Clustering Method for Mixed Categorical and Numerical Data

Prajapati Madhavi[1], J. S. Dhobi[2]

[1]*Student, Computer science and engineering, GEC Gandhinagar, Gujarat, India*
[2] *Professor, Computer science and engineering, GEC Gandhinagar, Gujarat, India*

ABSTRACT

Clustering is the process of discovering a set of categories to which objects should be assigned. A cluster is comprised of a number of similar objects collected or grouped together. The current requirements to cluster real world data sets are scalability, ability to handle any kind of data like categorical and numerical. Traditional algorithm can cluster categorical or numerical data but not the both. Various clustering algorithms have been developed to group data into clusters. However, these clustering algorithms work effectively either on pure numeric data or on pure categorical data, most of them perform poorly on mixed categorical and numerical data types in previous k-means algorithm was used but it is not accurate for large datasets. This study includes the discussion about the different clustering algorithm, its advantages and limitations. An efficient method is proposed for clustering both numerical and categorical data.

Keyword: *Mixed data clustering, Entropy based clustering, data mining*

1. INTRODUCTION

The process of grouping a set of physical or abstract objects into classes of *similar* objects is called clustering. A cluster is a collection of data objects that are *similar* to one another within the same cluster and are *dissimilar* to the objects in other clusters.

We are living in a world full of data. Every day, people encounter a large amount of information and store or represent it as data, for further analysis and management. One of the vital means in dealing with these data is to classify or group them into a set of categories or clusters.

Basically, classification systems are either supervised or unsupervised, depending on whether they assign new inputs to one of a finite number of discrete supervised classes or unsupervised categories, respectively.

In supervised classification, the mapping from a set of input data to a finite set of discrete class labels is modeled in terms of some mathematical function, where is a vector of adjustable parameters.

In unsupervised classification, called clustering or exploratory data analysis, no labeled data are available. The goal of clustering is to separate a finite unlabeled data set into a finite and discrete set of "natural," hidden data structures, rather than provide an accurate characterization of unobserved samples generated from the same probability distribution.

In cluster analysis a group of objects is split up into a number of more or less homogeneous subgroups on the basis of an often subjectively chosen measure of similarity (i.e., chosen subjectively based on its ability to create "interesting" clusters), such that the similarity between objects within a subgroup is larger than the similarity between objects belonging to different subgroups.

Conventional clustering techniques have been focused on a single type of attributes, either numerical or categorical attributes of datasets.

As mixed attribute type datasets are common in real life, clustering techniques for mixed attribute type datasets is required in various informatics fields such as bio informatics, medical informatics, geo informatics, information retrieval, to name a few.

However, conventional approaches are designed mainly for a single type attributes they are not appropriate for mixed attribute type datasets. Recently, some approaches to clustering for mixed attribute have been introduced by converting categorical attribute values to numerical ones and applying traditional clustering algorithms with only numerical values [1].

Recently, clustering techniques for mixed-type attribute datasets have been developed in an effort to use traditional algorithms without transformation. What these techniques have in common is their division of mixed type attributes into categorical attributes and numerical attributes, as well as their uses of data analysis techniques.

For example, k-prototype clustering, which is an extension of k-means clustering, is one of the clustering algorithms for mixed-attribute datasets without transformation. However, the clustering performance of k-prototype clustering depends on the selection of optimal cluster number k as an initial prototype for clustering, as the cluster number is selected randomly or user-defined. Therefore, it has been a challenge in clustering to determine the optimal cluster number.

The proposed clustering framework consists of three main steps (see Fig. 1). In Step 1, we use an entropy based similarity measure with only categorical attributes and extract candidate cluster numbers by evaluating the difference of values with entropy based similarity measure. We analyze the difference of total entropy among clusters in an exhaustive manner by reducing the number of clusters until all of clusters merge into one cluster and extract candidate cluster numbers by using the difference in entropy values.

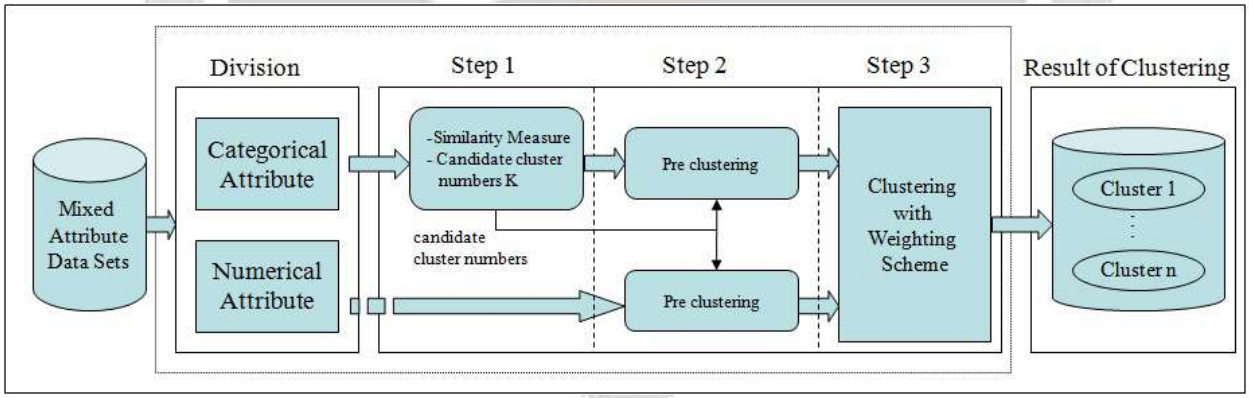

Fig-1 Overview of the Proposed Clustering Framework

Second, we apply the extracted candidate cluster numbers K from Step 1 to cluster the dataset using only numerical attributes (Step 2). Now, we have two clustering results, one by using only categorical attributes and the other by using numerical ones. Note that the number of clusters is decided solely by categorical attributes.

In Step 3, a weighting scheme is applied using the degree of balance in number of objects in the clusters. After the pre-clustering, we can compare how two clustering results are balanced. The main point of the weighting scheme is to put more weight onto the better-balanced clustering between categorical and numerical one. 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.

2. RELATED WORKS

For categorical attributes, Squeezer algorithm reads each tuple t in sequence over all dataset and determine it using the similarity values between t and clusters [2]. ROCK is an adaptive one of an agglomerative hierarchical clustering algorithm [3] and CACTUS is a fast summarization based algorithm [4].

Amir Ahmad and Lipika Dey [5] proposed a k-mean clustering algorithm for mixed numeric and categorical data. Ming-Yi Shih, et al. [6] proposed a two-step method for clustering mixed categorical and numerical data. It first constructs similarity or relationships among categorical attributes based on their co-occurrence and then those categorical attributes are converted into numeric data. Finally, the hierarchical and partitioning clustering algorithms used for clustering the data including converted into numeric data.

For the similarity measure, entropy concept has been used for categorical data in the literature. As an element of information theory, entropy is also a measure of the uncertainty with a random variable. The total entropy value is created using a classical entropy theory, Shannon Entropy [7]. Entropy based clustering is a method that finds similar objects in clusters based on their total entropy values and determines the number of clusters and identifies the location of the cluster center. The basic idea of entropy based clustering is that the lowest entropy value between two objects represents the highest similarity among objects [8].

3. PROPOSED FRAMEWORK

Our proposed framework starts with the division of mixed attribute datasets into categorical and numerical attributes sub datasets.

First we dividing mixed attribute type datasets into categorical and numerical attributes sub dataset. We measure the similarity of categorical attribute sub dataset by utilizing entropy based similarity measure using an agglomerative process. Based on the results of the similarity measure, we analyze the changes in total entropy total entropy value while building clusters in agglomerative way and extracting candidate cluster numbers, K (i.e., a list of desirable cluster numbers), for mixed attribute type dataset clustering.

As a criterion function, similarity measure between objects is one of the primary steps in clustering process. Entropy can be used to measure the uncertainty of random variables.
Similarity measure for numerical attribute Distance functions such as Euclidean distance are used as since they represent the inherent distance meaning between numerical attributes but they are not for categorical attribute.

For categorical attribute it is difficult to measure similarity in that its values cannot be directly compared each other because they are not ordered nor continuous, whereas numerical attributes are ordered and continuous.

The entropy is simply defined as follows:

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x)$$

Where p(x) is the probability mass function of the random variable x and X is the set of possible outcomes of *x*.

We consider that a dataset $X=SC+SN$ (Where SC is a subset of categorical attributes and SN is a subset of numerical attributes) in the presence of R objects.

Then, $SC=\{D1, D2, ..., Dcn\}$, where Di is ith categorical attribute. $SN=\{N1, N2, ..., Nnn\}$, where Ni is ith numerical attribute. Ati is a set of distinct values in ith categorical attribute(Di).

The definition of total entropy in a given dataset can be redefined as

$$H(S\ C) = -\sum_{i=1}^{cn} \sum_{v \in At_i} p(v) \log p(v)$$

Where p(v) is the probability of occurrence of value v in *i*th categorical attribute(Di).

The entropy criterion for optimal candidate cluster *CK* as follows:

$$OC(C^K) = \frac{1}{cn}[(H(SC) - \frac{1}{K}\sum_{k=1}^{K} H(C_k)]$$

Where $H(SC)$ is the total entropy in the given dataset, $\frac{1}{K}\sum_{k=1}^{K} H(C_k)$ is the average entropy of CK.

We can extract candidate cluster numbers for clustering by exploring the difference of each cluster's average entropy while clusters are merged in an agglomerative way.

Assume that each object in a dataset as a singleton cluster. If two highly similar clusters are merged into one, then the variance of average entropy will not change much.

We notate the difference of average entropy of clusters (Diffent) as follow

$$\text{Diffent}(Ca,Cb) = H(Ca \cup Cb) - 1/2\ [H(Ca)+H(Cb)] \geq 0$$
$$\text{Diffent}(Ca,Cb) = 0 \quad \text{Where } Ca \text{ is identical with } Cb$$

	N1	N2	N3	...	Nn
N1	Diff(N1, N1)	Diff(N1, N2)	Diff(N1,N3)	...	Diff(N1, Nn)
N2		Diff(N2, N2)	Diff(N2, N3)	...	Diff(N2, Nn)
N3			Diff(N3, N3)	...	Diff(N3,Nn)
.				...	...
					...
.					...
Nn					Diff(Nn, Nn)

Fig-2 Initializing Diff_{ent}

	N1	...	Ni	...	Nj	...	Nn
N1	0	...	Updated	...	Deleted	...	.
...		...	Updated	...	Deleted	...	.
Ni			0	Updated	**Merged**	Updated	Updated
...				...	Deleted	...	.
Nj					0	Deleted	Deleted
...						...	.
Nn							0

Fig-3 Snapshot of Merging Cluster

Let DK be the difference of entropy between AH(CK−1) and AH(CK). The algorithm computes the set of entropy values, D={ DK } for all 2≤K≤R until the number of cluster being one. By monitoring the value changes in DK, the algorithm determines the candidate cluster numbers as follows: 1) find a subset of D, i.e., DS, with all DK which satisfies DK−1<DK and DK>DK+1. 2) Select P (P is an input parameter) greatest values of DK values in DS then the set of K values becomes the candidate cluster numbers.

In Step 2, to determine appropriate weights we pre-cluster each type attributes based on the candidate cluster numbers K.

Using the numbers, the given dataset is clustered with only categorical and numerical attributes respectively.

In general, well-structured clustering shows that the numbers of objects in clusters are balanced. Our approach sets a priority on one type of attribute over the other by comparing the balance of clustering of one result with categorical attributes to that with numerical ones. We first give more weight for the better balanced attribute type between categorical attribute and numerical one to improve the results of the final clustering.

The weight condition of our mixed attribute clustering is defined as follows:

1. $\omega t = \omega c + \omega n$

Where ωc is weight for categorical attribute and ωn is weight for numerical attribute

2. $\omega t = 1$ *and* $0 \leq \omega c \leq 1, \quad 0 \leq \omega n \leq 1.$

After determining the weight, the final clustering is performed based on the following similarity measure using the weights:

$$S_M = \omega_C S_C + \omega_N S_N$$

Where SM is the similarity of mixed attribute type datasets (i.e., total) and SC, SN is the similarity of categorical and numerical attributes.

In Step 3, we cluster mixed attribute type datasets by using the candidates cluster numbers and weighting each type of attributes with different values.

With SM values, our algorithm utilizes the agglomerative hierarchical clustering method for the final result.

4. EXPERIMENTAL EVALUATION

Here we use weather dataset. The dataset has 5 attributes total (outlook, temperature, humidity, windy, play), 3(outlook, windy, play) of them are categorical and others are numerical. The dataset has 14 instances.

Following figure shows how this dataset clustered. It shows that two clusters are created using this approach one of them has 5 instances (36%) and other has 9 instances (64%).

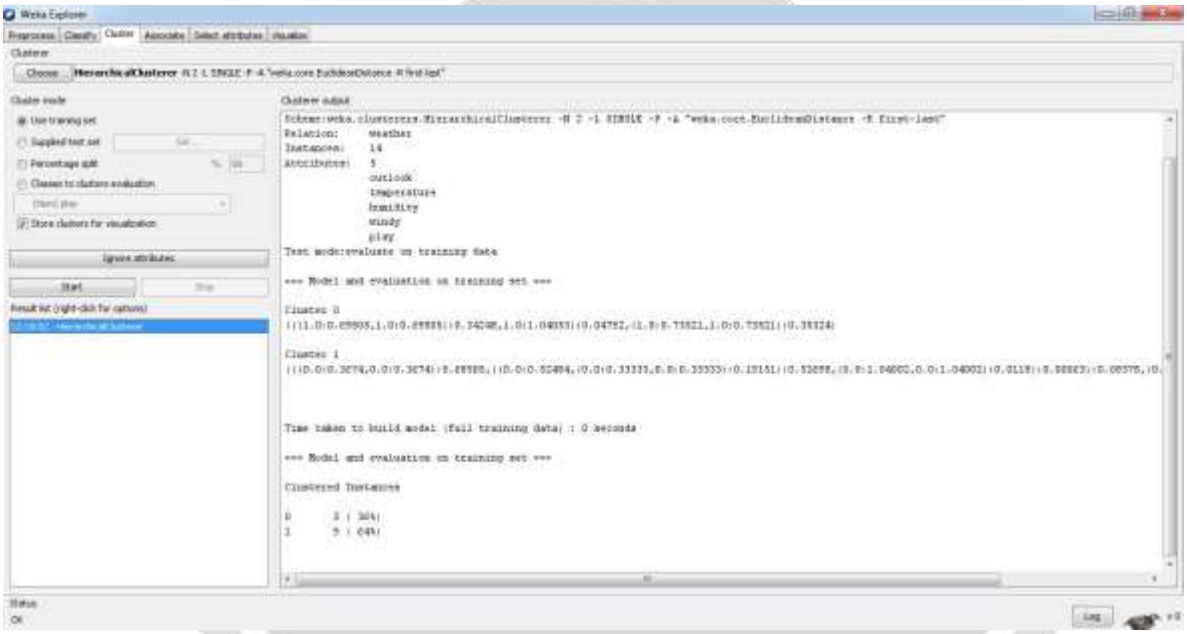

Fig-4 Output of Proposed algorithm

Following figure show how cluster visualize.

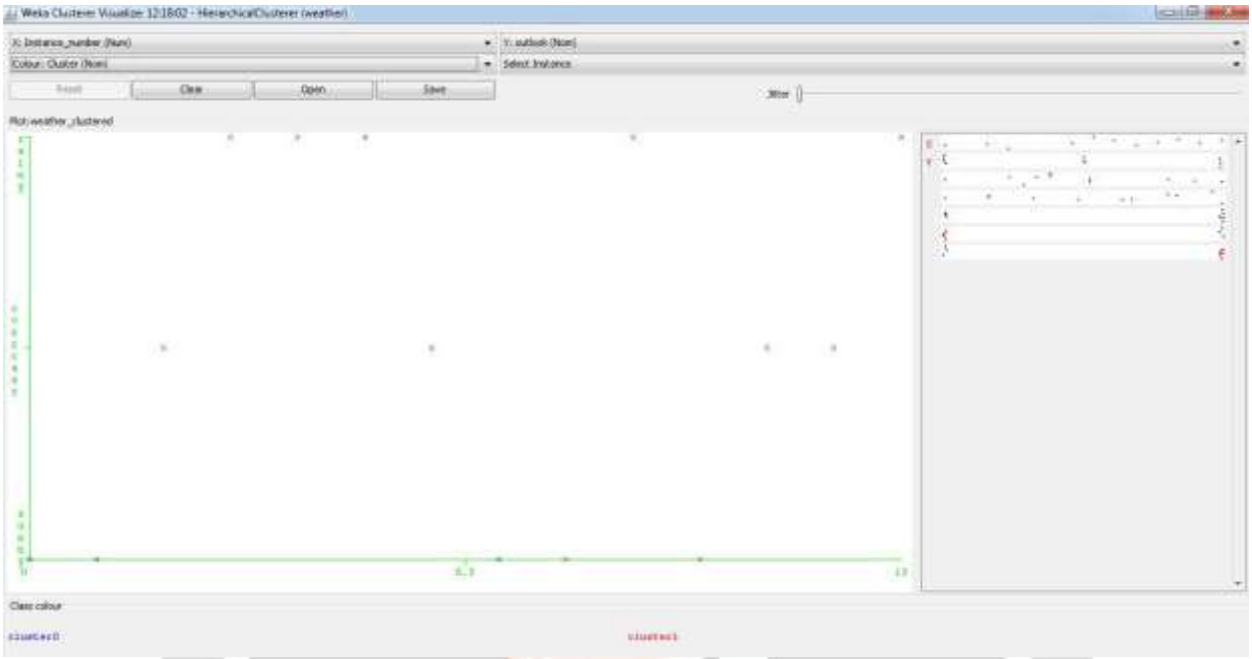

Fig-5 Visualization of Cluster

Fig. 7 shows that the comparison using the average accuracy between our approach and a conventional one. The conventional approach is transforming categorical data to numerical data without background knowledge of the categorical data set and using Euclidean distance as a similarity measure. Then, it uses k-means algorithm for clustering. The result of our approach is better than the conventional one.

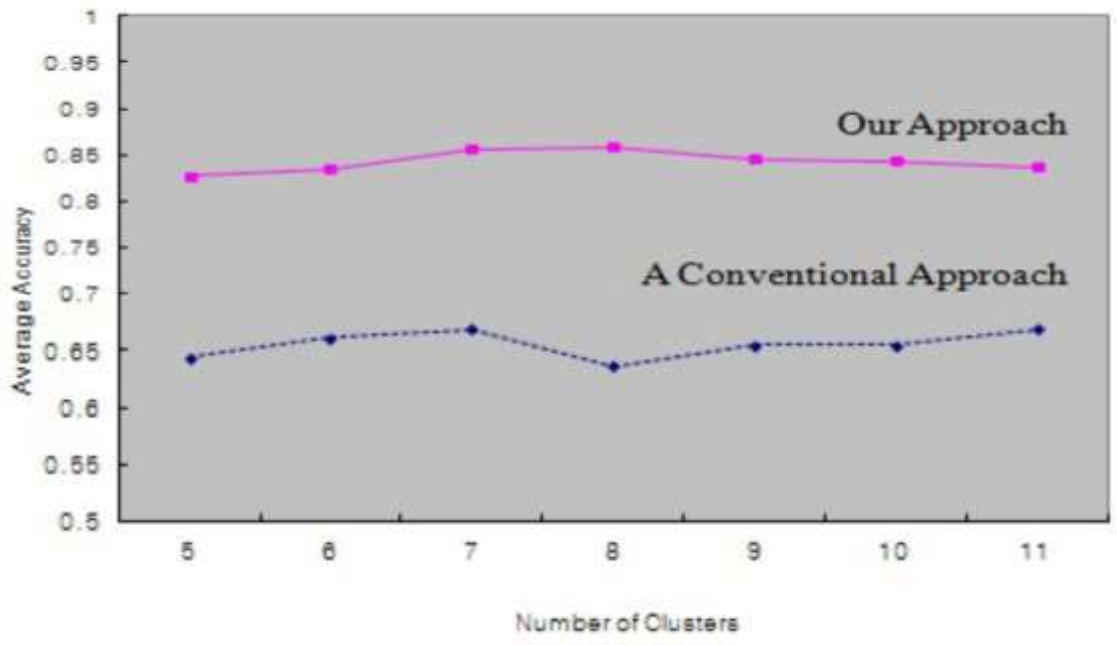

Fig-6 Comparison between our approach and conventional one

5. CONCLUSION

In this paper, we proposed a clustering framework for mixed attribute type dataset. Conventional clustering algorithms have focused on a single attribute type such as either numerical or categorical attribute. There exist approaches to clustering mixed attribute type datasets by transforming one type into the other. One of challenges in such approaches is the loss of information during the conversion due to the difference between two data types.

Our proposed framework uses a pre-clustering process focused on categorical attributes to better understand which type of attributes can be more influential in clustering mixed attribute type datasets. It divides dataset into categorical attribute and numerical attribute sub datasets. Based on the expected entropy as a similarity measure, it evaluates the average entropy between different numbers of clusters and then extracts candidate cluster numbers with the results of evaluating the difference. After pre-clustering, the balance of clustering is analyzed and used to determine weight values of each attribute type. Finally, clustering process is performed with the extracted candidate cluster numbers and weight values.

Our experimental results show 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.

As future work, we will research subspace clustering algorithms for large scale dataset with mixed attribute types, investigate feature selection techniques to detect correlation between different type attributes, and investigate other alternative weighting scheme algorithms to improve the proposed framework.

6. REFERENCES

[1] Yosr Naija, Salem Chakhar, Kaouther Blibech, Riadh Robbana, "Extension of Partitional Clustering Methods for Handling Mixed Data", ICDMW, pp. 257-266 IEEE Intl. Conf on Data Mining Workshops, 2008.

[2] Z. He, X. Xu, S. Deng, "Squeezer: An Efficient Algorithm for Clustering Categorical Data," Journal of Computer Science and Technology, vol. 17, no. 5, pp. 611-625, 2002.

[3] Sudipto Guha, Rajeev Rastogi, Kyuseok Shim, "ROCK: A Robust Clustering Algorithm for Categorical Attributes", Int. Conf. Data Engineering, pp.512-521, Sydney, Australia, Mar 1999.

[4] Venkatesh Ganti, Johannes Gehrke, Raghu Ramakrishnan, "CACTUS- Clustering Categorical Data Using Summaries", Int. Conf. Knowledge Discovery and Data Mining, pp. 73-83, 1999.

[5] Amir Ahmad, Lipika Dey, "A k-mean clustering algorithm for mixed numeric and categorical data", Data & Engineering, Vol 63, Issue 2, pp. 503-527, 2007

[6] Ming-Yi Shih, Jar-Wen Jheng, and Lien-Fu Lai., "A Two-Step Method for Clustering Mixed Categroical and Numeric Data", *Tamkang Journal of Science and Engineering, Vol. 13, No. 1, pp. 11_19 (2010)*

[7] C.E. Shannon, "A mathematical theory of communication," Bell System Technical Journal, 1948.

[8] Tao Li, Sheng Ma, Mitsunori Ogihara, "Entropy-Based Criterion in Categorical Clustering", Proceeding of the twenty-first international conference on Machine Learning, pp. 68, Canada, 2004

[9]Jamil Al-Shaqsi and Wenjia Wang "A Clustering Ensemble Method for Clustering Mixed Data", IEEE, 978-1-4244-8126-2/10, 2010

[10] M. V. Jagannatha Reddy, Dr.B.Kavitha "Efficient Ensemble Algorithm for Mixed Numeric and Categorical Data", IEEE , 978-1-4244-5967-4/10, 2010

[11] Frank Lin, William W. Cohen "A General and Scalable Approach to Mixed Membership Clustering", IEEE 12th International Conference on Data Mining, 2012

[12] K. Kalaivani, A. P. V. Raghavendra "Efficiency Based Categorical Data Clustering", IEEE, 978-1-4673-1344-5/12, 2012

[13] Takashi Furukawa, Shin-ichi Ohnishi, Takahiro Yamanoi "A study on a fuzzy clustering for mixed numerical and categorical incomplete data", IEEE, 2013

[14] Paul Blomstedt, Jing Tang, Jie Xiong, Christian Granlund, and Jukka Corander "A Bayesian Predictive Model for Clustering Data of Mixed Discrete and Continuous Type", IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 37, NO. 3, MARCH 2015

[15] Han Li, Chun Li, Jie Hu, Xiaodan Fan "A Resampling based Clustering Algorithm for Replicated Gene Expression Data", IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, VOL. X, NO. X, FEBRUARY 2015

[16] Dao Lam, Member, IEEE, Mingzhen Wei and Donald Wunsch "Clustering Data of Mixed Categorical and Numerical Type with Unsupervised Feature Learning", IEEE, 2015

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