Modes) and fuzzy clustering algorithms (Fuzzy C-Means and Fuzzy C-Modes). The resultant clusters are combined using ensemble clustering methods and evaluated by both f-measure and entropy measure. It is found that splitting is more beneficial and applying fuzzy clustering algorithms yields better results than traditional clustering algorithms.

## Keywords

Clustering, Ensemble clustering, Mixed dataset, Numeric dataset, Categorical dataset.

## 1. INTRODUCTION

Data mining [1] is the process used to analyze large quantities of data and gather useful information from them. It extracts the hidden information from large heterogeneous databases in many different dimensions and finally summarizes it into categories and relations of data. Clustering and classifications are the two main techniques of data mining followed by association rules, predictions, estimations and regressions.
Clustering is a process of grouping objects with similar properties. Any cluster should exhibit two main properties; low inter-class similarity and high intra-class similarity. Clustering is an unsupervised learning i.e. it learns by observation rather than examples. There are no predefined class label exists for the data points.

Clustering constitutes a major class of data mining algorithms. In general, the output produced by a clustering algorithm will be the assignment of data objects in dataset to different groups. In other words, it will be sufficient to identify each data object with a unique cluster label [2]. The clustering technique has been extensively studied in many fields such as pattern recognition, similarity search and trend analysis[3].

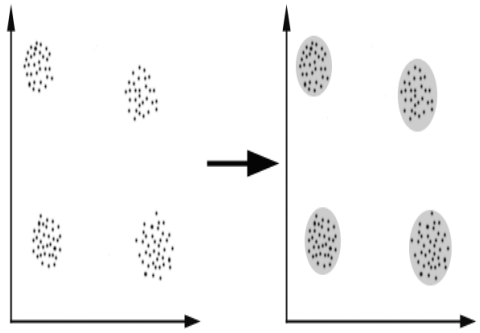

**Fig 1: Clustering Process**

## 2. ENSEMBLE CLUSTERING

Ensemble Clustering [4] is the method to combine several runs of different clustering algorithms to get a common partition of the original dataset, aiming for consolidation of results from a portfolio of individual clustering results. Ensemble clustering is a promising approach that combines the results of multiple clustering algorithms to obtain a consensus partition by merging different partitions based upon well-defined rules. Ensemble clustering methods [5] have been developed to improve the robustness and accuracy of clustering algorithms, as well as the ability to capture the structure of complex data. Ensemble methods can combine both different data and different clustering algorithms.

## 3. ENSEMBLE FUZZY CLUSTERING

In this paper, an unsupervised ensemble fuzzy clustering approach have been proposed that permit to dispose both of the flexibility of the fuzzy sets and the robustness of the ensemble methods. Use of ensemble clustering approach merges the results of two different fuzzy clustering algorithms namely Fuzzy C-Means and Fuzzy C-Modes. The overview of the ensemble fuzzy clustering algorithm framework is described in Figure 2 and the steps are as follows:

1) The mixed dataset is divided into two sub datasets: the pure categorical dataset and the pure numeric dataset.

2) Fuzzy C-Means clustering algorithm is applied for numeric data and Fuzzy C-Modes for categorical data to form the clusters.

3) Combine the resultant fuzzy clusters by applying ensemble clustering methods

4) Evaluate the ensemble fuzzy clusters using validation measures such as f-measure and entropy.

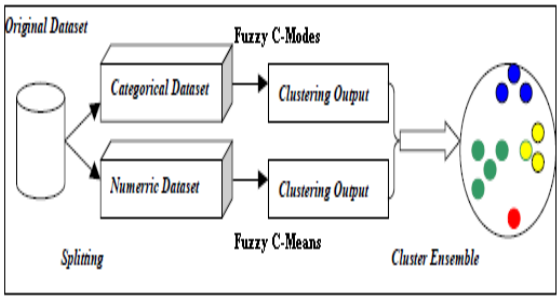


**Fig 2: Overview of Ensemble Fuzzy Clustering**

## 4. FUZZY CLUSTERING ALGORITHMS

The base clustering algorithms used for the ensemble fuzzy approach used here are Fuzzy C-Means and Fuzzy C-Modes.

### 4.1. Fuzzy C-Means

Fuzzy C-Means (FCM) [6] is a data clustering technique which allows one piece of data to belong to two or more clusters. A single point can have partial membership in more than one class. There can be no empty classes and no class that contains no data points. FCM attempts to find the most characteristic point in each cluster, which can be considered as the 'center' of the cluster. This algorithm works by assigning membership to each data point corresponding to each cluster center on the basis of distance between the cluster center and the data point.

Fuzzy C-Means algorithm is based on the minimization of the following objective function [6] :

$$J(U,V) = \sum_{i=1}^{n}\sum_{j=1}^{c}(\mu_{ij})^m \left\|x_i - v_j\right\|^2$$

where, '$||x_i - v_j||$' is the Euclidean distance between $i^{th}$ data and $j^{th}$ cluster center.

Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the update of membership $\mu_{ij}$ and the cluster centers $v_j$ by:

$$\mu_{ij} = 1 / \sum_{k=1}^{c}(d_{ij} / d_{ik})^{(2/m-1)}$$

$$v_j = (\sum_{i=1}^{n}(\mu_{ij})^m x_i) / (\sum_{i=1}^{n}(\mu_{ij})^m), \forall j = 1, 2, .....c$$

where '$n$' is the number of data points, '$m$' is the fuzziness index m € [1, ∞], '$c$' represents the number of cluster center, $v_j$ represents the $j^{th}$ cluster center, $\mu_{ij}$ represents the membership of $i^{th}$ data to $j^{th}$ cluster center, $d_{ij}$ represents the Euclidean distance between $i^{th}$ data and $j^{th}$ cluster center.

#### *4.1.1. Steps of Fuzzy C-Means algorithm*

1) Randomly assign values for each $\mu_{ij}$, where $\mu_{ij} \geq 0$

2) Repeat

3) Calculate the cluster centroids $v_j$ using the above formula

4) Using the newly derived $v_j$, update the value of $\mu_{ij}$ by using the above formula

5) Until the convergence is obtained.

### 4.2. Fuzzy C-Modes

In most fuzzy versions of clustering algorithms, the assigned memberships of data to a cluster are fuzzy, but the centroid itself is not fuzzy. In this paper, existing Fuzzy K-Modes algorithm is modified as Fuzzy C-Modes algorithm that uses fuzzy centroids for clustering categorical data. Fuzzy centroids are a set of fuzzy values that contain category values and their confidence degrees for each attribute.

To cluster categorical data, the proposed Fuzzy C-Modes algorithm extends the hard K-Modes algorithm based on the Fuzzy C-Means type procedure. It updates the cluster centers at each iteration by measuring the distance between each cluster centroid and each object.

Let $X = \{X1, X2, ....., Xn\}$ be a set of $n$ objects. Object $Xi$ is represented as $[x_i,1, x_i,2, ....., x_i, m]$ and $X_i = X_k$ if $x_{i.j} = x_{k,j}$ for $1 \leq j \leq m$. The Fuzzy C-Modes algorithm is to cluster the data $X$ into $k$ clusters by minimizing the objective function [9]:

$$F(W,Z) = \sum_{l=1}^{k}\sum_{i=1}^{n} W_{li}^{\alpha} d(Z_l, X_i)$$

Subject to

$$0 \leq W_{li} \leq 1 \; ; \; 1 \leq l \leq k \; ; \; 1 \leq i \leq n,$$

$$\sum_{l=1}^{k} W_{il} = 1 \; , \; 1 \leq i \leq n$$

and $$0 < \sum_{i=1}^{n} W_{li} < n, \; 1 \leq l \leq k$$

While $*_{li}$ is the membership degree of data $X_i$ to the $l^{th}$ cluster, and is an element of the $(k \times n)$ partition matrix

$W = [*_{li}]$. $C* = [C*_1, C*_2, .....C*_l, .....C*_k]$ and $C*_l$ is the *l*th cluster centre and the parameter $\alpha$ controls the fuzziness of the membership of each object.

The proposed Fuzzy C-Modes algorithm uses a simple matching dissimilarity measure for categorical objects as follows. Here, let *X* and *Y* be two categorical objects represented by $[x_1, x_2, .....x_m]$ and $[y_1, y_2, .....y_m]$ respectively.

$$d(X,Y) = \sum_{j=1}^{m} \delta(x_j, y_j)$$

Where

$$\delta(x_j, y_j) = \begin{cases} 0, x_j = y_j \\ 1, x_j \neq y_j \end{cases}$$

The measure *d* satisfies a metric space on the set of categorical objects, and is also a kind of generalized Hamming distance.

### *4.2.1. Steps of Fuzzy C-Modes algorithm*

1.Randomly assign a cluster label to each object, that is initialize the cluster membership $W^{(1)}$. Determine $C*^{(1)}$ such that $F(W^{(1)}, C*^{(1)})$ is minimized. Set t=1.

2.Determine $W^{(t+1)}$ such that $F(W^{(t+1)}, C*^{(t)})$ is minimized. If $F(W^{(t+1)}, C*^{(t)}) = F(W^{(t)}, C*^{(t)})$, then stop; otherwise set $t = t + 1$ and goto Step 3.

3.Determine $C*^{(t+1)}$ such that $F(W^{(t+1)}, C*^{(t+1)})$ is minimized. If $F(W^{(t+1)}, C*^{(t+1)}) = F(W^{(t+1)}, C*^{(t)})$, then stop; otherwise goto Step 2.

## 5. CLUSTER EVALUATION

One of fundamental challenges of clustering [12] is how to evaluate results, without auxiliary information. A common approach for evaluation of clustering results is to use validity indexes. Clustering validity approaches can use two criteria: External criteria (evaluate the result with respect to a pre-specified structure) and internal criteria (evaluate the result with respect a information intrinsic to the data alone).
A common approach for evaluation of clustering results is to use external cluster validity measures. The external validation measures [13] are used to measure the extent to which cluster labels affirm with the externally given class labels. Many different measures for evaluating the performance of clustering have been proposed. The clusters formed are evaluated and interpreted according to the distance between data points and cluster centers of each cluster.
In this paper, to evaluate the performance of the ensemble hard clustering and ensemble fuzzy clustering approach, two external cluster validation indices are used, namely f-measure and entropy measures [14].

## 6. RESULTS AND DISCUSSION

Various mixed dataset are taken from UCI Machine Learning Repository such as Abalone, Servo, Teaching Assistant Evaluation (TAE), Liver disorders and Thyroid to evaluate the performance of different clustering algorithms for splitted dataset as shown in Table 1. Experimental results are given below:

**Table 1. Dataset Description**

| DATASETS | ATTRIBUTES | DATA POINTS |
|---|---|---|
| Abalone | 8 | 4177 |
| Servo | 4 | 167 |
| Liver Disorders | 7 | 345 |
| Teaching Assistant Evaluation | 5 | 151 |
| Thyroid | 5 | 215 |

### 6.1.Ensemble Hard Clustering

The mixed dataset is splitted into pure numeric dataset and pure categorical dataset and applied K-Means clustering algorithm for numeric dataset and K-Modes clustering algorithm for categorical dataset and combined the clusters by the ensemble hard clustering method. Now the resultant cluster is evaluated to measure the performance and the results are given below [Figure 3 and Table 2].

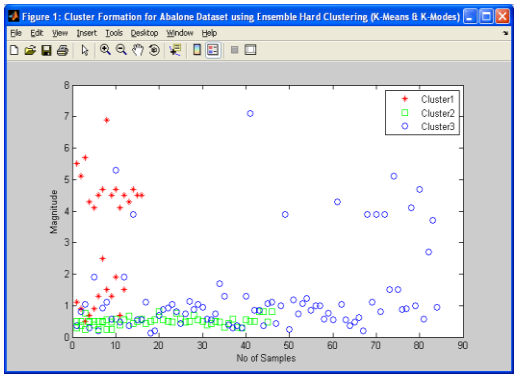


**Fig 3: Ensemble Hard Clustering**

**Table 2. Ensemble Hard Clustering**

| Clustering Measures | Dataset | Ensemble Hard Clustering |
|---|---|---|
| F-Measure | Servo | 0.4342 |
| | Abalone | 0.4802 |
| | Liver Disorders | 0.5150 |
| | TAE | 0.6168 |
| | Thyroid | 0.7715 |
| Entropy Measure | Servo | 0.4571 |
| | Abalone | 0.3114 |
| | Liver Disorders | 0.3118 |
| | TAE | 0.4210 |
| | Thyroid | 0.4354 |

## 6.2. Ensemble Fuzzy Clustering

Fuzzy C-Means clustering algorithm is applied for numeric dataset and Fuzzy C-Modes clustering algorithm for categorical dataset and combined the clusters by the ensemble fuzzy clustering method. Now the resultant cluster is evaluated to measure the performance and the results are given below [Figure 4 and Table 3].

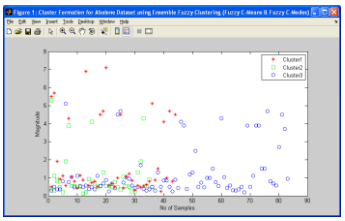

**Fig 4: Ensemble Fuzzy Clustering**

**Table 3. Ensemble Fuzzy Clustering**

| Clustering Measures | Dataset | Ensemble Fuzzy Clustering |
|---|---|---|
| F-Measure | Servo | 0.8963 |
| | Abalone | 0.8059 |
| | Liver Disorders | 0.8850 |
| | TAE | 0.9592 |
| | Thyroid | 0.9119 |
| Entropy Measure | Servo | 0.1677 |
| | Abalone | 0.0114 |
| | Liver Disorders | 0.1453 |
| | TAE | 0.1868 |
| | Thyroid | 0.0081 |

## 6.3. Comparing Ensemble Hard Clustering and Ensemble Fuzzy Clustering

Finally comparing the results of ensemble hard clustering with ensemble fuzzy clustering, it is found that applying fuzzy logic algorithms to both numeric and categorical dataset is more beneficial since it yields better results as shown below [Table 4, Figure 5 and Figure 6].

**Table 4. Ensemble Hard Clustering Vs Ensemble Fuzzy Clustering**

| Clustering Measures | Dataset | Ensemble Hard Clustering | Ensemble Fuzzy Clustering |
|---|---|---|---|
| F-Measure | Servo | 0.4342 | 0.8963 |
| | Abalone | 0.4802 | 0.8059 |
| | Liver Disorders | 0.5150 | 0.8850 |
| | TAE | 0.6168 | 0.9592 |
| | Thyroid | 0.7715 | 0.9119 |
| Entropy Measure | Servo | 0.4571 | 0.1677 |
| | Abalone | 0.3114 | 0.0114 |
| | Liver Disorders | 0.3118 | 0.1453 |
| | TAE | 0.4210 | 0.1868 |
| | Thyroid | 0.4354 | 0.0081 |

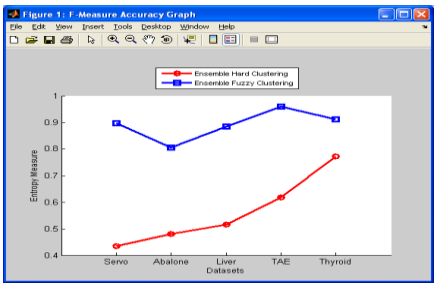


**Fig 5: F-Measure Graph**

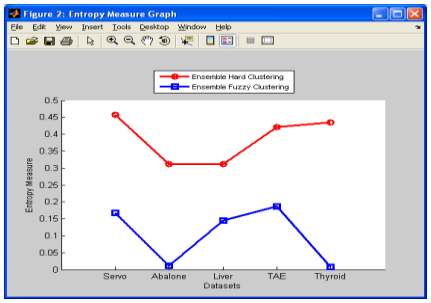


**Fig 6: Entropy Measure Graph**

## 7. CONCLUSION

In general, clustering aims at discovering groups and identifying interesting distribution and patterns in data sets. The conventional K-Modes algorithm is capable of efficiently clustering categorical data; however, it uses hard centroids for categorical attributes and a simple distance measure to classify boundary data. To address these shortcomings of the K-Modes algorithm, the existing fuzzy clustering algorithm is modified so that it uses fuzzy centroids for clustering categorical data.

In the proposed algorithm, the distance measure between data and fuzzy centroids is defined and the method for updating fuzzy centroids is formulated. The suitability of the proposed Fuzzy C-Modes algorithm (for clustering categorical dataset) and the resultant clusters of Ensemble Fuzzy Clustering is clearly experimented through different evaluation measures namely f-measure and entropy.

## REFERENCES

[1] Jain. A.K, Murty. M.N, and Flynn. P. J, "Data clustering: a review", ACM Computing Surveys, 1999.

[2] Kotsiantis. S, Pintelas. P, "Recent Advances in Clustering: A Brief Survey", WSEAS Transactions on Information Science and Applications, Vol. 1, No. 1 (73-81), 2004.

[3] Zengyou He, Xiaofei Xu, and Shengchun Deng "A Link Clustering based approch for Clustering Categorical Data", Department of Computer Science and Engineering, Harbin Institute of Technology, China

[4] Jagannatha Reddy. M.V and Dr. Kavitha. B, "Efficient Ensemble Algorithm for Mixed Numeric and Categorical Data", IEEE International Conference on Computational Intelligence and Computing Research, Dec, 2010.

[5] Roberto Avogadri and Giorgio Valentini, "Ensemble clustering with a fuzzy approach", Department of Science and Information (DSI), University of Milan, Italy.

[6] Velmurugan. T and Santhanam. T, "Clustering Mixed Data Points using Fuzzy C-Means Clustering Algorithm for Performance Analysis", International Journal on Computer Science and Engineering Vol. 2, No. 9, 2010.

[7] Aranganayagi. S and Thangavel. K, "Extended K-Modes with Probability Measure", International Journal of Computer Theory and Engineering, Vol. 2, No. 3, June, 2010.

[8] Aranganayagi. S and Thangavel. K, "Improved K-Modes for Categorical Clustering using Weighted Dissimilarity Measure", International Journal of Information and Mathematical Sciences, Vol. 2, No. 5, 2009.

[9] Michael K. Ng and Liping Jing, "A New Fuzzy K-Modes Clustering Algorithm for Categorical Data", International Journal of Granular Computing, Rough Sets and Intelligent Systems, Vol. 1, No. 1, 2009.

[10] Zhexue Huang and Michael K. Ng, "A Fuzzy K-Modes Algorithm for Clustering Categorical Data", IEEE Transactions on Fuzzy Systems, Vol. 7, No. 4, August 1999.

[11] Dae-Won Kim, Kwang H. Lee, and Doheon Lee, "Fuzzy clustering of categorical data using fuzzy centroids", Pattern Recognition Letters 25 (1263–1271), 2004.

[12] L.Jegatha Deborah, R.Baskaran, A.Kannan, "A Survey on Internal Validity Measure for Cluster Validation", International Journal of Computer Science & Engineering Survey (IJCSES) Volume 1, Issue No.2, November 2010.

[13] Erendira Rendon, Itzel Abundez, Alejandra Arizmendi and Elvia M. Quiroz, "Internal versus External cluster validation Indexes", International Journal of Computers and Communications, Issue 1, Volume 5, 2011.

[14] Satya Chaitanya Sripada and Dr. Sreenivasa M. Rao**,** "Comparison of Purity and Entropy of K-Means Clustering and Fuzzy C Means Clustering", Indian Journal of Computer Science and Engineering (IJCSE), Vol. 2, No. 3, June, 2011.

[15] Revati Raman Dewangan , Lokesh Kumar Sharma and Ajaya Kumar Akasapu, "Fuzzy Clustering Technique for Numerical and Categorical dataset", International Journal on Computer Science and Engineering , NCICT 2010, Special Issue.