International Journal of Scientific Engineering and Research (IJSER)
www.ijser.in
ISSN (Online): 2347-3878
Volume 3 Issue 2, February 2015

A Novel Multiple View Clustering Using TW-k-Means Algorithm

Divya. P[1], Ranjani. C[2]

[1]PG Scholar, Department of CSE (PG), Sri Ramakrishna Engineering College, Coimbatore, India

[2]Assistant Professor, Department of CSE (PG), Sri Ramakrishna Engineering College, Coimbatore, India

Abstract: ***Multiple view clustering has become an active research area in the field of data mining. However, still there is a scope for improvement in the performance of clustering. This paper proposes TW-k-means algorithm which is an automated two-level variable weighting clustering algorithm for multiview data, which can simultaneously compute weights for views and individual variables. In this algorithm, a view weight is assigned to each view to identify the compactness of the view and a variable weight is also assigned to each variable in the view to identify the importance of the variable. Both view weights and variable weights are used in the distance function to determine the clusters of objects. In the new algorithm, two additional steps are added to the iterative k-means clustering process to automatically compute the view weights and the variable weights. Experimental results show that the TW-k-means algorithm outperforms the other existing clustering algorithms in effective manner.***

Keywords: Clustering, TW-k-means algorithm, multiview data, variable weighting, view reduction

1. Introduction

Cloud Clustering is a fundamental technique of unsupervised learning in machine learning and statistics. It is generally use to find groups of similar objects in a set of unlabelled data. Multiview data are instances that have multiple views from different feature spaces. Here the data is observed from multiple dimensions. For example, Web pages can be represented with three views: a term vector view whose elements correspond to the occurrence of certain words in the web page text, a hyperlink graph view that shows other web pages which each web page points to, and a term vector view for the words in the anchor text.

Traditional methods take multiple views as a set of flat variables and do not consider the difference among various view [1], [2], [3]. In the case of multiview clustering, it takes the information from the multiple views and also considers the variances among different views which produces a more precise and efficient partitioning of data.

Variable weighting clustering calculates a weight for each variable [4], [5], [6]. This weight is used to determine the variables that are important and those that are unimportant. In traditional variable weighting clustering, there are various methods that calculate and consider the weights for individual variable but do not consider the discrepancies in views in the case of multiview data. Thus, those methods do not provide an efficient or accurate clustering. In multiview data, difference among views and also the importance of each variable in the view must be considered.

In this paper, a new automated view reduction algorithm for multiview data has been proposed. It is an enhancement to the usual k-means 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 weight of a variable 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 variable weights in a view shows the significance of variables in a view alone.

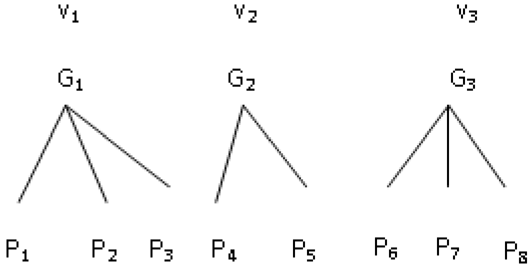

Figure 1: Multiview Concept

Fig.1 illustrates the multiview concept. Once the view weights are calculated, the views with weights below than the user given threshold is eliminated. Here a formula from an optimization model was derived which is used for calculating both view weights and variable weights. This is an extension to the basic k-means clustering process. The rest of the paper is organized as follows. Section 2 provides a survey of the related work on variable weighting. Section 3 gives the problem statement. Section 4 provides the detailed description of the proposed algorithm. Section 5 provides the experiments and results of performance of the proposed method. Sections 6 give the conclusion.

2. Literature Survey

There are two approaches exists in multiview learning. They are centralized and distributed. In centralized algorithms, more than one representation is taken into account in order to extract the data patterns. In distributed approach it first learns the hidden patterns in each representation separately and then from multiple patterns the optimal ones are learned [7].

In [9], Huang et al proposed a new method called W-k-means clustering algorithm that can calculate weights for each variables automatically. The variable weights are calculated based on the importance of the variables in clustering. These weights are used for selecting the variables in data mining applications where large data are involved. The optimal weights are found when algorithm converges. The computed variable weights are inversely proportional to the sum of the within cluster variances of the variable.

Jing and Huang [10] have proposed a new k-means type algorithm for clustering high dimensional objects in subspaces. The concept of weight entropy is used to assign weights to each dimension in the subset and different dimensions are considered to make different contributions to the identification of objects to the cluster. The within cluster scattering is brought to be the least and negative weight entropy is maximized at the same time so that more dimensions donate to the identification of a cluster. This can avoid problem of identifying clusters by a few dimensions with sparse data.

SYNCLUS [11] is the first clustering algorithm that used the concept of both view weights and variable weights. It is done in two steps. In the first step all the variables are assigned some weights randomly which is then partitioned into k clusters using k-means. The second step computes a new set of optimal weights by optimizing a weighted mean-square. These two process continues until the process converges to an optimal set of variable weights. This method computes the variable weights automatically. But the view weights are given by the users.

3. Problem Definition

The problem of finding clusters in variable groups and individual variables can be stated as follows. Let X={X1,X2,X3,……,Xn} be a set of n objects represented by the set P of s variables. Assume P is divided into Q views

Where Gt Gs= for s t and $\bigcup_{t=1}^{Q} G$t=P. Let V={V1, V2, V3,…..Vn} be a set of Q view weights, where Vt indicates the weight that is assigned to the ith view and $\sum_{t=1}^{Q} V$t=1. Let R={Rj} be a set of a variable weights, where Rj indicates the weight that is assigned to the ith variable and $\sum_{j \in G_t} R$j=1 (1≤t≤Q) , $\sum_{j=1}^{s} R$j=Q. Assume that X contain k clusters from G. Along with that we have to find the important views from the view weight matrix V=[Vt]Q and identify the important variables from the variable weight matrix R=[Rj]s.

In the new method, the two types of weights are used for different aims. In W-k-means [9], the variable weights are used to recognize the subset of variables in which clustering structure occur and eliminate the effect of insignificant (or noisy) variables. In the new method, we assume that the cluster structures occur in variable groups G and use variable weights Rt to identify the subset of variables in each variable group Gt. Meanwhile, the variable group weights V are used to identify the importance of cluster structures among these variable cluster structures. If the variable group contains insignificant cluster structures, a small variable group weight should be assigned to eliminate the effect of such variable groups. On the contrary, if the variable group contains significant cluster structures, a big variable group weight should be assigned so as to enhance the effect of such variable group. The group with weights less than a threshold value is eliminated and then the weight values are recalculated. These variable groups are termed as views.
View weights can be easily distinguished than the variable weights as the number of views is much smaller than the number of variables. View weights are determined in the view level alone while the variable weights within a view will be determined.

4. Methodology

In this paper, a new algorithm called TW-K-Means algorithm is proposed which are discussed below

A. *TW-k-Means Clustering Algorithm*

The proposed TW-k-Means algorithm involves the following steps.

- Enter the threshold view weight value
- Initialize i=0, k centroid view weights and variable weights
- Calculate the distance of each object to each centroid
- Calculate the new values of view weight and variable weights based on the above calculated distances
- For the first iteration, eliminate the view with weigh below threshold view weight else go to next step
- Assign the object to that cluster from which it has minimum distance
- Recalculate the centroid values
- If the centroid value changes then go to step 3 else stop the process

B. *The Optimization Model*

The clustering process to partition the X into k clusters that considers both view weights and variable weights are represented as a minimization of the following objective function.

$$P\ (U,\ C,\ R,\ V) = \sum_{o=1}^{k} \sum_{i=1}^{n} \sum_{t=1}^{Q} \sum_{j \in G_t} u_{i,o} v_t\, r_j\, d(x_{i,o}\,, c_{i,o}) + \mu \sum_{j=1}^{k} r_j \log(r_j) + \gamma \sum_{t=1}^{Q} v_t \log(v_t) \quad (1)$$

Subject to $\sum_{o=1}^{k} u_{i,o}$=1, $u_{i,l} \in \{0,1\}$,1≤ $i \le n$
$\sum_{i=1}^{Q} v_t$=1, 0≤ $v_t \le 1$, 0≤ $r_j \le 1, 1 \le t \le Q, \sum_{j \in G_t} r_j = 1$

Where U is an n×k partition matrix whose elements $u_{i,o} = 1$ indicates that object i is allocated to cluster o. C={C_1, C_2, …….C_n} is a set of k vectors representing the centers of k clusters. V={V_1, V_2,....V_Q} are Q weights for Q views. R={r_1, r_2,…r_s} are s weights for s variables. γ>0, $\mu > 0$ are two given parameters $(x_{i,j}, c_{o,j})$ is a distance measure on the jth variable between the ith object and centre of Oth cluster.If the variable is numerical then d$(x_{i,j}, c_{o,j})$ = $(x_{i,j}, c_{o,j})^2$. If the variable is categorical then d$(x_{i,j}, c_{o,j})$= 0 $x_{i,j}$ =$c_{o,j}$ and 1 if $x_{i,j} \neq c_{o,j}$.
The first term in the objective function is the sum of the within cluster dispersions. The next two terms are negative weight entropies. The two positive parameters γ, μ are used to control the strength of motivations for clustering on more views and variables.

1. Problem P1: Fix C=C^, R=R^ and V=V^ and solve the reduced problem P (U,C^,R^,V^).
2. Problem P2: Fix U=U^, R=R^ and V=V^ and solve the reduced problem P (U^,C,R^,V^).
3. Problem P2: Fix U=U^, C=C^ and V=V^ and solve the reduced problem P (U^,C^,R,V^).

4. Problem P2: Fix U=U^, C=C^ and R=R^ and solve the reduced problem P (U^,C^,R^,V).

To solve problem 1 consider

$u_{i,o}$=1 (2)

If $D_0 \leq D_l$ for $1 \leq l \leq k$ where

$$D_t = \sum_{i=1}^{Q} \sum_{j \in G_t} v_t \, r_j \, d(x_{i,j}, c_{l,j})$$

and $u_{i,l}$ = 0 for s ≠ l

Problem 2 is solved by

$$c_{o,j} = \frac{\sum_{i=1}^{n} u_{i,o}}{\sum_{i=1}^{n} u_{i,o}} x_{i,j} \quad (3)$$

For $1 \leq o \leq k$.
If the variable is categorical then
$c_{o,j} = a_j^r$ where a_j^r is the mode of the variable values of the jth variable in cluster c.

The solution to problem 3 is given by considering the following.
Let C=C^, U=U^ and V=V^ be fixed.

P(U^,C^,R,V^) is minimized iff

$$r_j = \frac{e^{\frac{-F^j}{\mu}}}{\sum_{h \epsilon G_t} e^{\frac{-F^h}{\mu}}} \quad (4)$$

Where

$$F_j = \sum_{o=1}^{k} \sum_{j \in G_t} u'_{i,o} \, v'_t \, d(x_{i,j}, c_{o,j})$$

The solution to problem 4 is as follows. Let C=C^, U=U^ and R=R^ be fixed. Then, P(U^, C^, R^, V) is minimized iff

$$v_t = \frac{e^{\frac{-F_j}{\mu}}}{\sum_{h=1}^{Q} e^{\frac{-D_t}{\gamma}}} \quad (5)$$

where

$$D_t = \sum_{o=1}^{k} \sum_{i=1}^{n} \sum_{j \in G_t} u'_{i,o} \, r'_t \, d(x_{i,j}, c'_{o,j})$$

TW-k-Means Algorithm

The algorithm that minimizes the objective function (1) is given as follows:
Input: Cluster numbers k, threshold view weight T_v and the input parametersμ *and* γ.
Output: Finest values of U, C, V and R.
1. Choose the centers C^0 and randomly initialize V^0 and R^0
2. Let t=0
3. Calculate U, C, R and V.
4. Prioritize the views in the order of view weights calculated
5. Eliminate those views whose weights fall below $T_{v.}$
Repeat
Update Ut+1 by (2);
Update Ct+1 by (3);
Update Rt+1 by (4);
Update Vt+1 by (5);
t=t+1;
until the objective function attains its confined least value.

The input parameters γ and μ are used to control the distribution of the two types of weights V and R. We can show that the objective function (1) can be minimized with respect to v and r iff $\gamma \geq 0$.If $\gamma > 0$, according to (10), v is inversely proportional to D. The smaller $D_{t,}$ the larger v_t, the more important the corresponding views. If $\gamma = 0$, this will produce clustering result with only one important view. It may not be desirable for high-dimensional data sets.

If $\mu > 0$, according to (8), r is inversely proportional to F. The smaller F, the larger r, and more important the corresponding variable is. If $\mu = 0$ clustering results with only one important variable in a view.

5. Results and Discussion

Experimental results were discussed here and Characteristics of Real- Life Data Set are given. The Water Treatment Plant data set came from the daily measures of sensors in an urban waste water treatment plant [12]. This data set contains 527 instances and 38 features. The 38 features can be partitioned into four views.

The first 22 attributes that describes the input conditions of the plant is considered to be the first view. Attributes from 23 to 29 portrays the output demands. Third view is illustrated by the attribute 30 to 34 because it depicts the performance input values. Final view is described by the last four features. The graphical representations of the clustering results are shown below.

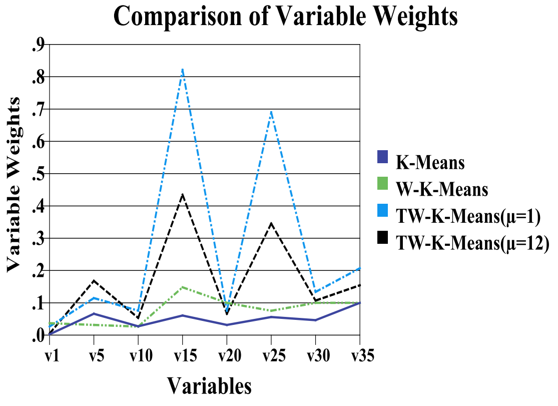

Figure 2: Comparison of variable weights

In fig 2 it is observed that as μ value gets increased, the variable weight gets decreased gradually. This result can be explained using equation (4).

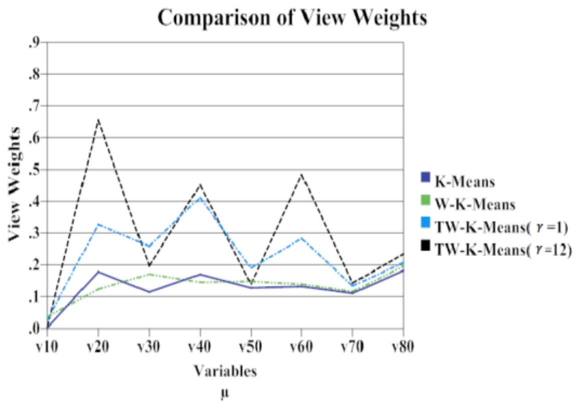

Figure 3: Comparison of view weights

In fig 3 it is observed that as γ increases the variance of view weights increased rapidly. This result can be explained using equation (5).

From above analysis, it can be summarized that the following method can be used to control two types of weight distributions in TW-k-means by setting different values of μ and γ. The experiments have been conducted for two different values of μ and γ for varying values of γ and μ respectively.

Large μ makes more variables contribute to the clustering while small μ makes only important variables contribute to the clustering.

Large γ makes more views contribute to the clustering while small γ makes only important views contribute to the clustering.

Precision

Precision is calculated as the fraction of correct objects among those that the algorithm believes belonging to the relevant class.

Precision =TP/(TP+FP)

Where, TP is the True Positive, FP is the False Positive.

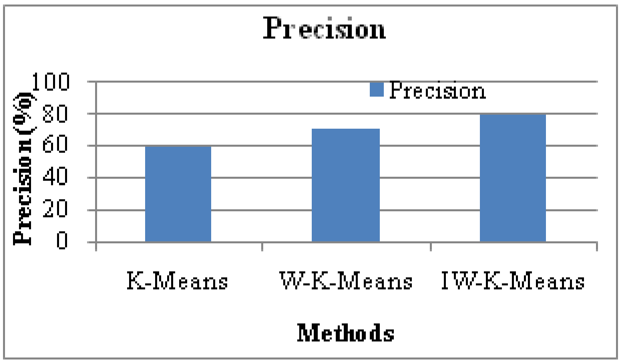

Figure 4: Precision Comparison

In Fig 4, X-axis represents the methods such as existing system i.e., k-means, W-k-means, and the proposed method TW-k-means and Y axis represents the precision rate. From the graph it is inferred that the proposed system is effective in terms of precision rate.

Recall

Recall is the fraction of actual objects that were identified.

Recall =TP/(TP+FN)

Where, FN is the False Negative.

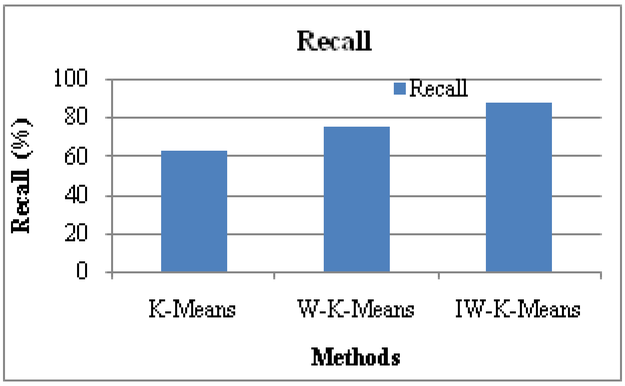

Figure 5: Recall comparison

In Fig 5, X axis represents the methods such as existing system i.e., k-means, W-k-means, and the proposed method TW-k-means and Y axis represents the precision rate. From the graph it is inferred that the proposed system is effective in terms of recall rate.

Accuracy

Accuracy is defined as the number of correctly classified objects.

$$\text{Accuracy}=100\frac{\sum_{i=1}^{k} a_i}{N}$$

Where, ai is the number of points in each cluster ci. N is the number of points in the dataset.

Table 1: Comparison of accuracy rates of dataset considering all views

Algorithm	Clustering accuracy %
k-means	0.945
W-k-means	0.957
TW-k-means	0.985

From Table 1 it is observed that the clustering accuracy of TW-K-Means is better than WK-Means and K-Means. It is observed that as the number of objects increases the accuracy of the proposed algorithm remains efficient than the other two. The graphical representation of the Table 1 is plotted below.

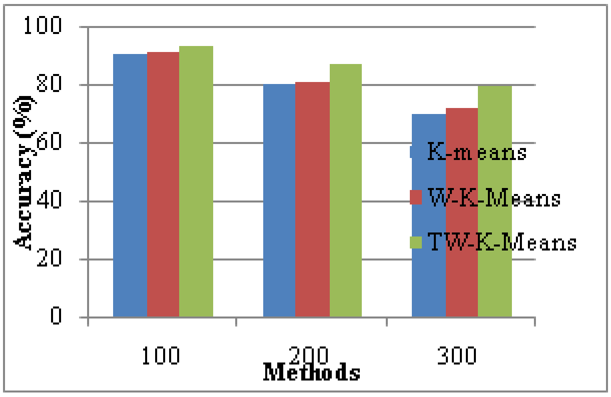

Figure 6: Accuracy Comparison

To sum up, it is possible to control two types of weight distributions in the TW-k-means algorithm by setting different values of γ and μ and TW-k-means is superior to the other two clustering algorithms in multiview data.

C. Characteristics of Real Life Data Set

The image segmentation dataset consists of 2310 objects drawn randomly from a database of seven outdoor images. The dataset contains 19 features which can be divided into two views [12]. The first view consists of Shape view. It contains nine features about the shape information of the seven images. The second view is RGB view: contains 10 features about the RGB values of the seven images.

Precision

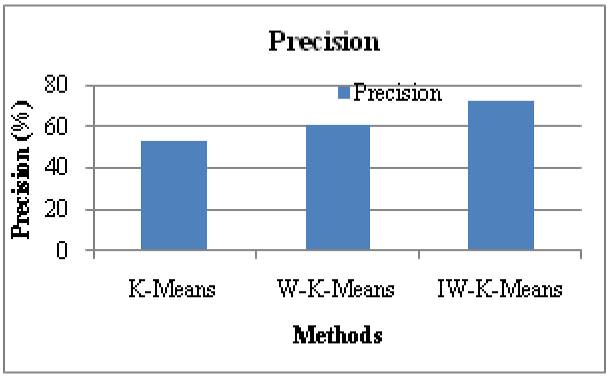

Figure 7: Precision graph

In Fig 7 it is observed that the precision rate is high for the proposed system TW-k-means algorithm.

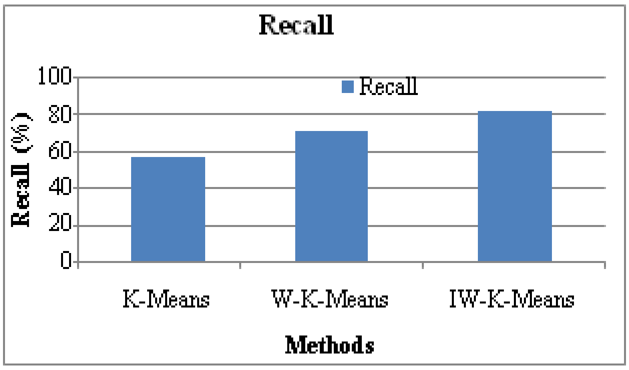

Figure 8: Recall Graph

In Fig 8 it is observed that the recall rate is high for the proposed system TW-k-means.

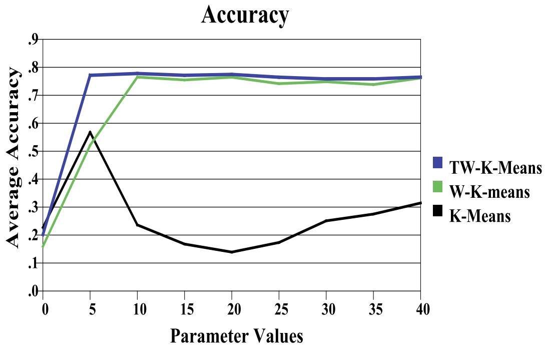

Figure 9: Accuracy comparison

6. Conclusion

In this paper, an efficient algorithm for clustering multiview data has been proposed. 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. The insignificant views are eliminated based on the input threshold weight. Therefore, this algorithm can obtain better clustering results than individual variable weighting clustering algorithms from multiview data. The performance of the TW-k-means algorithm is compared with other two clustering algorithms and the results have shown that the proposed algorithm significantly outperformed the other clustering algorithms. It also compared the effect of control parameters on the view weights and variable weights.

References

[1] S. Bickel and T. Scheffer, 2004“Multi-view Clustering,” Proc. IEEE Fourth Int’l Conf. Data Mining, pp. 19-26.

[2] B. Long, P. Yu, and Z. Zhang, 2008“A General Model for Multiple View Unsupervised Learning,” Proc. Eighth SIAM Int’l Conf. Data Mining (SDM ’08).

[3] D. Zhou and C. Burges, 2007 “Spectral Clustering and Transductive Learning with Multiple Views,” Proc. 24th Int’l Conf. Machine Learning, pp. 1159-1166.

[4] E. Fowlkes, R. Gnanadesikan, and J. Kettenring,1988 “Variable Selection in Clustering,” J. Classification, vol. 5, pp. 205-228.

[5] G. Tzortzis and C. Likas, 2010 “Multiple View Clustering Using a Weighted Combination of Exemplar-Based Mixture Models,” IEEE Trans. Neural Networks, vol. 21, no. 12, pp. 1925-1938.

[6] G. De Soete, 1986 “Optimal Variable Weighting for Ultrametric and Additive Tree Clustering,” Quality and Quantity, vol. 20, pp. 169-180.

[7] L. Jing, M. Ng, and Z. Huang, 2007 “An Entropy Weighting k-Means Algorithm for Subspace Clustering of High-Dimensional Sparse Data,” IEEE Trans. Knowledge and Data Eng., vol. 19, no. 8, pp. 1026-1041

[8] R. Gnanadesikan, J. Kettenring, and S. Tsao, 1995 “Weighting and Selection of Variables for Cluster Analysis,” J. Classification, vol. 12, pp. 113-136.

[9] V.R. de Sa, 2005 “Spectral Clustering with Two Views,” Proc. IEEE 22nd Int’l Workshop Learning with Multiple Views (ICML), pp. 20-27.

[10] Z. Deng, K. Choi, 2010 F. Chung, and S. Wang, “Enhanced Soft Subspace Clustering Integrating Within-Cluster and Between-Cluster Information,” Pattern Recognition, vol. 43, no. 3, pp. 761-781.

[11] Z. Huang, 1998“Extensions to the k-Means Algorithms for Clustering Large Data Sets with Categorical Values,” Data Ming and Knowledge Discovery, vol. 2, no. 3, pp. 283-304.

[12] A. Frank and A. Asuncion, 2010 “UCI Machine Learning Repository,”http://archive.ics.uci.edu/ml.