International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
Volume 3– No.5, July 2012 – www.ijais.org

Hybrid Clustering Algorithm based on Mahalanobis Distance and MST

V. Valli Kumari
Computer Science & Systems Engineering,
Andhra University,
Visakhapatnam
Andhra Pradesh, India, 5300 03

BHVS Ramakrishnam Raju
Information Technology,
SRKR Engineering College,
Bhimavaram, , India, 534204

Azad Naik
Department of Computer Science
George Mason University
Fairfax, VA 22030 USA

ABSTRACT
Most of the clustering algorithms are based on Euclidean distance as measure of similarity between data objects. Theses algorithms also require initial setting of parameters as a prior, for example the number of clusters. The Euclidean distance is very sensitive to scales of variables involved and independent of correlated variables. To conquer these drawbacks a hybrid clustering algorithm based on Mahalanobis distance is proposed in this paper. The reason for the hybridization is to relieve the user from setting the parameters in advance. The experimental results of the proposed algorithm have been presented for both synthetic and real datasets.

General Terms
Data Mining, Clustering, Pattern Recognition, Algorithms.

Keywords
Minimum Spanning Tree, Fuzzy, Mahalanobis.

1. INTRODUCTION
Categorizing the set of data items into natural groups (clusters) is called Clustering. Many algorithms have been proposed for clustering. These algorithms are mainly divided into partitional and hierarchical algorithms [1, 2]. The most popular partitional algorithms are K-Means and Fuzzy C-Means algorithms because of their simplicity. But on the other hand these algorithms suffer from the drawbacks like, random selection of initial cluster centers and number of clusters should be known a prior. On the other hand hierarchical clustering algorithms are not efficient for bulky datasets. Hierarchical algorithms are classified into agglomerative and divisive approaches. These algorithms also require the number of clusters as an input parameter to terminate the algorithm. To alleviate the burden of selecting the input parameters, a graph based divisive algorithm based on minimum spanning tree (MST) [3, 4] is used in this paper.

Clustering process starts with creation of distance (similarity) matrix. The similarity between the data points is assessed using distance measure between the data points. The most common measure for similarity is Euclidean distance. The Euclidean distance is generally used for low dimensional data sets. It is also called as the L_2 norm. The Euclidean distance is the usual manner in which distance is measured in real world. Euclidean distance is easy and fast to implement but it has some drawbacks [5]: it is sensitive, if the variables involved in the dataset are at different scales and it does not account for the relation between the variables. The clustering algorithms using Euclidean distance as a similarity measure are subjective to the magnitude of variables involved in the dataset. Mahalanobis is another distance function [6] that is used for similarity measure, if the variables in the data set are correlated and if the variability is to be included in the distance metric. The covariance matrix provides normalization of the data relative to their stretch. Hence data need not be normalized. In this paper, a hybrid clustering algorithm using MST based on Mahalanobis distance measure is tested.

2. RELATED WORK
Cluster analysis is a complicated problem due to the number of similarity measures exists and there is no universal solution for the entire situation in the domain. Several clustering algorithms were developed based on different distance measures for variety of cluster shapes [7-11]. But the results of these algorithms were contradictory when applied for clustering of other shapes. Gustafson-Kessel (GK) clustering algorithm [7] and Gath-Geva (GG) clustering algorithm [8] were proposed to discover non-spherical clusters. A modified Mahalanobis distance with preserved volume was used in GK algorithm. It is based on fuzzy clustering algorithm for partitioning the data sets with different geometrical shapes. However, if the prior information about the cluster is not known, the algorithm will carry the singular problem for the inverse covariance matrix. The Gaussian distance can only be used for the data with multivariate normal distribution in GG algorithm. It was pointed out that clustering techniques proposed for well separated clusters fails when applied for overlapping clusters [12].

Center-based clustering approaches, like, Fuzzy C-Means (FCM) [1], Possibility C-Means (PCM) [13], and Fuzzy Possibility C-Means (FPCM) [14] algorithms, use Euclidean distance function to measure the similarity between the two data points for hyper-spherical clusters. However, more sophisticated approaches rely on a cluster-specific Mahalanobis distance, making it possible to find clusters of hyper-ellipsoidal shape. This similarity measure will unwind the limitation that all clusters have the same size [15]. But on the other hand Mahalanobis distance measure reduces the sturdiness of the clustering algorithm. The extended versions of FCM, PCM, and FPCM, are Fuzzy C-Means based on Alternative Mahalanobis distance (FCM-AM) algorithm, the Possibility C-Means based on Alternative Mahalanobis distance (PCM-AM) algorithm, the Fuzzy Possibility C-Means based on Alternative Mahalanobis distances (FPCM-AM) algorithms [16], respectively. These algorithms were based on the local and global Mahalanobis distances.

3. METHODOLOGY

3.1 Formal Preliminaries

Given a set of data objects $S = \{x_1, x_2, ...x_N\}$, where $x_i = (x_{i1}, x_{i2}, ..., x_{il})^T \in \Re^l$ is a feature vector. The goal of the clustering algorithm is to organize the data set *S* into K groups, such that the similarity among data objects with in cluster and the variation of data objects in different clusters are maximized. The techniques used for clustering are based on the measures of similarity between data objects. The similarity between the objects is measured using distance between the data objects. A distance metric is a real-valued function d is defined as, $d : S \times S \to \Re$ such that for distinct $(x_i, x_j) \in S$,

(i) $d(x_i, x_j) \geq 0$,

(ii) $d(x_i, x_j) = d(x_j, x_i)$ and

(iii) $d(x_i, x_j) = 0$.

The Euclidean distance is the most common distance metric used for low dimensional data sets. The Euclidean distance is defined as:

$$d_{euclidean}(x_i, x_j) = \sqrt{\sum (x_i - x_j)^2},$$

$$\forall i, j = 1 : N \quad (1)$$

Euclidean measure is helpful in low dimensions, where as for high dimensional data and for categorical variables it performs poorly [17]. Each attribute is treated as totally different from all of the attributes [18].

Mahalanobis distance is a well-known statistical distance function that can used for similarity measure, if the variables in the data set are correlated and if the variability is to be included in the distance metric. The covariance matrix provides normalization of the data relative to their stretch. Hence data need not be normalized [19, 20]. Mahalanobis distance between two samples (x_i, x_j) is defined as:

$$d_{Mahalanobis}(x_i, x_j) = \sqrt{\left((x_i - x_j)^T \sum^{-1} (x_i - x_j)\right)} \quad (2)$$

Where, $\sum^{-1}$ is the inverse of covariance matrix. The Mahalanobis metric is not dependent upon the scales of variables. The Mahalanobis distance or its square can be used to measure closeness of an object from another object.

In the case of Σ = I, Mahalanobis distance is the same as Euclidean distance:

$$d_{Mahalanobis}(x_i, x_j) = \sqrt{\left((x_i - x_j)^T I^{-1} (x_i - x_j)\right)}$$

$$= \sqrt{(x_i - x_j)^2} = d_{euclidean}(x_i, x_j)$$

3.2 Minimum Spanning Tree Clustering

Let $G(S) = (V, E)$ be the undirected graph to represent the given data objects *S*, where *V* = *S* and $E = \{(x_i, x_j) \mid x_i, x_j \in S, i \neq j\}$. The weight or edge length represents the similarity between the objects. The application of MST in clustering was proposed by Zahn [21]. A Minimum Spanning Tree (MST) of *S* is constructed by using either Kruskal's algorithm [22] or Prim's algorithm [23] to initiate the clustering procedure. The weight of any edge in MST is greater than the given threshold (δ), then by removing such *(K-1)* edges (inconsistent) from MST results in *K* number of sub trees, called clusters. The clusters produced may contain data points vary from a few to very large in number. If data points in any cluster are fewer than some threshold then that cluster is eliminated from the clustering process. The centers of each subtree represent initial representatives of the clusters in the splitting stage of the proposed clustering process. Hence the number of clusters need not be set initially in this algorithm. The resulting clusters are merged based on fuzzy similarity measure [24] to find optimal number of clusters. This is explained in the next section

3.3 Fuzzy similarity Merging

The merging of similar clusters offers an automatic approach for cluster validation. The cluster merging method used in this paper is based on fuzzy similarity between pair of clusters [24]. This similarity is based on compactness or dispersion within the cluster and separation between the clusters. The result of this merging is optimal partitioning from that of over partitioning.

The fuzzy dispersion of cluster *i* is defined as:

$$FDISP(T_i) = \left(\frac{1}{n_i} \sum_{x \in T_i} \mu_i^m \|x - c_i\|^2 \right)^{1/2} \quad (3)$$

Where $n_i = |T_i|$ and μ_i denotes i^{th} row of membership matrix μ.

The separation or dissimilarity between the clusters is measured as:

$$DISM(T_i, T_j) = \|C_i - C_j\| \quad (4)$$

The merging criterion for merging two similar clusters is based on:

$$SIMI(T_i, T_j) = \frac{FDISP(T_i) + FDISP(T_j)}{DISM(T_i, T_j)} \forall i, j = 1 : K$$

and $i \neq j$ (5)

It can be seen that $SIMI(T_i, T_j)$ is the ratio of compactness to the separation between the two clusters. This ratio can be used to measure the similarity between the two clusters. Then the cluster validity index is calculated for the analysis of clustering. This index is calculated as explained in fallowing section.

3.4 Validity Ratio

Validity Ratio is used to evaluate the clustering results. In this paper the validity ratio, this is based on compactness to deal with the internal cohesion among the data elements and

isolation to measure separation between the clusters [25]. The compactness is measured by Intra-cluster distance as fallows:

$$Intra = \frac{1}{N}\sum_{i=1}^{K}\sum_{x \in C_i}\left\|x - c_i\right\|^2 \quad (6)$$

Where N is the number of data items in the dataset, K is the number of clusters, and c_i is the centre of cluster.

The isolation or separation is measured by Inter-cluster distance, which is defined as the minimum of the pair wise distance between any two cluster centers given by,

$$Inter = \min\left(\left\|c_i - c_j\right\|^2\right),$$

$$i = 1:K-1 \; and \; j = i+1:K \quad (7)$$

In the evaluation of the proposed clustering algorithm, the validity ratio used is as proposed by Ray and Turi [25]:

$$validity = \frac{Intra}{Inter} \quad (8).$$

4. RESULTS

Experiments were conducted on both artificial and real data sets to test the performance of the proposed algorithm. The artificial data sets are used in our experiments, since they are easily manageable. A 2-D data set Data_B is generated with 163 data points containing 3 well separated clusters. Another artificial data set, Ruspini [26] was used in the testing of the proposed algorithm. Ruspini data set contains 75 data points each with 2D points and distributed in 4 clusters. The Ruspini data set is a simple, well-known example that is commonly used as a benchmark dataset in evaluating clustering methods [27].

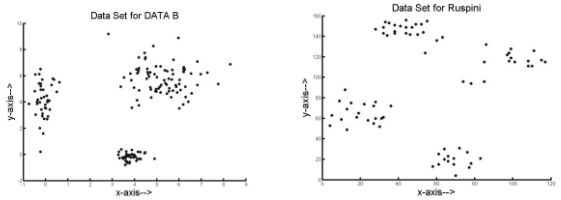
The real data sets used in our experiments are collected from UCI [28]. The data sets are Iris and Wine datasets. Iris data set expresses different categories of iris flowers, having 150 objects with 4 numeric attributes, namely sepal length, sepal width, petal length, and petal width. It has three classes, i.e. Setosa, Versicolor and Virginica, each containing 50 objects. It is known that two classes Versicolor and Virginica have some overlap while the class Setosa is well separated from the other two. Thus, we can accept that there are 2 or 3 [29] clusters in the Iris data set. The Wine dataset contains 178 observations, 13 attributes for each observation and distributed into 3 classes. The details of both artificial and real datasets are presented in table 1.

The proposed algorithm is tested for both Mahalanobis and Euclidean distance measures. The validity ratios determined by the proposed algorithm for both the distance metrics are compared. The validity ratios measured by the proposed algorithm for datasets are as shown in table 2. For Data_B, Ruspini the number of clusters produced by the proposed algorithm is same as that of actual number of clusters present in dataset.

The optimal number of clusters determined by the proposed algorithm for Iris data is 2 as against actual number of clusters present in the data set (3), which is acceptable according to [29]. Except for Iris data set, the validity ratios for Data_B, Ruspini and Wine data sets measured by the proposed algorithm using Mahalanobis distance are lower than the validity ratios measured by using Euclidean distance. Experiments showed that the usage of Euclidean Distance or Mahalanobis Distance affects the expected results. The user parameters required for the proposed algorithm are fuzziness index and the threshold to decide the outlier cluster. Hence this algorithm requires minimum user involvement in order to get desired output.

Table 1. Data sets

Name	Data Size	No of Attributes	Actual no. of Clusters
Data_B	163	2	3
Ruspini	75	2	4
Iris	150	4	3
Wine	178	13	3



(a) (b)

Fig. 2. Data distribution in artificial Datasets (a) Data_B and (b) Ruspini

Table 2. Validity Ratios and number of clusters produced by proposed algorithm for Mahalanobis and Euclidean distance Measures

Name of the Data Set	Validity Ratio	
	Mahalanobis Measure	Euclidean Measure
Data_B	0.0085	0.0519
Ruspini	0.000035	0.0439
Iris	0.2022	0.0658
Wine	0.000144	0.0751

5. CONCLUSIONS

Most of the clustering algorithms are based on Euclidean distance as measure of similarity between data objects. These algorithms also require initial setting of parameters as a priori, for example the number of clusters. The most common measure for similarity is Euclidean distance. Euclidean distance is easy and fast to implement but it has some drawbacks: it is sensitive, if the variables involved in the dataset are at different scales and it does not account for the relation between the variables. To alleviate the burden of selecting the input parameters, a graph based divisive algorithm using minimum spanning tree (MST) is proposed in this paper. To overcome drawbacks involved in using the Euclidean distance as a similarity measure, Mahalanobis distance is used as a distance function in proposed algorithm. The results showed that the usage of Euclidean Distance or Mahalanobis Distance affects the expected results and also the number of clusters need not be set as a prior for the

proposed algorithm. On the downside using Mahalanobis distance as a similarity measure, is computationally expensive because the inverse of the covariance matrix is to be computed every time a pattern changes its cluster domain. Our future work will focus on this aspect of Mahalanobis distance.

6. REFERENCES

[1] Jain A.K. and Murty M.N. and Flynn, P.J. Data Clustering: A Review. In: ACM Computing Surveys, Number 31, vol. 3, pp. 264-323, 1999

[2] Bezdek, J. C. Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, NY, 1981.

[3] Oleksandr Grygorash, Yan Zhou, Zach Jorgensen. Minimum Spanning Tree Based Clustering Algorithms, Proceedings of the 18th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'06), 2006.

[4] Vathy-Fogarassy Á., Feil B., Abonyi J. Minimal Spanning Tree based Fuzzy Clustering. Transactions on Enformatika, Systems Sciences and Engineering, Volume 8, ISSN: 1305-5313, pp. 7–12, 2005.

[5] Ertöz, L., Steinbach, M., Kumar, V., Finding Clusters of Different Sizes, Shapes, and Densities in Noisy, High Dimensional Data, D. Barbara, C. Kamath, (Eds.), Proceedings of the Third SIAM International Conference on Data Mining, Volume 3, San Francisco: SIAM, 2003.

[6] Hill, T., Lewicki, P., Statistics: Methods and Applications, A Comprehensive Reference for Science, Tulsa: StatSoft Inc., 2006.

[7] D. Gustafson, W. Kessel, Fuzzy clustering with a fuzzy covariance matrix, Proceedings of the IEEE Conference Decision Control, pp. 761–766, 1979.

[8] I. Gath, A. Geva, Unsupervised optimal fuzzy clustering, IEEE Trans. Pattern Anal. Mach. Intell. 11 pp 773–781, 1989.

[9] Y. Man, I. Gath, Detection and separation of ring-shaped clusters using fuzzy clustering, IEEE Trans. Pattern Anal. Mach. Intell., vol 16, No.8 , pp 855–861, 1994.

[10] R.N. Dave, Use of the adaptive fuzzy clustering algorithm to detect lines in digital images, Intell. Robots Comput. Vision VIII 1192, pp600–611, 1989

[11] F. Höppner, Fuzzy shell clustering algorithms in image processing: fuzzy c-rectangular and 2-rectangular shells, IEEE Trans. Fuzzy syst, Vol 5 , pp599–613, 1997

[12] S. Bandyopadhyay,An automatic shape independent clustering technique, Pattern Recognition, vol 37, pp 33–45, 2004

[13] Pal, N. R., Pal, K., & Bezdek, J. C. A possibilistic approach to clustering IEEE Transactions on Fuzzy Systems, vol 1, No 2, pp 98-110, May, 1993.

[14] Pal, N. R., Pal, K., & Bezdek, J. C. A mixed c-mean clustering model, Proceedings of the Sixth IEEE International conference on Fuzzy System , vol 1, pp 11-21, July. 1997.

[15] A. Keller and F. Klawonn. Adaptation of Cluster Sizes in Objective Function Based Fuzzy Clustering. In:C.T. Leondes, ed. Database and Learning Systems IV, pp 181–199. CRC Press, Boca Raton, FL, USA 2003

[16] Hsiang-Chuan Liu, Fuzzy Partition Clustering Algorithms Based on Alternative Mahalanobis Distances, Journal of Educational Measurement and Statistics, 13-32, 2008.

[17] Larose, D.T., Discovering Knowledge in Data: An Introduction to Data Mining, New Jersey: John Wiley and Sons, 2005

[18] Ertöz, L., Steinbach, M., Kumar, V. "Finding Clusters of Different Sizes, Shapes, and Densities in Noisy, High Dimensional Data", January 24 2003, D. Barbara, C. Kamath, (Eds.), Proceedings of the Third SIAM International Conference on Data Mining, Volume 3, 2003, San Francisco: SIAM

[19] Besset, D.H., Object-Oriented Implementation of Numerical Methods: An Introduction with Java and Smalltalk, California: Morgan Kaufmann, 2004.

[20] Eyob, E., Social Implications of Data Mining and Information Privacy: Interdisciplinary Frameworks and Solutions, Pennsylvania: Idea Group Inc., 2009

[21] Zahn, C. T. Graph-theoretical methods for detecting and describing gestalt clusters, IEEE Trans. Comput. C-20 (Apr.), pp. 68–86, 1971.

[22] Kruskal, J.B. On the shortest spanning subtree of a graph and the traveling salesman problem, In American Mathematical Society, vol.7, pp. 48-50, 1956.

[23] Prim, R. Shortest connection networks and some generalizations, Bell System Technical Journal, Vol. 36, pp. 1389-1401., 1957.

[24] Xuejian Xiong, Kap Luk Chan. Similarity-Driven Cluster Merging Method for Unsupervised Fuzzy Clustering. In Proceedings of UAI'2004. pp.611-627, 2004

[25] Ray S., Turi R.H. Determination of Number of Clusters in K–Means Clustering and application in colour Image Segmentation, Proc. 4th Intl. Conf. ICAPRDT '99, pp. 137-143, Calcutta India, 1999

[26] Ruspini Dataset : http://www.unc.edu/~rls/s754/data/ruspini.txt

[27] Pearson, R. K., Zylkin, T., Schwaber, J.S., Gonye, G.E. Quantitative evaluation of clustering results using computational negative controls. Proc. 2004 SIAM International Conference on Data Mining, Lake Buena Vista, Florida, 2004.

[28] Real Datasets: http://archive.ics.uci.edu/ml/index.html

[29] Frank Höppner, F. Klawonn, R. Kruse, and T. Runkler. Fuzzy Cluster Analysis-Methods for Classification, Data Analysis and Image Recognition", John Wiley & Sons, LTD, 1999.