ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering
(An ISO 3297: 2007 Certified Organization)
Vol. 3, Issue 10, October 2015

Clustering using Mahalanobis with Deterministic Initialization

Shubham Goyal[1], Monotosh Manna[2]

[1] M.Tech Scholar, Dept. of Computer Science and Engineering, YIT, Jaipur, India

[2] Assistant Professor, Dept. of Computer Science and Engineering, YIT, Jaipur, India

ABSTRACT: Mahalanobis distance has been an appreciated choice for clustering in mixed Gaussian distribution field. With elliptical cluster formation, it has only one area of improvement that is, choosing proper initialization. We deal with this problem by proposing a deterministic initialization for k means to be used with Mahalanobis distance for fast yet effective results. We further compare our proposal with Melynkov and Melynkov's algorithm. Also, evaluation of the proposal with benchmark datasets is done.

KEYWORDS: Clustering; K-means; Mahalanobis Distance; Initializations of k-means; Performance Comparison.

I. INTRODUCTION

Clustering is the most popular method for many applications like data compression, image processing, text processing, network topology decisions, etc. Originally, clustering as an unsupervised learning method is a technique to group similar objects together. This can be used to identify patterns in data, hence data analysis. The notion of a "cluster" cannot be precisely defined, which is one of the reasons why there are so many clustering algorithms. [1]

Hundreds of clustering algorithms have been around, out of which a 50 years + old k-means clustering method is still one of the top 10 popular clustering techniques [2]. The major issues of k-means include that it can recognize only hyper-spherical structures in an m-dimensional geometric data space. This geometry is bound by the limitations of the geometric properties of the distance measure being used that is, Euclidean distance in traditional k-means. Mahalanobis distance is capable of recognizing elliptical structures. [3] suggested how Mahalanobis can be used with k-means. The challenge in using Mahalanobis is estimating the initial values of means and covariance of each cluster. [3] have proposed a method for estimating the initial values. The results of clustering Iris dataset using this algorithm show that the method is effective. Yet, the technique suggested is time-consuming and few parameters need to be decided as per dataset characteristics. This paper proposes to use a proposed deterministic initialization method to compute the initial estimates.

II. RELATED WORK

A. Popular Initializations of k-means

Kmeans++[4] helps in faster convergence of k-means by reducing the number of iterations and better sum of SME. It initializes the number of clusters k before proceeding for clustering data. Kauffman and Rousseeuw[5] suggested selecting the first centroid and further examining which points in the database on being selected as the next centroid produces the greatest reduction in the SSE. The same process goes for all seeds. Celebi and Kingravi [6] proposed a deterministic initialization of k means using hierarchical clustering modifying the initialization proposed by Su and Dy[7]. Macqueen [8] proposed two methods, first of which takes the first k points of the dataset as centroids and the remaining data points as the members of the cluster with nearest centroid. The same process is repeated for several iterations. Its cost related limitation was then improved by the second method which chose the initial centres randomly from among the data points. Bradley and

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering
(An ISO 3297: 2007 Certified Organization)
Vol. 3, Issue 10, October 2015

Fayyad[9] used Macqueen's second method for initialization producing j subsets of intermediate centroids each with k points. MinMax k-means[10] aims at minimizing the maximum intra cluster variance instead of the sum of intra cluster variances by picking random set of centroids and uses association of weights to cluster depending on the variance such that a weighted version of the sum of the intra cluster variation criterion is derived. Authors in [11] also follows perturbation discipline subdividing the input space into equal intervals based on the weights in vectors.The El Agha initialization [12] first finds the boundaries o data points and then divides the area covered by points into k rows and k columns forming a 2-D grid. It then generates initial k means centroids depending on the overall shape of data.

B. Use of Mahalanobis distance:

Authors in [13] propose a work aiming at learning Mahalanobis distance with given must-link and cannot-link information and uses it in various applications such as image segmentation, data clustering, face pose estimation etc. Authors in [14] use a class including data as input and transfers data to a high dimensional space with a methodology somewhat similar to Kernel methods. The clusters formed using Mahalanobis distance are then tested on two different heuristics-centre based and KNN based algorithm. With this evaluation, enhancement in classification success rates is achieved. Jain and Mao also did their share in hyperellipsoidal clustering as discussed in [15]

C. Mahalanobis in k-means:

One of the first works in this direction was done by Andrew Ceroili. Ceroili[16] proposed a non parametric technique of clustering using k-means and Mahalanobis distance and proves it an exceptional choice in non hierarchical cluster analysis. Also, Mahalanobis distance overcomes the doubt of variable standardization by yielding a scale invariant classification and is easy to implement. Tarsitano[17] suggested a method of implementing k- means using Mahalanobis presuming common variance matrices. It refines the Art et al's algorithm[18].

Authors in [3] have suggested a method to compute estimates of mean and co variances of desired clusters through probabilistic initialization. This initialization method consists of picking centroids according to a ranking system based on distances. Thereafter, fixed number of points is assigned to a cluster based on distance from chosen centroid. The algorithm used for initialization can be briefly put as

Step 1: Compute sum of w smallest distances for each point from all other points, as $S_{i,w}$.
Step 2: Assign a random k^{th} center according to probability inversely proportional to $S_{i,w}$
Step 3: Assign Dpoints nearest to the current center to current cluster C_k.
Step 4: For current cluster, compute the estimated mean and covariance
Step 5: Update membership of points in the core of the current cluster according to a probability coverage criterion.
Step 6: Repeat Steps 3, 4 and 5 upto f times.
Step 7: For all points not assigned to any cluster, compute estimated Mahalanobis distance from centre of C_k and include 'r' points in C_k for edge of the cluster.
Step 8: Repeat Steps 2 to 7 for $k = 1\ to\ K$.

III. DEVELOPMENT OF IDEA

The basic idea of the thesis is to provide an efficient clustering algorithm using k-means and Mahalanobis distance for Gaussian models. Since Mahalanobis requires a proper initialization, we first propose a deterministic initialization method for k-means. The idea is to use knowledge of the dataset properties to have a fast yet effective initialization. For this, we start by focusing on every attribute of the dataset and calculate$Range(A)$, which is the difference between the lowest and highest value of the attribute A. Let $V = \{V_i, V_2, V_3, \ldots, V_m\}$ be the set of values for attribute A. $Range$ cannot be generalized since it depends on the domain knowledge of the dataset. Therefore,

$$Range_A = max_A - min_A \qquad \text{eqn(1)}$$

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering
(An ISO 3297: 2007 Certified Organization)
Vol. 3, Issue 10, October 2015

***Algorithm 1*:**
Input: Prominent attribute 'p'
Step 1: Initial clusters are formed using following conditions
For any data point, if $min_p + j * \delta_p \leq value_p < min_p + (j + p) * \delta_p$, then the data point belongs to cluster j.
Step 2: Centroid of each cluster is computed as mean of all cluster points

This algorithm depends on users' understanding of the characteristics of the dataset. The prominent attribute should be significant enough to reflect the inclination of data pattern. In case, this cannot be judged (may be done to absence of expert knowledge); then a self-learning version of initialization is to be used. This is given in Algorithm 2.

Algorithm 2:
Step 1: Variation of each dimension is computed as, $\delta_i = \frac{Range_i}{k}$, $1 \leq i \leq m$, where $range_i$ is defined as in eqn(1)
Step 2: Initial clusters are formed using following conditions
For any data point, if $min_1 + j * \delta_1 \leq value_1 < min_1 + (j + 1) * \delta_1$, then the data point belongs to cluster j.
Step 3: Centroid of each cluster is computed as mean of all cluster points
Step 4: Record SSE
Step 5: For every secondary dimension, $2 \leq j \leq m$, repeat the same.
Step 6: Retain the clustering with minimum SSE

Very briefly, the proposal is to use the above initialization for initial estimates of range. Thereafter, continue clustering as traditional k-means along with Mahalanobis distance. The algorithm can be outlined as:

Algorithm 3:
Step 1: Estimating mean and covariance
Step 1.1: Apply the proposed initialization Algorithm 2 to obtain initial K clusters and centroids
Step 1.2: Compute estimated mean and covariance of each cluster
Step 2: Mahalanobis K-Means clustering
Step 2.1: For each point, compute Mahalanobis distance from each of the centroids. Assign the point to the cluster as per minimum distance.
Step 2.2: Update the estimates of mean and covariance using points currently assigned to each cluster.
Step 2.3: Repeat Steps 2.1 and 2.2 until clusters are stable.

IV. ANALYSIS OF PROPOSED INITIALIZATION

A prominent attribute is selected as per dataset characteristics and only values of that attribute is considered to decide initial centres. This takes O(n) time, with inner operation of complexity O(1). Thus, overall runtime of proposed initialization can be estimated to O(n^2).

Once centres are decided, Euclidean distance is used to decide the cluster membership. This takes O(n^2) time with inner operations of O(n^2) time with inner operations of O(m^2) complexity.

V. PERFORMANCE COMPARISON

Proposed method can be compared with that of Melynkov and Melynkov[3] in terms of runtime and accuracy of results. The comparison is briefed in Table1

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering
(An ISO 3297: 2007 Certified Organization)
Vol. 3, Issue 10, October 2015

Table 1: Comparison between Melynkov and Melynkov's algorithm and proposed algorithms

Attributes for Comparison	Melynkov and Melynkov's algorithm[3]	Proposed Algorithm 1	Proposed Algorithm2
Time Complexity of Initialization method	$O(n^3)$	$O(n^2)$	$O(mn^2)$
Inner operations	$O(m^3 + n)$	$O(m^2)$	$O(m^2)$
Parameter dependency	D, f, w have to be decided	Prominent attribute is to be decided	No extra parameters
Probabilistic behavior	Cluster memberships depend on probability criteria and hence require many runs for proper results	Deterministic Method	Deterministic Method
Accuracy over Iris dataset	At $D = w = 20, f = 5, R = 0.904$	At $p = 3\ to\ 4, R = 0.904$	At $p = 3\ to\ 4, R = 0.904$

We observe that our proposed algorithm on comparison with Melynkov and Melynkov's[12], saves much time without losing its accuracy.

VI. RESULTS OF THE PROPOSED ALGORITHM ON IRIS DATASET

Fisher's Iris dataset is a very famous dataset taken from UCI[19] which consists of 150 instances of Iris flowers taken from 3 different species; 50 elements each of Setosa, Versicolor and Virginica Iris. We test our proposed algorithm on all the three clusters, each belonging to Setosa, Versicolor and Virginica species respectively with varying dimensions and record the RandIndex.

For k=3 and dimension=1, we observe that the clusters are very well formed and no mixing of data points occur between any two clusters. For k=3 and dimension=2, good cluster formations take place but the data points start to mix within nearby clusters. For k=3 and dimension= 3 and 4, figures are identical and data points get mixed up between clusters and no clarity is observed but the highest possible accuracy of RandIndex=0.904 is achieved. The identical figures sum up the fact that with highest accuracy and increasing dimensions, visual interpretation remains same, that is, mixed clusters. Figure 1 shows the visual evaluation of our experiments for 3 and 4 dimensions.

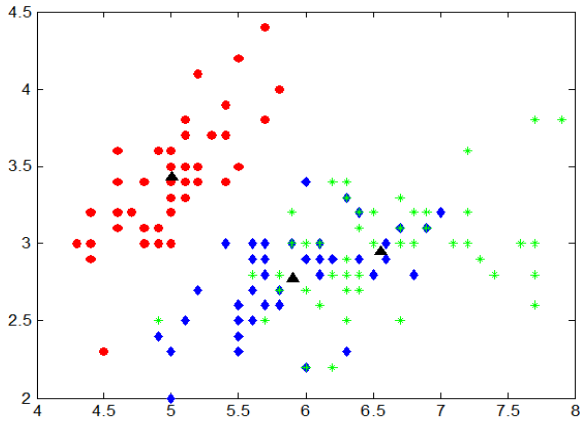

Figure 1: Results of the proposed algorithm over Iris dataset for k=3 and p=3 or 4

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

The computed RandIndex for different attribute taken as prominent one in Iris dataset are tabulated in Table 2. It can be observed that different attributes yield different values of RandIndex. At p=3 and 4, same accuracy of 0.904 is achieved, which is the highest possible accuracy till date for Iris dataset.

Table 2: Tabular representation of the results of the proposed algorithm using Iris dataset

Number of clusters (k)	Priority Dimension (p)	Calculated RandIndex	Structure of formed clusters	Mixing of data points between clusters
3	1	0.5392	Proper	No
3	2	0.1998	Semi-Proper	Just started
3	3	0.9039(Highest)	Improper	Yes
3	4	0.9039(Highest)	Improper	Yes

VII. CONCLUSION

The utility of Mahalanobis distance in k-means for clustering of Gaussian mixture models is evident. It can effectively identify elliptical cluster shapes, hence can replace Euclidean distance measure in traditional k-means. The major issue is proper initialization of the variables used for Mahalanobis computations. A fast initialization technique has been proposed for this purpose. Results show that the proposed method achieves same level of accuracy with much gain in runtime.

REFERENCES

1. Estivill-Castro, Vladimir (20 June 2002). 'Why so many clustering algorithms — A Position Paper'. ACM SIGKDD Explorations Newsletter 4 (1): 65–75. doi:10.1145/568574.568575
2. X. Wu et al. 'Top 10 algorithms in data mining', Knowledge and Information Systems, Volume 14, Issue 1, pp 1-37, January 2008.
3. Igor Melnykova and Volodymyr Melnykov,'On K-means algorithm with the use of Mahalanobis distances', Statistics and Probability Letters 84 (2014) 88–95
4. D.Arthur, and S. Vassilvitskii, 'k-means++: The advantages of careful seeding', in Proceedings of the 18th annual ACM-SIAM symposium on discrete algorithms, pp. 1027–1035, 2007
5. Leonard Kaufman and Peter J. Rousseeuw, 'Finding groups in data : An introduction to cluster analysis', Wiley series in Probability and Statistics,2005
6. M. Emre Celebi and Hassan A. Kingravi, 'Deterministic Initialization of the K-Means Algorithm using Heirarchical Clustering' in International Journal of Pattern Recognition and Artificial Intelligence, 26(7):1250018, 2012
7. T. Su and J. G. Dy, 'In Search of Deterministic Methods for Initializing K-Means and Gaussian Mixture Clustering,' Intelligent Data Analysis, vol. 11, no. 4, pp. 319–338,2007
8. J. MacQueen, 'Some methods for classification and analysis of multivariate observations', In Proc. 5th Berkeley Symp. Mathematical Statistics and Probability, 1967
9. P. S. Bradley and U. Fayyad, 'Refining initial points for k-means clustering', Proceedings of the 15th int. conf. on machine learning, pp. 91–99, 1998.
10. Grigorios Tzortzis and Aristidis Likas 'The MinMax k-Means clustering algorithm' in Pattern Recognition Letters, Volume 47, Issue 7, July 2014, Pages 2505-2516
11. Rehab Duwairi, Mohammed Abu. Rahmeh 'A novel approach for initializing the spherical K-means clustering algorithm' in Simulation Modeling practice and Theory papers, Volume 54, May 2015, Pages 49–63
12. Mohammed El Agha and Wesam M. Ashour 'Efficient and Fast Initialization Algorithm for K-means Clustering' in I.J. Intelligent Systems and Applications, 2012, 1, 21-31 Published Online February 2012 in MECS (http://www.mecs-press.org/) DOI: 10.5815/ijisa.2012.01.03
13. Shiming Xiang, , Feiping Nie, Changshui Zhang, 'Learning a Mahalanobis distance metric for data clustering and classification' in Pattern Recognition Letters, Volume 41, Issue 12, December 2008, Pages 3600–3612
14. Bahadir Durak, 'A Classification Algorithm Using Mahalanobis Distance Clustering Of Data With Applications On Biomedical Data Sets', A Thesis Submitted To The Graduate School Of Natural And Applied Sciences Of Middle East Technical University, 2010.
15. J. Mao and A.K. Jain ' A self-organizing network for hyperellipsoidal clustering (hec)'. Ieee transactions on neural networks, 7(Jan), 1996 ,Pg16–29.
16. Andrea Cerioli in 'K-means Cluster Analysis And Mahalanobis metrics: a problematic match or an overlooked opportunity?' in Statistica Applicata Vol. 17, n. 1, 2005
17. Agostino Tarsitano, 'Mahalanobis metrics for k-means algorithms' presented at Convegno intermedio SIS, Napoli, 2003.
18. D Art, R. Gnanadesikan, J.R Kettenring (1982), ' Data-based metrics for cluster analysis',Utilitas Mathematica, 21A, 75-99.
19. UCI Machine Learning Repository, http://ics.uci.edu/ mlearn/MLRepository.html