OPEN ACCESS

International Journal Of Modern Engineering Research (IJMER)

MK-Prototypes: A Novel Algorithm for Clustering Mixed Type Data

N. Aparna[1], M. Kalaiarasu[2]

[1, 2](Department of Computer Science, Department of Information Technology, Sri Ramakrishna Engineering College, Coimbatore)

Abstract: *Clustering mixed type data is one of the major research topics in the area of data mining. In this paper, a new algorithm for clustering mixed type data is proposed where the concept of distribution centroid is used to represent the prototype of categorical variables in a cluster which is then combined with the mean to represent the prototype of clusters with mixed type variables. In the method, data is observed from different views and the variables are grouped into different views. Those instances that can be viewed differently from different viewpoints can be defined as multiview data. During clustering process the differences among views are ignored in usual cases. Here, both views and variables weights are computed simultaneously. The view weight is used to determine the closeness or density of view and variable weight is used to identify the significance of each variable. With the intention of determining the cluster of objects both these weights are used in the distance function. In the proposed method, enhancement to the k-prototypes is done so that it automatically computes both view and variable weights. The proposed algorithm MK-Prototypes algorithm is compared with two other clustering algorithms.*

Keywords: *clustering, mixed data, multiview, variable weighting, view weighting, k-prototypes.*

I. Introduction

Clustering is a fundamental technique of unsupervised learning in machine learning and statistics. It is generally used to find groups of similar items in a set of unlabeled data. The aim of clustering is to divide a set of data objects into clusters so that those data objects that belongs to the same cluster are more similar to each other than those in other clusters [1-4]. In real world, datasets usually contain both numeric and categorical variables [5,6]. However, most existing clustering algorithms assume all variables are either numeric or categorical , examples of which include the k-means [7], k-modes [8], fuzzy k-modes [9] algorithms. Here, the data is observed from multiple outlooks and in multiple types of dimensions. For example, in a student data set, variables can be divided into personal information view showing the information about the student's personal information, the academic view describing the student's academic performance and the extra-curricular view which gives the extra-curricular activities and achievements made by the student.

Traditional methods take multiple views as a set of flat variables and do not take into account the differences among various views [10], [11], [12]. In the case of multiview clustering, it takes the information from multiple views and also considers the variations among different views which produces a more precise and efficient partitioning of data.

In this paper, a new algorithm Multi-viewpoint K-prototypes (MK-Prototypes) for clustering mixed type data is proposed. It is an enhancement to the usual k-prototypes algorithm. In order to differentiate the effects of different views and different variables in clustering, the view weights and individual variables are applied to the distance function. Here while computing the view weights, the complete set of variables are considered and while calculating the weights of variables in a view, only a part of the data that includes the variables in the view is considered. Thus, the view weights show the significance of views in the complete data and the variables weights in a view shows the significance of variables in a view alone.

II. Related Works

Till date, there exist a number of algorithms and methods to directly deal with mixed type data. In [13], Cen Li and Gautam Biswas proposed an algorithm, Similarity-based agglomerative clustering(SBAC) that works well for data with mixed attributes. It adopts a similarity measure proposed by Goodall [14] for biological taxonomy. In this method, while computing the similarity, higher weight is assigned to infrequent attribute value matches. It does not make any suppositions on the underlying features of the attribute values. An agglomerative algorithm is used to generate a dendrogram and a simple distinctness heuristic is used to extract a partition of the data.Hsu and Chen proposed CAVE [15], a clustering algorithm based on the Variance and Entropy for clustering mixed data. It builds a distance hierarchy for every categorical attributes which needs domain expertise.Hsu et al.[16] proposed an extension to the self-organizing map to analyze mixed data where the distance hierarchy is automatically constructed by using the values of class attributes.

In [17] Chatziz propsed KL-FCM-GM algorithm in which data derived from the clusters are in the Guassian form and is designed for the Guass-Multinomial distributed data.

Huang presented a k-prototypes algorithm [18] where k-means is integrated with k-modes to partition mixed data. Bezdek et al. considered the fuzzy nature of the objects in his work the fuzzy k-prototypes[19] and Zheng et al. proposed [20] an evolutionary type k-prototypes algorithm by introducing an evolutionary algorithm framework.

III. Proposed System

The motivation for the proposed system is on one hand to provide a better representation for the categorical variable part in a mixed data since the numerical variables can be well represented using the mean concept itself. On the other hand it considers the importance of view and variables weights in the process of clustering. The concept of distribution centroid represents the cluster centroid for the categorical variable part. Huang's strategy of evaluation is used for the computation of both view weights and variable weights.

A. The distribution centroid

The idea of distribution centroid for a better representation of categorical variables is stimulated from fuzzy centroid proposed by Kim et al.[21]. It makes use of a fuzzy scenario to represent the cluster centers for the categorical variable part.

For Dom(V_j)={$\{v_i^1, v_i^2, v_i^3, \dots v_i^t\}$, the distribution centroid of a cluster o, denoted as $C_o^{'}$, is represented as follows

$$C_o^{'} = \{c_{o1}^{'}, c_{o2}^{'}, \dots, c_{oj}^{'}, \dots c_{om}^{'}\} \tag{1}$$

where

$$c_{oj}^{'} = \{\{b_j^1, w_{oj}^1\}, \{b_j^2, w_{oj}^2\}, \dots \{b_j^k, w_{oj}^k\}, \dots \{b_j^t, w_{oj}^t\}\} \tag{2}$$

.

In the above equation

$$w_{oj}^k = \sum_{i=1}^{n} \mu(x_{ij}) \tag{3}$$

where

$$\mu(x_{ij}) = \begin{cases} \frac{u_{io}}{\sum_{i=1}^{n} u_{io}} & \text{if } x_{ij} = b_j^k \\ o & \text{if } x_{ij} \neq b_j^k \end{cases} \tag{4}$$

Here, u_{io} is assigned the value 1, if the data object x_i belongs to cluster o and as 0, if the data object x_i do not belong to cluster o

From the above mentioned equations it is clear that the computation of distribution centroid considers the number of times each categorical value repeat in a cluster. Thus to denote the center of a cluster it takes into account the distribution features of categorical variables

B. Weight calculation using Huang's approach

Weight of a variable identifies the effect of that variable in clustering process. In 2005, Huang et al. proposed an approach to calculate the weight of variable [22]. According to their method, the weight is computed by minimizing the value of objective function.

The standard for assigning weight of variable is to allocate a larger value to a variable that has a smaller sum of the within cluster distances (WCD), and vice versa. This principle is given by

$$w_j \propto \frac{1}{D_j} \tag{5}$$

where w_j is the significance of the variable j, ∝ is the mathematical symbol denoting direct proportionality, and D_j is the sum of the within cluster distances for this variable.

C. Multiview concept

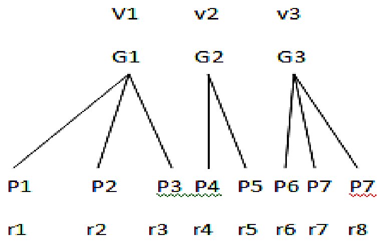

FIGURE 1 : Multiview concept

In 2013, Chen Et Al [23] proposed Tw-K-Means where the concept of multiview data was introduced. The above figure illustrates the multiview concept. During the process of clustering, the differences among different views are not considered. In the process of multiview clustering, in addition to variable weights, the variables are grouped according to their characteristic properties. Each group is termed as a view and a weight is assigned to each view. The view weight is assigned according to Huang's approach.

D. The proposed algorithm

The proposed algorithm, MK-prototypes put together the concepts in section 3.1, section 3.2, section 3.3. The figure 2 describes the steps involved in the algorithm:
Steps in the proposed algorithm:

1. Compute the distribution centroid to represent the categorical variable centroid
2. Compute the mean for the numerical variables
3. Integrate the distribution centroid and mean to represent the prototype for the mixed data
4. Compute the view weights and variable weights.
5. Measure the similarity between the data objects and the prototypes
6. Assign the data object to that prototype to which the considered data object is the closest
7. Repeat steps 1-6 until an effective clustering result is obtained.

E. The optimization model

The clustering process to partition the dataset X into k clusters that considers both view weights and variable weights is represented according to the framework of [23] as a minimization of the following objective function.

$$P(U,Z,R,V) = \sum_{o=1}^{k}\sum_{i=1}^{n}\sum_{t=1}^{Q}\sum_{s\in G_t} u_{i,o} v_t r_s d(x_{i,s}, z_{o,s}) \qquad (6)$$

subject to $\sum_{o=1}^{k} u_{i,o} = 1$, $u_{i,l} \in \{0,1\}$, $1 \le i \le n$

$$\sum_{i=1}^{Q} v_t = 1, \quad 0 \le v_t \le 1, \quad 0 \le r_j \le 1, \quad 1 \le t \le Q, \quad \sum_{j\in G_t} r_j = 1$$

where U is an n x k partition matrix whose elements $u_{i,o}$ are binary where $u_{i,o} = 1$ indicates that object i is allocated to cluster o.$Z = \{Z_1, Z_2, \ldots. Z_k\}$ is a set of k vectors on behalf of the centers of the k clusters.$V = \{V_1, V_2, \ldots. V_Q\}$ are Q weights for Q views. $R = \{r_1, r_2 \ldots r_s\}$ are s weights for s variables.$d(x_{i,s}, z_{o,s})$ is a distance or dissimilarity measure on the s^{th} variable between the i^{th} object and the center of the o^{th} cluster.

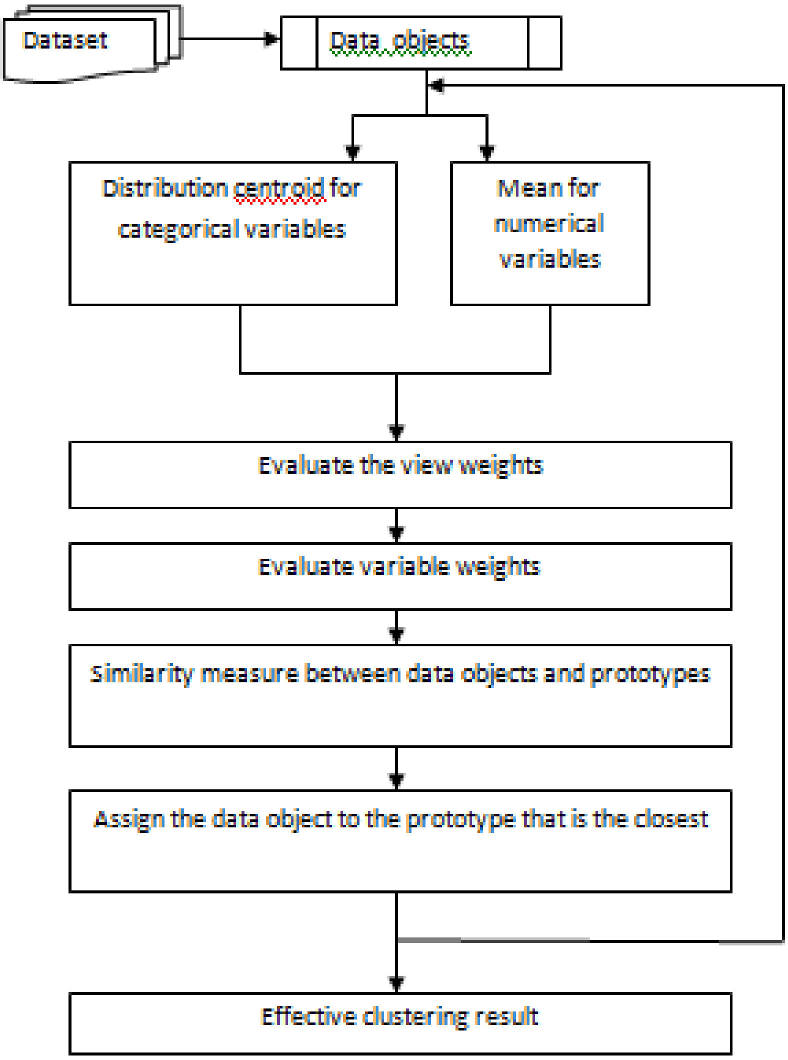
Dataset
Data objects
Distribution centroid for categorical variables
Mean for numerical variables
Evaluate the view weights
Evaluate variable weights
Similarity measure between data objects and prototypes
Assign the data object to the prototype that is the closest
Effective clustering result

FIGURE 2. Flowchart for the proposed algorithm

In order to minimize the equation, the problem is divided into four sub-problems:

1. Sub-problem 1: Fix Z=Z^,R=R^ and V=V^ and solve the reduced problem P(U,Z^,R^,V^).
2. Sub-problem 2: Fix U=U^, R=R^ and V=V^ and solve the reduced problem P(U^,Z,R^,V^).
3. Sub-problem 3: Fix Z=Z^, U=U^ and V=V^ and solve the reduced problem P(U^,Z^,R,V^).
4. Sub-problem 4: Fix Z=Z^, R=R^ and U=U^ and solve the reduced problem P(U^,Z^,R^,V).

The sub-problem 1 is solved by:

$$u_{i,o} = 1 \tag{7}$$

if

$$\sum_{s=1}^{m} v_t r_s \mathrm{d}(x_{i,o}, z_{i,0}) \leq \sum_{s=1}^{m} v_t r_s \,\mathrm{d}(x_{i,o}, z_{e,0}) \tag{8}$$

where 1≤ $e \leq k$

$$u_{i,o} = 0 \; where \; e \neq o$$

The sub-problem 2 is solved for the numeric variable by

$$z_{o,s} = \frac{\sum_{i=1}^{n} u_{i,o} x_{i,s}}{\sum_{i=1}^{n} u_{i,o}} \tag{9}$$

and for the categorical variables by $z_{o,s} = c'_{i,s}$ which is already defined.

$d(x_{i,s}, z_{o,s}) = |x_{i,s} - z_{o,s}|$ if the sth variable is a numeric variable .

$d(x_{i,s}, z_{o,s}) = \varphi(x_{i,s}, z_{o,s})$ if the sth variable is a categorical variable .

where $\varphi(x_{i,s}, z_{o,s}) = \sum_{k=1}^{t} \delta(x_{i,s}, b_j^k)$ and $\delta(x_{i,s}, b_j^k)$ is 0 if $x_{i,s} \neq b_j^k$ and $w_{o,j}^k$ if $x_{i,s} = b_j^k$.

The solution to the sub-problem 3 is as followed:

Let Z=Z^, U=U^ and V=V^ be fixed . Then the reduced problem P(U^,Z^,R,V^) is minimized if

$$r_s = \frac{1}{\sum_{h \epsilon G_t} \left[\frac{D_s}{D_h}\right]^{\frac{1}{\gamma}}} \tag{10}$$

where

$$D_s = \sum_{o=1}^{k} \sum_{i=1}^{n} u'_{i,o}\, w'_t\, d(x_{i,s}, z'_{o,s}) \tag{11}$$

Sub-problem 4 is solved as follows

$$w_t = \frac{1}{\sum_{t=1}^{h} \left[\frac{F_s}{F_t}\right]^{\frac{1}{\mu}}} \tag{12}$$

where

$$F_s = \sum_{o=1}^{k} \sum_{i=1}^{n} \sum_{s \epsilon G_t} u'_{i,o} r'_s d(x_{i,s}, z'_{o,s}) \tag{13}$$

Having presented the detailed computations required for calculating the important variables, the proposed algorithm
MK-Prototypes can be described as given below:

1. Choose the number of iterations, number of clusters k, value of μ and γ, randomly choose k distinct data objects and convert them into initial prototypes and initialize the view weights and variable weights.
2. Fix Z', R', V' as Z^t, R^t, V^t respectively and minimize the problem P(U, Z', R', V') to obtain U^{t+1}.
3. Fix U', R', V' as U^t, R^t, V^t respectively and minimize the problem P(U', Z, R', V') to obtain Z^{t+1}.
4. Fix U', Z', V' as U^t, Z^t, V^t respectively and minimize the problem P(U', Z', R, V') to obtain R^{t+1}.
5. Fix U', Z', R', V as U^t, Z^t, R^t respectively and minimize the problem P(U', Z', R', V) to obtain V^{t+1}.
6. If there is no improvement in P or if the maximum iterations is reached, then stop. Else increment t by 1 , decrement number of iterations by 1 and go to Step 2.

IV. Experiments on Performance Of Mk-Prototypes Algorithm

In order to measure the performance level of the proposed algorithm, it is used to cluster a real-world dataset Heart (disease). The dataset is taken from UCI Machine Learning Repository.
The proposed algorithm is compared with k-prototypes and SBAC algorithm. They are well known for clustering mixed type data. In this paper, the clustering accuracy is measured using one of the most commonly used criteria. The clustering accuracy r is given by

$$r = \frac{\sum_{i=1}^{k} a_i}{n} \tag{14}$$

where $a_i$is the number of data objects that occur in both the ith cluster and its corresponding true class and n is the number of data objects in a data set.
Higher the value of r , the higher the clustering accuracy . A perfect clustering gives a value of r=1.0.

A. Dataset description

The Heart disease data set is a mixed dataset. It contains 303 patient instances. The actual data set contains 76 variables out of which 14 are considered usually. In the proposed algorithm, in order to define three views 19 out of 76 variables are considered here. It consists of seven numeric variables and twelve categorical variables.

These 19 variables can be naturally divided into 3 views.

1. Personal data view: It includes those variables which describes a patient's personal data.
2. Historical data view: It includes those variables which describes a patient's historical data like the habits.
3. Test output view: It includes all those variables which describes the results of various tests conducted for the patient.

Here, G_1, G_2, G_3 represents the three views personal, historical, test output respectively.

B. Results and analysis

Below are the graphical representations of the clustering results. Fig 3 shows the variation in variable weights for varying μ values and fixed γ values. Fig 4 shows the variation in view weights for varying μ values and fixed γ values.

From Table 1, it is observed that as μ increased, the variance of V decreased rapidly. This result can be explained from equation (10) as μ increases, V becomes flatter. The graphical representation of the Table 1 has been shown below.

Table 1: Variable weights vs γ value For fixed μ value

μ \ γ	1	4	12
10	0.01	0	0
15	0	0.01	0
20	0.7	0	0.05
25	0.03	0.4	0.1
30	0	0.1	0.02
35	0.02	0	0

Table 2 shows that as γ increased, the variance of view weights decreased rapidly. This result can be explained from equation (11) as γ increases, W becomes flatter. The graphical representation has been shown below.

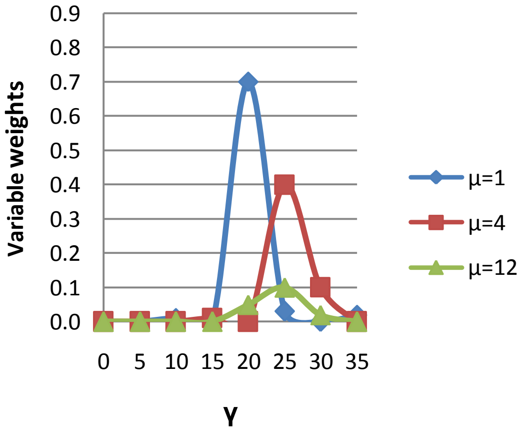

Fig 3: Variable weights vs γ value for fixed μ value

Table 2: View weights vs μ value for fixed γ value

γ / μ	1	4	12
10	0.05	0.075	0.01
20	0.075	0.14	0.015
30	0.095	0.05	0.04
40	0.16	0.06	0.01
50	0.04	0.07	0.005
60	0.05	0.04	0.02
70	0.07	0.035	0.01

From above analysis, it can be summarized that the following method can be used to control two types of weight distributions in MK-Prototypes algorithm by setting different values of γ and μ.

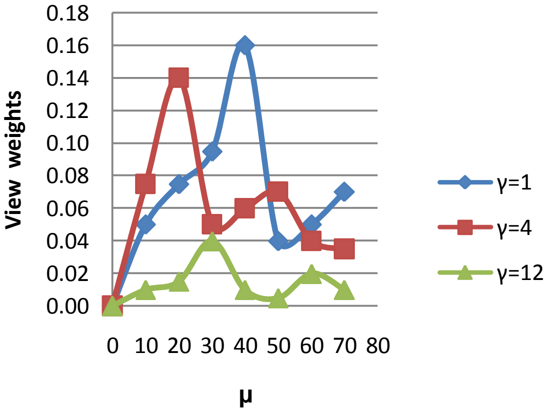

Figure 4: View weights vs μ value for fixed γ value

The experiments have been conducted for three different values of μ and γ for varying values of γ and μ respectively.

1. Large μ makes more variables contribute to the clustering while small μ makes only important variables contribute to the clustering.
2. Large γ makes more views contribute to the clustering while small γ makes only important views contribute to the clustering.

Table 3: Comparison of accuracy rates of dataset considering all views

Algorithms	Clustering accuracy %
k-prototypes	0.521
SBAC	0.747
MK-Prototypes	0.846

From the above table, it is clear that the proposed algorithm has a better clustering accuracy than the existing k-prototypes and SBAC.

V. Conclusion

Mixed type data are encountered everywhere in the real world. In this paper, a new algorithm, Multiview point based clustering algorithm for mixed type data has been proposed. When compared with the existing algorithms the proposed algorithm has many significant contributions. The proposed algorithm encapsulates the characteristics of clusters with mixed type variables more efficiently since it includes the distribution information of both numeric and categorical variables.

It also takes into account the importance of various variables and views during the process of clustering by using Huang's approach and a new dissimilarity measure.

It can compute weights for views and individual variables simultaneously in the clustering process. With the two types of weights, dense views and significant variables can be identified and effect of low-quality views and noise variables can be reduced.

Because of these contributions the proposed algorithm obtains higher clustering accuracy, which has been validated by experimental results.

REFERENCES

[1] Z.X. Huang, Extensions to the k means algorithm for clustering large datasets with categorical values, Data Min. Knowl. Discovery2 (3) (1998) 283–304.

[2] A.K.Jain, R.C.Dubes, Algorithms for Clustering Data, Prentice-Hall, New Jersey, 1988.

[3] A.K.Jain, M.N.Murty, P.J.Flynn, Data clustering: a survey, ACM Comput. Surv. 31 (3) (1999) 264–323.

[4] J.W.Han, M.Kamber, Data Mining Concepts and Techniques, Morgan Kaufmann, SanFrancisco,2001.

[5] C.Hsu,Y.P.Huang, Incremental clustering of mixed data based on distance hierarchy, Expert Syst. Appl. 35 (3) (2008) 1177–1185. [6] C.Hsu, S.Lin, W.Tai, Apply extended self-organizing map to cluster and classify mixed-type data, Neurocomputing 74 (18) (2011) 3832–3842.

[7] S.Lloyd, Least squares quantization in PCM, IEEE Trans. Inf. Theory 28 (2) (1982) 129–137.

[8] Z.X.Huang, Extensions to the k-meansalgorithm for clustering large datasets with categorical values, Data Min. Knowl. Discovery 2 (3) (1998) 283–304. [9]Z.X.Huang, M.K.Ng, A fuzzy k-modes algorithm for clustering categorical data, IEEE Trans. Fuzzy Syst. 7 (4) (1999) 446–452.

[10] J. Mui and K. Fu, "Automated Classification Of Nucleated Blood Cells Using A Binary Tree Classifier," IEEE Trans. Pattern Analysis And Machine Intelligence, Vol. 2, No. 5, Pp. 429-443, May 1980.

[11] J. Wang, H. Zeng, Z. Chen, H. Lu, L. Tao, And W. Ma, "Recom: Reinforcement Clustering Of Multitype Interrelated Data Objects,"Proc. 26th Ann. Int'l ACM SIGIR Conf. Research And Development In Informaion Retrieval, Pp. 274-281, 2003.

[12] S. Bickel And T. Scheffer, "Multi-View Clustering," Proc. IEEE Fourth Int'l Conf. Data Mining, Pp. 19-26, 2004.

[13] C.Li, G.Biswas, Unsupervised Learning with Mixed Numeric And Nominal Data, IEEE Trans. Knowl. Data Eng.14 (4) (2002) 673–690.

[14] D.W.Goodall, A New Similarity Index Based On Probability, Biometrics 22 (4) (1966) 882–907.

[15] C.C.Hsu, Y.C.Chen, Mining Of Mixed Data With Application To Catalog Marketing, Expert Syst. Appl. 32 (1) (2007) 12–27.

[16] C. Hsu, S.Lin, W.Tai, Apply Extended Self-Organizing Map To Cluster And Classify Mixed-Type Data, Neurocomputing 74 (18) (2011) 3832–3842.

[17] S.P.Chatzis, A Fuzzy C-Means-Type Algorithm For Clustering Of Data With Mixed Numeric And Categorical Attributes Employing A Probabilistic Dissimilarity Functional, Expert Syst. Appl. 38 (7) (2011) 8684–8689.

[18] Z.X.Huang, Clustering Large Datasets with Mixed Numeric and Categorical Values, In: Proceedings Of The First Pacific-Asia Knowledge Discovery And Data Mining Conference, 1997, Pp.21–34.

[19] J.C.Bezdek, J.Keller, R.Krisnapuram, Fuzzy Models And Algorithms For Pattern Recognition And Image Processing, Kluwer Academy Publishers, Boston, 1999. [25].

[20] Z.Zheng, M.G.Gong, J.J.Ma, L.C.Jiao, Unsupervised Evolutionary Clustering Algorithm For Mixed Type Data, In : Proceedings Of The IEEE Congresson Evolutionary Computation (CEC), 2010, Pp.1–8.

[21] W.Kim, K.H.Lee, D.Lee, Fuzzy Clustering Of Categorical Data Using Fuzzy Centroid, Pattern Recognition Lett.25 (11) (2004) 1263–1271.

[22] Z.X.Huang, M.K.Ng, H.Q.Rong, Z.C.Li, Automated Variable Weighting In K-Means Type Clustering, IEEE Trans. Pattern Anal. Mach. Intell.27 (5) (2005) 657–668.

[23] Xiaojun Chen, Xiaofei Xu, Joshua Zhexue Huang, And Yunming Ye, Tw-K-Means: Automated Two-Level Variable Weighting Clustering Algorithm For Multiview Data, IEEE Transactions On Knowledge And Data Engineering, Vol. 25, No. 4, April 2013, pp 932-945