International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)
ISSN: 0976-1353 Volume 7 Issue 1 –MARCH 2014.

Automated Two Level Variable for Multiview Clustering

R.Gandhimathi,
PG Student,
Department of CSE,
Chettinad College of Engg & tech,
Karur, India.
gandhi.banumsc@gmail.com

Ms.T.Seetha,
Assistant Professor,
Department of CSE,
Chettinad college of Engg & tech,
Karur, India.
seethasns@gmail.com

***Abstract* – Clustering is used to identify the relationship among different objects from large volume of data. The clustering analysis is feasible only when the groups are formed with important features. The existing K-Means clustering processing time and the computation cost is high. The proposed two level variable weighting algorithm calculates weights for both views and variables to identify the important Properties. Due to automatic calculation, clustering can be done effectively in a smaller number of steps. The performance of Two-variable K-means is compared with five clustering algorithms to measure clustering accuracy in multiview data.**

***Index terms*- K-Means, high dimensional data analysis, Multiview clustering**.

I. INTRODUCTION

Cluster is a collection of data objects in which objects are similar to one another within the same cluster and dissimilar to objects belonging to other clusters. Clustering technique can be applied in fields of engineering, economics, and logistics. Clustering analysis many times be necessary to change preprocessing data until the result achieves request meaning.

The multiview dataset collection contains set of different features extracted from the same unanalyzed data. This data is meant for exploring the results of multiple clustering runs against each other and against different features to get interesting patterns. In weighted clustering every data point is assigned a real valued weight.

Clustering algorithm is a program seeks to identify and reveal presence of vectors in high point dimensional involving two or more quantities in space, share some dissimilarity quality.

Various methods in Clustering classified as partitioning, density based, Hierarchial and Grid based. Partitioning Methods conduct partitions in single level partition on datasets.

Hierarchical method is level by level search in cluster. In Density based the general idea is based on density to form a cluster. Grid Based method minimize object space into limited number of cells to make grid structure.

In weighted clustering, it is used to predict output from independent variables based on systematic data by modeling conditional probability density. The automated clustering algorithms on weighted multiview data calculate weights for variables and views. Multiview clustering uses facts from many views and relation between distinct views into examination and process more accurate and strong data partitioning.

II. RELATED WORK

Synclus(Synthesized clustering)is the first variable weighting multiview clustering algorithm, which considers both the compactness of the view and importance of the variable. It calculates the variable weight automatically and view weights are assigned by the user. The WCMM (Weighted Combination of Mixture Models)method automatically calculates view weights but, it does not consider variable weights.

Therefore the existing two algorithms are not scalable to large data sets and takes more time to respond. Recent Multiview clustering utilize all multiple views in order to construct correct and split the data.

III. EXISTING SYSTEM

A. *K-Means Clustering*

The K-means algorithms for partitioning, where each cluster center represented by mean value of objects in the cluster.

Input: K: Number of clusters

D: Dataset having n objects.

Methods:

1. Arbitrarily choose k objects from d as initial cluster centers.
2. Repeat(1)
3. Assign each object to cluster to object most similar, based on the mean value of objects in the cluster.
4. Update the cluster means, calculate the mean value of objects in each cluster until there is no change.

B. *Weighted k-means clustering:*

In K-Means clustering only similarity is measured and data points are grouped into single cluster. Weighted K-Means with additional steps calculate weights for every individual variables. Weight values are calculated using the given formula .

$$\hat{w}_j = \begin{cases} 0 & that\ is\ D_j = 0 \\ \dfrac{1}{\sum_{t=1}^{h}\left[\frac{D_j}{D_t}\right]^{\frac{1}{\beta-1}}} & \end{cases}$$

$$D_j = \sum_{l=1}^{k}\sum_{i=1}^{n} \hat{u}_{i,l}\, d(x_{i,j} - z_{l,j})$$

Where,

- ui,l means object i is assigned to cluster l
- d(x i,j , zl,j) is the distance between objects x and z
- h is the number of variables Dj such that Dj≠0
- z l,j is the value of the variable j of the centroid of the cluster l

$$P(\hat{U}, \hat{Z}, W) = \sum_{j=1}^{m} w_j^{\beta} \sum_{l=1}^{k}\sum_{i=1}^{n} \hat{u}_{i,l}\, d(x_{i,j}, Z_{l,j}) = \sum_{j=1}^{m} w_j^{\beta} D_j$$

C. *Entropy based Weighted k-means clustering:*

In weighted K means weights for variables are fixed. Entropy based W-*k*-means collaborate with W-k--means additionally evaluate entropy value with the following optimization function:

$$P(U, Z, W) = \sum_{l=1}^{k}\sum_{i=1}^{n}\sum_{j=1}^{m} u_{i,l}\, \emptyset_j\, d\left(x_{i,j}, Z_{l,j}\right) + \eta \sum_{j=1}^{m} \emptyset_j \log\left(\emptyset_j\right)$$

Where

$$\emptyset_j = \frac{exp - \left\{\frac{-F_j}{\eta}\right\}}{\sum_{j=1}^{m} exp - \left\{\frac{-F_j}{\eta}\right\}}$$

IV. PROPOSED SYSTEM

The proposed TW-K-Means two level variable weighting k-means cluster algorithm

distinguish the impacts of different views and variables in clustering. The proposed algorithm calculate both the view weight and variable weight automatically. The view weights reflect the importance of the views in the multiview dataset and the variable weights in a view only reflect the importance of the variables in the view.

A. *Two level variable weighting algorithm:*

- Input: Number of k clusters.
- Output: Optimal values of u, z, v, w.
- Randomly choose k cluster centers
- Select the number of data points.
- For t=1 to T Assign weight values to each datas.
- Calculate the view weight and variable weight
- u-data partition , z – how many number of clusters
- Partition the data upto cluster l

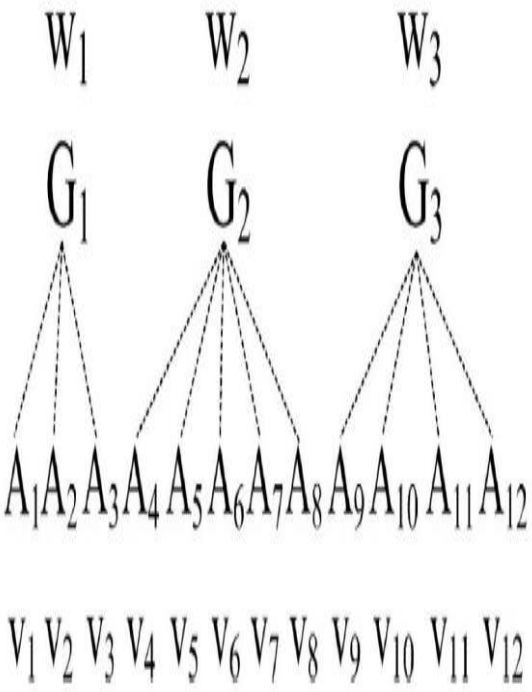
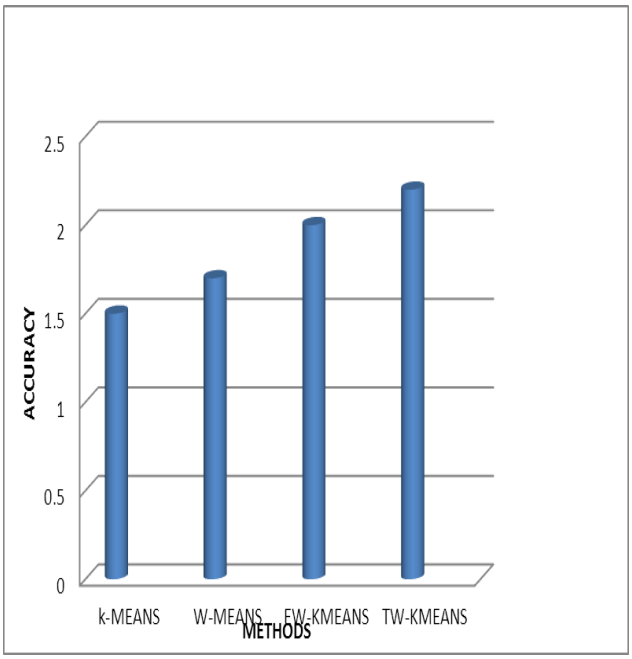


Fig 4.1 Two level Variable Weighting

Here A1,A2.A3 are attributes for the individual variables and they are viewed and v1,v2,v3 are weights for individual variables. The attributes of the variables are grouped as G1,G2,G3 and they are viewed and weight is assigned as W1,W2,W3 for given data. So both weights calculated simultaneously.

B. *Performance Evaluation*:

Performance are evaluated based on the K-means, W-Means, Ew-K-means, Tw-K-Means. Paired t-test comparing is evaluated and accuracy is measured.



V. CONCLUSION

In Standard variable weighting methods are not equally balanced, since only variable weights are considered. Comparing with traditional weighting methods, in two level variable weighting, the variable weights used to identify important variables and view weights used to identify close cluster structures within these views. So clustering results is efficient compared to individual variable weighting clustering algorithm. This work can be further extended to implement on

fuzzy techniques to automatically group variables in clustering process.

REFERENCES

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

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

[3] Z. Huang, M. Ng, H. Rong, and Z. Li, "Automated Variable Weighting in k-Means Type Clustering," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 27, no. 5, pp. 657-668, May 2005.

[4] C. Domeniconi, D. Gunopulos, S. Ma, B. Yan, M. Al-Razgan, and D. Papadopoulos, "Locally Adaptive Metrics for Clustering High Dimensional Data," Data Mining and Knowledge Discovery, vol. 14, no. 1, pp. 63-97, 2007.

[5] L. Jing, M. Ng, and Z. Huang, "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, 2007.

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

[7] D. Lashkari and P. Golland, "Convex Clustering with Exemplar Based Models," Advances in Neural Information Processing Systems, vol. 20, pp. 825-832, 2008.

[8]Z. Deng, K. Choi, F. Chung, and S. Wang, "Enhanced Soft Subspace Clustering Integrating Within-Cluster and Between-Cluster Information," Pattern Recognition, vol. 43, no. 3, pp. 767-781, 2010.

[9] P. Hoff, "Model-Based Subspace Clustering," Bayesian Analysis,vol. 1, no. 2, pp. 321-344, 2006.

[10] M.B. Blaschko and C.H. Lampert, "Correlational Spectral Clustering,"Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR '08), pp. 1-8, 2008.

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

.