International Journal of Innovative Research in Science and Engineering
Vol. No.2, Issue 02, February 2016
www.ijirse.com
ISSN: 2454-9665

AN ENTROPY BASED WEIGHTING TW-K-MEANS ALGORITHM FOR MULTIVIEW DATACLUSTERING

R.Tamilselvan[1], K.Philip Vinod[2], Dr.C.Palanisamy[3]

[1]*Research Scholar, Department of IT, PSG College of Technology, Coimbatore, (India)*

[2]*Assistant Professor, Department of CSE, JNN Institute of Engineering, Thiruvallur, (India)*

[3]*Professor and Head , Department of IT, Bannari Amman Institute of Technology, Sathyamangalam,*

ABSTRACT

This paper proposes a new TW-k-means type clustering algorithm that can automatically calculate variable weights. A new step is introduced to the TW-k-means clustering process to iteratively update variable weights based on the current partition of data and a formula for weight calculation is proposed. In this algorithm, a variable 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. An automated two-level variable weighting clustering algorithm for multiview data, which can simultaneously compute weights for views and individual variable. We used two real-life data sets to investigate the properties of two types of weights in TW- k-means and investigated the difference between the weights of TW-k-means and the weights of the individual variable weighting method. The experiments have revealed the convergence property of the view weights in TW-k-means. 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 weight. Experimental results on both synthetic and real data have shown that the new algorithm outperformed the standard TW-k-means type algorithms in recovering clusters in data.

Index Terms: Variable weighting, Clustering, View weight, Multiview Data, TW-k-means.

I. INTRODUCTION

Clustering is a process of partitioning a set of objects into clusters such that objects in the same cluster are more similar to each other than objects in different clusters according to some defined criteria. The k-means type clustering algorithms [1], [2] are widely used in real world applications such as marketing research [11] and data mining to cluster very large data sets due to their efficiency and ability to handle numeric and categorical variables that are ubiquitous in real databases. A major problem of using the k-means type algorithms in data mining is selection of variables. The TW-k-means type algorithms cannot select variables automatically because they treat all variables equally in the clustering process. In practice, selection of variables for a clustering problem such as customer segmentation is often made based on under-standing of the business problem and data to be used. Tens or hundreds of variables are usually extracted or derived from the database in the initial selection which form a very high-dimensional space. It is well-known that an interesting clustering structure

ISSN: 2454-9665

usually occurs in subspace defined by a subset of the initially selected variables. To find the clustering structure, it is important to identify the subset of variables.

Multiew data are instances that have multiple views from different feature spaces. It is the result of integration of multiple types of measurements on observations from different perspectives and different types of measurements can be considered as different views. For example, the variables of the nucleated blood cell data [3] were divided into views of density, geometry, "color" and texture, each representing a view of particular measurements on the nucleated blood cells. In a banking customer data set, variables can be divided into a demographic view representing demographic information of customers, an account view showing the information about customer accounts, and the spending view describing customer spending behaviors. In this paper, we propose TW-k-means, a novel two-level variable weighting k-means clustering algorithm for multiview data. In the TW-k-means algorithm, to distinguish the impacts of different views and different variables in clustering, the weights of views and individual variables are introduced to the distance function. The view weights are computed from the entire variables, whereas the weights of variables in a view are computed from the subset of the data that only includes the variables in the view. Therefore, the view weights reflect the importance of the views in the entire data, while the variable weights in a view only reflect the importance of variables in the view. We present an optimization model for the TW- k-means algorithm and introduce the formulae, derived from the model, for computing both view weights and variable weights. We define the TW-k-means algorithm as an extension to the standard k-means clustering process with two additional steps to compute view weights and variable weights in each iteration.

Since the two steps do not require intensive computation, the new clustering algorithm remains efficient in clustering large high dimensional multiview data. Compared with SYNCLUS and WCMM, TW-k-means can automatically compute both view weights and individual variable weights. Moreover, it is a fast clustering algorithm which has the same computation complexity as k-means. Two sets of experiments on five real-life data sets have been conducted, one was used to investigate the properties of two types of weights in TW-k-means, the other was used to verify the performance of TW-k-means in classification. With the first experiment [4], we discuss how to control two types of weight distributions, illustrate the differences of the weights in TW-k-means and the individual variable weighting method, and demonstrate the convergence property of view weights in TW-k-means. In the second experiment, we compared TW-k-means with five clustering algorithms and the results have shown that the TW-k-means algorithm significantly outperformed the other five in four evaluation indices.

The rest of this paper is organized as follows. In section II, Project Related work for Weighting Clustering, cluster ensemble and Multiview Clustering. In section III, Clustering algorithm. The Experimental results are given in section IV. This paper concludes with section V.

II. PROJECT RELATED WORK

2.1 Weighting Clustering

The proposed the W- k-means clustering algorithm that can automatically compute variable weights in the k-means clustering process. W-k-means extends the standard k-means algorithm with one additional step to compute variable weights at each iteration of the clustering process. The variable weight is inversely

ISSN: 2454-9665

proportional to the sum of the within-cluster variances of the variable. As such, noise variables can be identified and their affects on the clustering result are significantly reduced. The new algorithm we propose in this paper weights both views and individual variables and is an extension to W- k-means. A multivariate Dirichlet process mixture model which is based on a cluster model for multivariate means and variances. The model is learned by a Markov chain Monte Carlo process. However, its computational cost is prohibitive. Bouveyron et al. [5] proposed the GMM model which takes into account the specific subspaces around which each cluster is located, and therefore limits the number of parameters to estimate. Tsai and Chiu [6] developed a variable weights self-adjustment mechanism for k-means clustering on relational data sets, in which the variable weights are automatically computed by simultaneously minimizing the separations within-clusters the traditional variable weighting clustering methods only compute weights for individual variables and ignore the differences in views in the multiview data. Therefore, they are not suitable for clustering of multiview data.

2.2 Cluster Ensemble

The purpose of cluster ensemble is to build a robust clustering portfolio that can perform as good as if not better than the single best clustering algorithm across a wide-range of data sets. Different clustering algorithm may take a different approach.

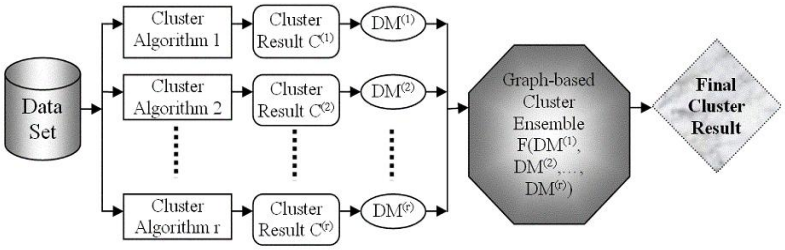

Fig. 1. Cluster Ensemble Architecture.

For example, K-means is to group the data set so that the total Mean Square Error to the center of each cluster is minimum while graph-based partitioning clustering is to partition the graph into K parts based on the minimum edge weight cuts. Thus a cluster ensemble can be used to generate many cluster results using various clustering algorithms and then integrate them using a consensus function to support the various yield stable results. We present a two phase clustering combination strategy. At the first step, various clustering algorithms are run against the same data sets to generate clustering results. At the second step, these clustering results are combined by an auto associative additive system based on the distance matrix of graph clustering. The diagram below summarizes our approach. In our approach, a distance matrix is first constructed based on the cluster results from each individual cluster in algorithm; these distance matrices are combined to form a master distance matrix [7]. Then a weighted graph is constructed from the master distance matrix and a graph-based partitioning algorithm is applied to the graph for the final clustering results. Graph-based clustering uses various kinds of geometric structure or graphs for analyzing data.

2.3 Multi view Clustering

The variable weighting multiview clustering, as a combination of variable weighting method clustering and multiview clustering method, is a new direction for clustering of multiview data. To our knowledge, cluster is

ISSN: 2454-9665

the first clustering algorithm that uses weights for both views and individual variables in the clustering process [8]. The SYNCLUS clustering process is divided into two stages. Starting from an initial set of variable weights, SYNCLUS first uses the k-means clustering process to partition data into k clusters. It then estimates a new set of optimal weights by optimizing a weighted mean-square, stress-like cost function. The two stages iterate until the clustering process converges to an optimal set of variable weights. SYNCLUS only computes variable weights automatically and the view weights are given by users. Another weakness of SYNCLUS is that it is time consuming [9] so it cannot process large data sets. The proposed a weighted combination of exemplar-based mixture models for clustering multiview data that assigns different weights to the views and learns those weights automatically. In each view, the data is modeled using exemplar based mixture models, called convex mixture mode is (CMMs) [10]. However, this method does not consider individual variable weights so it cannot capture the differences among variables in a view. To sum up, the current two variable weighting multi-view clustering methods cannot automatically compute weights for both views and individual variables. Moreover, they are not scalable to large data sets. The proposed method can automatically compute two types of weights and it retains the efficiency of the k-means algorithm.

III. CLUSTERING ALGORITHM

In the two-level variable weighting method, the variable weights V are used to identify the important variables in each view, and the view weights W are used to identify compact cluster structures within these views. If the view contains compact cluster structures, a large view weight is assigned so as to enhance the effect of such view; on the contrary, if the view contains loose cluster structures, a small view weight is assigned to eliminate the effect of such view. Compared with the traditional variable weighting method[11],[12], the new method can take both individual variables and multiple views into consideration and capture the differences among different views and variables. Moreover, the traditional variable weighting methods suffer from unbalanced phenomenon: the view with more variables will play more important role than the view with less variables. In the two-level variable weighting method, the view weights will be only determined in the view level, while the variable weights will be only determined in a view. Therefore, the two levels of variable weights will eliminate the unbalanced phenomenon and compute more objective weights.

ALGORITHM: K-means

Input:

K: The number of clusters,

D: A data clustering n objects

Output:

A set of K clusters

Method:

1. Arbitrarily choose K objects from D as the initial cluster centers.
2. repeat
3. (re) assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
4. Update the cluster means i.e., calculate the mean value and variance of the objects for each clusters(λ);

ISSN: 2454-9665

5. Until no change;

Repeat

Update the partition matrix in λ;

Update the cluster centers in K;

Update the dimension weights in D;

Until (the objective function obtains its local minimum value);

The input parameter γ>0 is used to control the size of the weights as follows:

- γ > 0. In this case, according D_{li} is inversely proportional to D_{li}. The smaller D_{li} the larger γ_{li}, the more important the corresponding dimension.
- γ>0. λ_{li} is equal to one, indicating that the index has the smallest value of D_{li}. The other weights. Each cluster contains only one important dimension. It may not be desirable for high-dimensional data sets.
- γ < 0. In this case is proportional to D_{li}. The larger D_{li} the larger λ_{li}. This is contradictory to the original idea of dimension weighting. Therefore, γ cannot be smaller than zero.

Since the additional algorithm we have used in Weighted K-Means algorithm is an extension to the k-means algorithm by adding a new step to calculate the variable weights in the iterative process, it does not seriously affect the scalability of the k-means type algorithms in clustering high dimensional data; therefore, it is suitable for data mining applications. The data sets of γ in subspace clustering will be shown in Section IV [23].

ALGORITHM: TW- K-means

Input:

The number of clusters K and two positive real parameters λ, ή;

Output:

Optimal values of U, Z, V, and W;

Randomly choose K cluster centers Z^0;

Method:

for $t = 1$ to T do

$w_t^0 \leftarrow 1/T$

for all $j \in G_t$ do

$v_j^0 \leftarrow 1/|G_t|$

end for

end for

$r \leftarrow 0$

Repeat

Update the partition matrix in $w_t^0 \leftarrow 1/T$;

Update the cluster centers in $j \in G_t$;

Update the dimension weights in $r \leftarrow 0$;

Until (the objective function obtains its local minimum value);

ISSN: 2454-9665

The input parameter $t = 1$ is used to control the size of the weights as follows:

TW- k-means algorithm is an extension to the k-means algorithm by adding two additional steps to calculate the view weights and individual variable weights in the iterative process [13],[14]. It does not seriously affect the scalability of the k-means clustering process in clustering large data. If TW-k-means algorithm needs r iterations to converge, the computational complexity of the algorithm is weight. Therefore, TW-k-means has the same computational complexity as k-means.

IV. EXPERIMENTAL RESULTS

This section presents the evaluation of the proposed link based method, using a variety of validity indices and real data sets. The quality of data partitions generated by this technique is assessed against those created by different categorical data clustering algorithms and cluster ensemble techniques.

4.1 Data Sets

To investigate the performance of the TW-k-means algorithm in classifying real-life data, we selected three data sets from UCI Machine Learning Repository [23], the Multiple Features data set, the Internet Advertisement data set and the Image Segmentation data set. With these data, we compared TW- k-means with four individual variable weighting clustering algorithms, i.e., W-k-means [20], EW-k-means (see Section 5.1), LAC [21] and EWKM [22], and a weighted multiview clustering algorithm WCMM [15]. The Multiple Features data set contains 2,000 patterns of handwritten numerals that were extracted from a collection of Dutch utility maps. These patterns were classified into 10 classes (“0”-“9”), each having 200 patterns. Each pattern was described by 649 features that were divided into the following six views:

1. mfeat-fou view: contains 76 Fourier coefficients of the character shapes;
2. mfeat-fac view: contains 216 profile correlations;
3. mfeat-kar view: contains 64 Karhunen-Love coefficients;
4. mfeat-pix view: contains 240 pixel averages in 2 x3windows;
5. mfeat-zer view: contains 47 Zernike moments;
6. mfeat-mor view: contains 6 morphological variables.

Here, we use G1;G2;G3;G4;G5, and G6 to represent the six views.

The Internet Advertisement data set contains a set of 3,279 images from various web pages that are categorized either as advertisements or nonadvertisements (i.e., two classes). The instances are described in six sets of 1,558 features, which are the geometry of the images (width, height, aspect ratio), the phrases in the url of the pages containing the images (base url), the phrases of the images url (image url), the phrases in the url of the pages the images are pointing at (target url), the anchor text, and the text of the images alt (alternative) html tags (alt text). All views have binary features, apart from the geometry view whose features are continuous. Details for the construction of the data set can be found in [16],[17]. The Image Segmentation data set consists of 2,310 objects drawn randomly from a database of seven outdoor images.

4.2 Experiment Designs

With the three real-life data sets introduced in the last section, we carried out two experiments to compare TW-k-means with five clustering algorithms, i.e., W- k-means, k-means, LAC, EWKM and WCMM. The purpose of

ISSN: 2454-9665

the first experiment was to select proper parameter values for comparing the clustering performance of six algorithms in the second experiment. In each experiment, the number of clusters for all clustering algorithms were set as the actual number of classes of the used data set. In the first experiment, we set the parameter values of four clustering algorithms as 30 integers from 1 to 30. For TW-k-means, we set t as 30 integers from 1 to 30 and t as 12 values of {10;20;30;40;50;60;70;80;90;100;110;120}. Since the clustering results of the five clustering algorithms excluding WCMM were affected by the initial cluster centers, we randomly generated 100 sets of initial cluster centers for each data set. For each parameter setting, we ran each of the five clustering algorithms to produce 100 clustering results on each of the three data sets. For WCMM, we set t as eight values {1;1:5;2;2:5;3;3:5;4;4:5}. Since WCMM can find global optima, we only ran WCMM once. In the second experiment, we first set the parameter values for six algorithms by selecting those with the best results in the first experiment. Similar to the first experiment, we produced 100 results for each of the five clustering algorithms excluding WCMM and 1 result for WCMM on each data set [17],[18].

In order to compare the classification performance, we used precision, recall, f-measure and accuracy to evaluate the results [19]. Precision is calculated as the fraction of correct objects among those that the algorithm believes belonging to the relevant class. Recall is the fraction of actual objects that were identified. F-measure is the harmonic mean of precision and recall and accuracy is the proportion of correctly classified objects. All four indices use the corresponding actual classification as the reference classification.

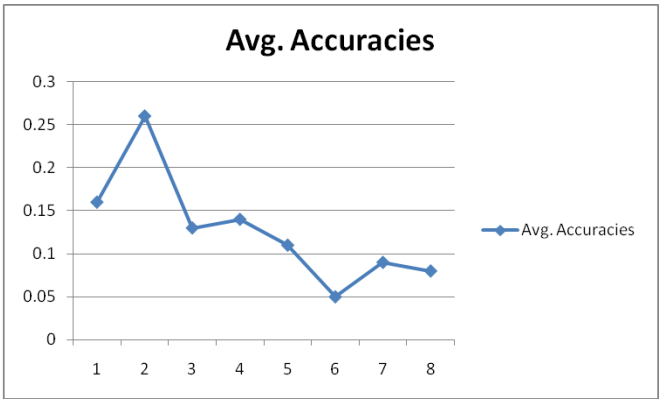

Fig. 2. The six Clustering algorithm for multiple Features data set

4.3 Results and Analysis

As an example, Fig. 2 draws the average clustering accuracies of six clustering algorithms on the Multiple Features data set in the first experiment. From these results, we can observe that TW-k-means produced better results with large value of than the other five algorithms. When was large, it produced relatively stable results with the change of WCMM produced the worst results, which indicates that WCMM failed to recover the clusters from this high-dimensional multiview data. EWKM [18],[19] produced unstable and worse results than W- k-means, LAC and T W- k-means. E W- k-means produced similar results as W- k-means, which indicates that the regularization term affects the result not too much. In the second experiment, we set the parameter values of six clustering algorithms as shown in Table 1 summarizes the total 1,503 clustering results.

From these results, we can see that TW-k-means significantly out-performed the other five algorithms in almost all results, especially on the Multiple Features and Internet Advertisement data sets. Although TW-k-means is

extension to EW-k-means, the introduction of weights on views improved its results. WCMM[20][21] produced the worst results on all three data sets. We used all five real- life data sets to compare the scalability of TW-k-means with the other five clustering algorithms. The average time costs of six clustering algorithms. We can see that the execution time of TW- k-means was only more than EW- k-means, and significantly less than the other four clustering algorithms. This result indicates that TW-k-means scales well to high-dimensional data

TABLE 1 Summary of Clustering on Three Real-Life Data Sets by Six Clustering Algorithms

S. No.	Data	Evaluation indices	W-k-means	EW-k-means	LAC	EWKM	WCMM	TW-k-means
1	MF	Precision	-0.06(.10)*	-0.07(.10)*	-0.07(.09)*	-0.24(.08)*	-0.59(.00)*	-0.79(.09)*
		Recall	-0.09(.09)*	-0.09(.09)*	-0.09(.08)*	-0.36(.12)*	-0.56(.00)*	-0.82(.08)*
		F-measure	-0.08(.10)*	-0.08(.10)*	-0.08(.08)*	-0.41(.12)*	-0.59(.00)*	-0.80(.09)*
		Accuracy	-0.09(.09)*	-0.09(.09)*	-0.09(.08)*	-0.36(.12)*	-0.56(.00)*	-0.82(.08)*
2	IA	Precision	-0.06(.19)*	-0.16(.20)*	-0.14(.20)*	-0.22(.19)*	-0.56(.00)*	-0.72(.12)*
		Recall	-0.14(.04)*	-0.10(.07)*	-0.10(.08)*	-0.13(.06)*	-0.33(.00)*	-0.72(.07)*
		F-measure	-0.23(.04)*	-0.17(.12)*	-0.17(.12)*	-0.21(.09)*	-0.47(.00)*	-0.69(.11)*
		Accuracy	-0.14(.04)*	-0.10(.07)*	-0.10(.08)*	-0.13(.06)*	-0.33(.00)*	-0.72(.07)*
3	IS	Precision	-0.03(.07)*	-0.04(.08)*	-0.03(.07)*	-0.03(.09)*	-0.37(.00)*	-0.62(.09)*
		Recall	-0.03(.05)*	-0.03(.03)*	-0.03(.05)*	-0.03(.05)*	-0.41(.00)*	-0.64(.05)*
		F-measure	-0.01(.07)*	-0.02(.05)*	-0.01(.07)*	-0.02(.07)*	-0.40(.00)*	-0.60(.07)*
		Accuracy	-0.03(.05)*	-0.03(.03)*	-0.03(.05)*	-0.03(.05)*	-0.41(.00)*	-0.64(.05)*

The value in brackets is the standard deviation of 100 results. "*" indicates that the difference is significant

V. CONCLUSION

In this paper, we have presented k -means, a new TW-k-means type algorithm that can calculate variable weights automatically. Based on the current partition in the iterative k-means clustering process, the algorithm calculates a new weight for each variable based on the variance of the within cluster distances. The new weights are used in deciding the cluster memberships of objects in the next iteration. The optimal weights are found when the algorithm converges. An innovative two-level variable weighting clustering algorithm for clustering of multiview data. Given multiple-view data, TW-k-means can compute weights for views and individual variables simultaneously in the clustering process. With the two types of weights, compact views an d important variables can be identified and effect of low-quality views and noise variables can be reduced. Therefore, TW-k-means can obtain better clustering results than individual variable weighting clustering algorithms from multiview data. We used two real-life data sets to investigate the properties of two types of weights in TW-k- means. We discussed the difference of the weights between TW-k-means and EW-k-means algorithms.

The experimental results on both synthetic data and real data sets have shown that the TW-k -means algorithm out-performed the k -means type algorithms in recovering clusters in data. The synthetic data experiments have demonstrated that the weights can effectively distinguish noise variables from the normal variables. We compared TW-k-means with five clustering algorithms on three real-life data sets and the results have shown that the TW-k-means algorithm significantly outperformed the other five clustering algorithms in four evaluation indices. As such, it is a new variable weighting method for clustering of multiview data. In the future, we will combine the two-level variable weighting method with other techniques such as fuzzy techniques,

subspace clustering techniques, semi-supervised techniques etc. so as to apply our method to more applications. Moreover, we will investigate approaches that can automatically group variables in the clustering process.

VI. ACKNOWLEDGMENTS

The authors would like to thank all staff in Department of Information Technology in PSG College of Technology, Coimbatore Dist. India and Dr.M.Madheswaran Principal, Mahendra Engineering College, Namakkal, India for providing the valuable comments and suggestions of online an entropy based weighting k-means clustering for multiview data, data partitioning, High dimensional data, Big Data Analytics and suggestions on experiments.

REFERENCES

[1] J. Grambeier and A. Rudolph, "Techniques of Cluster Algorithms in Data Mining," Data Mining and Knowledge Discovery, vol. 6, pp. 303-360, 2002.

[2] Z. Huang, "Extensions to the K-Means Algorithm for Clustering Large Data Sets with Categorical Values," Data Mining and Knowledge Discovery, vol. 2, pp. 283-304, 1998.

[3] M.J. Zaki and M. Peters, "Clicks: Mining Subspace Clusters in Categorical Data via Kpartite Maximal Cliques," Proc. Int'l Conf. Data Eng. (ICDE), pp. 355-356, 2005.

[4] D. Barbara, Y. Li, and J. Couto, "COOLCAT: An Entropy-Based Algorithm for Categorical Clustering," Proc. Int'l Conf. Information and Knowledge Management (CIKM), pp. 582-589, 2002.

[5] A.P. Topchy, A.K. Jain, and W.F. Punch, "Clustering Ensembles: Models of Consensus and Weak Partitions," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 27, no. 12, pp. 1866-1881, Dec. 2005.

[6] C. Domeniconi and M. Al-Razgan, "Weighted Cluster Ensembles: Methods and Analysis," ACM Trans. Knowledge Discovery from Data, vol. 2, no. 4, pp. 1-40, 2009.

[7] L.I. Kuncheva and D. Vetrov, "Evaluation of Stability of K-Means Cluster Ensembles with Respect to Random Initialization," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 28, no. 11, pp. 1798-1808, Nov. 2006.

[8] A.P. Topchy, A.K. Jain, and W.F. Punch, "A Mixture Model for Clustering Ensembles," Proc. SIAM Int'l Conf. Data Mining, pp. 379-390, 2004.

[9] X.Z. Fern and C.E. Brodle y, "Random Projection for High Dimensional Data Clustering: A Cluster Ensemble Approach," Proc. Int'l Conf. Machine Learning (ICML), pp. 186-193, 2003.

[10] Z. Yu, H.-S. Wong, and H. Wang, "Graph-Based Consensus Clustering for Class Discovery from Gene Expression Data," Bioinformatics, vol. 23, no. 21, pp. 2888-2896, 2007.

[11] S. Bickel and T. Scheffer, "Multi-view Clustering," Proc. IEEE Fourth Int'l Conf. Data Mining, pp. 19-26, 2004.

[12] V.R. de Sa, "Spectral Clustering with Two Views," Proc. IEEE 22nd Int' l Work shop Lea rnin g with Multi pl e View s (ICML) ,pp. 20-27, 2005.

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

[14] K. Chaudhuri, S. Kakade, K. Livescu, and K. Sridharan, "Multi-view Clustering via Canonical Correlation Analysis," Proc. 26th Ann. Int'l Conf. Machine Learning, pp. 129-136, 2009.

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

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

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

[18] G. De So ete, "OV WTR E: A Program for Optimal Variable Weighting for Ultrametric and Additive Tree Fitting, " J.Classification, vol. 5, no. 1, pp. 101-104, 1988.

[19] V. Makarenkov and P. Legendre, "Optimal Variable Weighting for Ultrametric and Additive Trees an d k -Means Partitioning: Methods and Software," J. Classification, vol. 18, no. 2, pp. 245- 271, 2001.

[20] D. Modha and W. Spangler, "Feature Weighting in k-Means Clustering," Machine Learning, vol. 52, no. 3, pp. 217-237, 2003.

[21] C. Bouveyron, S. Girard, and C. Schmid, "High Dimensional Data Clustering," Computational Statistics and Data Analysis, vol. 52, no. 1, pp. 502-519, 2007.

[22] C.-Y. Tsai and C.-C. Chiu, "Developing a Feature Weight Self-Adjustment Mechanism for a k-Means Clustering Algorithm," Computational Statistics and Data Analysis, vol. 52, no. 10, pp. 4658-4672, 2008.

[23] Z. Huang, "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, 1998.

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

[25] Sumit Gupta "SECURITY ISSUES IN WIRELESS AD-HOC NETWORK:A REVIEW" International Journal of Advance Research in Science and Engineering-2319-8354

Tamilselvan R received the B.E (Computer Science and Engineering) from Muthayammal Engineering College, Rasipuram, Namakkal District Tamilnadu, India in 2005 Anna University Chennai and M.E(Software Engineering) from Bannari Amman Institute of Technology, Sathyamangalam, Erode District, Tamilnadu, India in Anna University of Technology Coimbatore 2010, and is currently an Research Scholar in Department Information Technology at PSG College of Technology, Coimbatore District, Tamilnadu, India. He majors in computer science area and familiar with Data Mining area. His research interest in Data mining, Data Clustering, Image processing and Big Data Analytics.

ISSN: 2454-9665

	**Philip Vinod K** received the B.Tech (Information Technology) from Annai Mathammal Sheela Engineering College, Namakkal District Tamilnadu, India in 2003 Periyar University, Salem and M.Tech(Information Technology) from Sathyabama University in 2011 Chennai and is currently an Faculty in the JNN Institute of Engineering Uthukottai, Thiruvallur District. He majors in computer science area and familiar with Data Warehousing and Data Mining area. His research interest in Data mining, Data Clustering, Image processing Wireless networks And Artificial Intelligence.
	**Dr.C.Palanisamy** received the B.E (Electronics and Communication Engineering) from K.S.R College of Technology, Tiruchengode, Namakkal District Tamilnadu, India in 1998 University of Madras and M.E(Communication System) from Thiyagarajar College of Engineering Madurai District, Tamilnadu, India in Madurai Kamaraj University 2000 and Ph.D degree in Data mining at Anna University Chennai and is currently an Professor and Head in Department of Information Technology at Bannari Amman Institute Technology, Sathyamangalam, Erode District, Tamilnadu, India. He majors in computer science area and familiar with Data Mining area.