A Comparative Study on K Means and PAM Algorithm using Physical Characters of Different Varieties of Mango in India

Bhaskar Mondal
Department of Computer Science and Engineering
National Institute of Technology Jamshedpur
Jamshedpur, India- 831014

J. Paul Choudhury, Ph.D
Department of Information Technology
Kalyani Government Engineering College
Kalyani, West Bengal, India- 741235

ABSTRACT

Clustering is the most important and popular technique for finding pattern and relationships in databases. In this paper a comparative study has been done on the clustering techniques like k-means and k-mediod (PAM) with difference distance measures to classify the different varieties of mango based on physical characters of fruit. As the purity of result of a clustering algorithm depend upon the distance measure technique used in that algorithm we have validate the result using different distance measure also. Classification of agricultural data is still remains a challenge due to its high dimension and noise. This type of study may be helpful for the agricultural research as well as for the field of science and technology.

General Terms

Clustering.

Keywords

Clustering, k-means, k-mediod, PAM, distance.

1. INTRODUCTION

The clustering techniques are proposed for partitioning a collection of data objects $X = \{x_1, x_2, x_3 \dots \dots x_{m-1}, x_m\}$ into k number of subsets or "clusters (C_i) where $\{C_i: 1 \leq i \leq k\}$" so that objects are more closely related to one another in same cluster than objects assigned to different clusters. Grouping is done on the basis of similarities or dissimilarities (distance,d_{ij}) between objects [2]. The number of groups (k) may be user defined and it's an unsupervised technique as no pre classified data is provided as training set. Clustering can be used to discover interesting patterns in the data, or to verify pureness of predefined classes. There are various examples of both these applications in the microarray literature. [1][10]. It is important to have knowledge of difference between clustering and classification. The classification techniques are supervised and some collection of pre-classified data objects should be provided, the problem is to label a newly encountered data records. Typically, the given labeled (training) patterns are used to learn the descriptions of classes which in turn are used to label a new pattern.

The clustering method may roughly divide into two types namely partitioning and hierarchical methods. In partitioning method classes are mutually exclusive each object belonging to one cluster. Each cluster may be represented by a centroid or a cluster representative. *k-means*[4] Each cluster is represented by the center of the cluster and *k-medoids* or PAM (Partition around medoids) [3] Each cluster is represented by one of the objects in the cluster are some hierarchical clustering methods.

On the other hand the hierarchical clustering methods are most commonly used. There are two types of hierarchical methods agglomerative and divisive method. The construction of an agglomerative hierarchical clustering, it repetitively finds a pair of closest points and merges them into one cluster, where a point is either an individual object or a cluster of objects until only one cluster remains. The hierarchy is build up in a series of N-1 agglomerations. The *divisive* methods starts with all objects in a single cluster and at each of N-1 steps divides some clusters into two smaller clusters, until each object resides in its own cluster. The divisive method is less popular one.

The different types of mangos are harvested within a year. One common question is that, how does mango of all corn are categorized by its size? The size of mango is dependent on different parameters like nature of fruit weight, length, breadth, width, stone weight, peel weight, and presentence of pulp[13][14]. Here the comparative study of clustering methods has been done base on physical characters of fruits of different varieties of mango available in gangetic West Bengal. Comparison is made in respect accuracy and ability to handle high dimension of data.

Section II describes preliminaries of the difference distance measures k-means and k-mediod clustering algorithms. The details of the proposed scheme are described in section III. Section IV and V presents the experimental results and security analysis. The conclusion and future scope of proposed scheme are presented in Section VI.

2. PRELIMINARIES

In this paper a comparative study has been done on the clustering algorithm like k-means and k-mediod (PAM) with difference distance measures to classify the agricultural (mango) data set. For measuring distance Euclidean distance, City block (Manhattan) distance, Chebyshev Distance, Minkowski Distance of Order 2 and 3, Bray Curtis (Sorensen) distances are used. The author will like to put the distance measure techniques first as distance calculation is the most important step of clustering algorithm.

2.1. Euclidean distance:

Euclidean distance (Minkowski Distance of Order 2) gives distance between two points on Cartesian coordinate. It calculates distance as root of square differences between coordinates of a pair of objects.

The expiration given by

$$d_{ij} = \sqrt[2]{\sum_{k=1}^{n}\left(x_{ik} + x_{jk}\right)^2}$$

2.2. City block (Manhattan) distance:

Manhattan or Minkowski Distance of Order 1) distance is also known as Manhattan distance, boxcar distance, absolute value distance. It represents distance between points in a city road grid. It examines the absolute differences between coordinates of a pair of objects.

The expiration given by

$$d_{ij} = \sqrt[2]{\sum_{k=1}^{n}|x_{ij} - x_{ij}|}$$

2.3. Correlation Distance:

The correlation between vectors i and j are defined as follows:

$$r_{ij} = \frac{\frac{1}{n}\sum_i x_i y_i - \mu_i \mu_j}{\sigma_i \sigma_j}$$

where μ_i and μ_j are the means of i and j respectively, and σi and σj are the standard deviations of i and j. The numerator of the equation is called the covariance of i and j, and is the difference between the mean of the product of i and j subtracted from the product of the means. Note that if i and j are standardized, they will each have a mean of 0 and a standard deviation of 1, so the formula reduces to:

$$d_{ij} = \frac{1}{n}\sum_i x_i y_i$$

Whereas euclidean distance was the sum of squared differences, correlation is basically the average product. There is a further relationship between the two. If we expand the formula for euclidean distance, we get this:

$$d_{ij} = \sqrt{\sum_{i}^{n}(x_i - y_i)^2} = \sqrt{\sum_i x_i^2 + \sum_i y_i^2 + 2\sum_i x_i y_i}$$

But if i and j are standardized, the sums Σx^2 and Σy^2 are both equal to *n*. That leaves Σxy as the only non-constant term, just as it was in the reduced formula for the correlation coefficient. Thus, for standardized data, we can write the correlation between i and j in terms of the squared distance between them:

$$d_{ij} = 1 - \frac{d^2(i,j)}{2n}$$

2.4. K-Means Algorithm:

K-means has a rich and diverse history as it was independently discovered in different scientific fields by Steinhaus (1955) [6], Lloyd (1957) [7], Ball & Hall (1965) [8] and McQueen (1967) [9]. Even though K-means was first proposed over 50 years ago, it is still one of the most widely used algorithms for clustering. Ease of implementation, simplicity, efficiency, and empirical success are the main reasons for its popularity.

The aim of k-means clustering algorithm is to divide a data set *X* into disjoint clusters by optimizing am objective function

$$E = \sum_{i=1}^{k}\sum_{x \epsilon C_i} d(x, m_i).$$

Here m_i is the center or centriod of cluster C_i, while $d(x, m_i)$ is the distance between a point x and cluster centre m_i. The objective function E attempts to minimize the distance between each object from the cluster center in a cluster. Initially we assign randomly k numbers of cluster centers. Then it starts assigning each records of data set to the cluster whose centre is closest one using some distance measure and recomputed the centers. The process continues until the center of cluster stops changing.

Considering a data set X with m objects i.e. $X = \{x_i : 1 \le i \le m\}$

Step 1: initialize k number of cluster centers randomly based on some prior knowledge.

Step 2: cluster the cluster prototype matrix D (distance matrix of distances between cluster centers and data objects.) of size k x m.

Step 3: Assign each object in the data set to nearest cluster i.e.

$$x_j \in C_n \; if \; d(x_j, C_n) \le d(x_j, C_i) \; \forall \; 1 \le j \le k, j \ne n \; where \; j = 1,2,3, \ldots \ldots m.$$

Step 4: Calculate the average of element of each cluster and change the k cluster centers by their averages.

Step 5: Again calculate the cluster prototype matrix M.

Step 6: Repeat Step 3, 4 and 5 until there is no change for each cluster.

2.5. PAM Algorithm:

PAM uses a k-medoid method to identify the clusters. PAM selects k objects arbitrarily from the data as medoids. In each step, a swap between a selected object O_i and a non-selected object O_h is made as long as such a swap would result in an improvement of the quality of clustering .To calculate the effect of such a swap between O_i and O_h a cost C_{ih} is computed, which is related to the quality of partitioning the non-selected objects to k clusters represented by the medoids. So, at this stage it is necessary first to understand the method of partitioning of the data objects when a set of k-medoids are given

The objective of PAM(Partitioning Around Medoids)[5] is to determine a representative object (medoid) for each cluster, that is, to find the most centrally located objects within the clusters. The PAM algorithm consists of two parts. The first build phase follows the following algorithm:

Step 1: Input: Database of object D.

Step 2: Select arbitrarily k representative objects. Mark these objects as “selected” and mark the remaining as “non-selected”.

Step 3: Repeat until no more objects are to be classified.

a. Assign each remaining object to the cluster of the nearest medoid

b. Do for all selected object O_m.

  i. Do for all non-selected objects O_i.

```
        Compute $C_{mi}$ (cost of swapping)
        End do
End do
   c.  If $C_{imin,hmin}$<0
   d.  Then mark $O_i$ as non-selected and Oh as selected.
Step 4: Go to step 3 until there is no change of mediod.
```

2.6. Determining the Number of Clusters

Automatically determining the number of clusters has been one of the most difficult problems in data clustering. Usually, clustering algorithms are run with different values of *K*; the best value of *K* is then chosen based on a criterion function. Figueiredo and Jain [10] used the minimum message length (MML) criteria in conjunction with the Gaussian mixture model (GMM) to estimate *K*. Their approach starts with a large number of clusters, and gradually merges the clusters if this leads to a decrease in the MML criterion. Gap statistics [11] is another commonly used approach for deciding the number of clusters. The key assumption is that when dividing data into an optimal number of clusters, the resulting partition is most resilient to the random perturbation. Dirichlet Process (DP) [12] introduces a non-parametric prior for the number of clusters. It is often used by probabilistic models to derive a posterior distribution for the number of clusters, from which the most likely number of clusters can be computed. Its key idea is to introduce a non-parametric Bayesian prior for the number of clusters. In spite of these objective criteria, it is not easy to decide which value of *K* leads to more meaningful clusters. 2.6.1(a) shows a 2-dimensional synthetic dataset generated from a mixture of six Gaussian components. The true labels of the points are shown in 2.6.1(e). When a mixture of Gaussians is fit to the data with 2, 5 and 6 components, shown in 2.6.1(b)-(d), respectively, each one of them seems to be a reasonable fit. This data is 2-dimensional, so we can easily visualize and assess how many clusters are good. But, this cannot be done when the data is high dimensional.

3. COMPARATIVE STUDY

We used physical characters of fruits of different varieties of mango to make a comparison study between k-means and PAM algorithms as well as four different distance majors for comparing the efficiency of these clustering algorithms with different distance majors. Brief description is given below with a sample data set:

Table 1. A sample of data set of physical characteristics of mango

Total Weight (g)	**Length (cm)**	**Breadth**	**stone Weight**	**Peel Weight**	**Pulp Weight**
169.83	9.95	5.78	41.00	36.43	92.40
258.33	10.57	7.08	37.57	31.23	189.53
697.57	14.90	9.58	60.73	64.43	572.27
234.40	9.97	6.05	43.23	31.37	159.80
463.63	12.48	7.27	43.50	47.77	372.23
300.43	10.00	7.53	41.10	43.23	229.30

4. EXPERIMENTAL RESULTS

Here we apply k-means and PAM algorithms on characters of fruits of different varieties of mango data set to classify it into thirteen equivalent classes. We use three distance majors separately Euclidean distance, Manhattan distance, Correlation distance.

Results of k-means and PAM using		
Total Number of records in dataset = 195		
Clustering Algorithm	**Correctly Classified**	**Average Accuracy**
K-means with Euclidean distance	165	84.61
K-means with Manhattan distance	174	89.23
K-means with Correlation distance	149	76.41
K-medoids with Eucledian distance	147	75.83
K-medoids with Manhattan distance	165	84.61
K-medoids with Correlation distance	173	88.71

We observe that k-means giving the maximum accuracy when we are using Manhattan distance & PAM algorithm is showing better accuracy with Correlation distance.

5. FUTURE SCOPE & CONCLUSION:

Organizing data into sensible groupings arises naturally in many scientific fields. It is, therefore, not surprising to see the continued popularity of data clustering. While a large number of clustering algorithms have been published and new ones continue to appear, there is no best algorithm. Most algorithms, including the popular K-means, and k-medoids are admissible algorithms. Indeed, the search for a best clustering algorithm is fruitless and contrary to the exploratory nature of clustering. The challenge in data clustering is to incorporate domain knowledge in the algorithm, find appropriate representation and measure of similarity, validate clusters, devise a rational basis for comparing methods, combine ‘multiple looks” of the same data, and develop efficient algorithms for clustering large datasets.

We have tried to obtain accurate results of clustering by using two popular clustering algorithms using three distance metrics (eucledian, manhattan and correlation). From the experimental results it can be concluded that on changing the value of the distance metric, the results of the clustering algorithm changes. The optimum number of clusters in our case is 13.That is if we divide the input data into 13 clusters, it produces the best and most likely division of all the mango varieties. This value has been decided by carefully examining the number of elements in each cluster for the two algorithms k-means and k-medoids using the three distance metrics, euclidean, manhattan and correlation distance.

This approach can be applied for other fruits for its recognition or agricultural result.

6. REFERENCES

[1] Alizadeh A., Eisen M.B, Davis R.E, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature. 2000; 403(6769):503–511.

[2] Guha, S., Rastogi, R., and Shim K. (1998). *CURE: An Efficient Clustering Algorithm for Large Databases*. In Proceedings of the ACM SIGMOD Conference.

[3] L. Kaufman and P. J. Rousseeuw, Finding Groups in Data: an Introduction to Cluster Analysis, John Wiley & Sons, 1990.

[4] MacQueen, J.B. (1967). *Some Methods for Classification and Analysis of Multivariate Observations*. In Proc. Of 5th Berkley Symposium on Mathematical Statistics and Probability, Volume I: Statistics, pp. 281–297.

[5] Nielsen T.O, West R.B, Linn S.C, et al. Molecular characterisation of soft tissue tumours: a gene expression study. Lancet2002.

[6] Steinhaus, 1956] STEINHAUS, H. 1956. Sur la division des corp materiels en parties. *Bulletin of acad. polon. sci.*

[7] Lloyd, 1982] LLOYD, S. 1982. Least squares quantization in PCM. *Ieee transactions on information theory.*

[8] Ball & Hall, 1965] BALL, G., & HALL, D. 1965. *ISODATA, a novel method of data anlysis and pattern classification*. Tech. rept. NTIS AD 699616. Stanford Research Institute, Stanford, CA.

[9] MacQueen, 1967] MACQUEEN, J. 1967. Some methods for classification and analysis of multivariate observations.

[10] Figueiredo & Jain, 2002] FIGUEIREDO,MARIO, & JAIN, ANIL K. 2002. Unsupervised learning of finite mixture models.

[11] Tibshirani *et al.* , 2001] TIBSHIRANI, R.,WALTHER, G., , & HASTIE, T. 2001. Estimating the number of clusters in a data set via the gap statistic. *Journal of the royal statistical society.*

[12] Ferguson, 1973] FERGUSON, THOMAS S. 1973. A bayesian analysis of some nonparametric problems. *Annals of statistics.*

[13] J. Paul Choudhury, Satyendra Nath Mandal, Dilip Dey, S.R. Bhadra Choudhury, Growth Estimation with Simulated Annealing considering weather parameters using Factor and Principal Component Analysis, Proceedings of National Conference on Methods & Models in Computing, Department of Computer and System Sciences, Jawaharlal Nehru University, New Delhi, pp 184-197, December 2007

[14] J. Paul Choudhury, Satyendra Nath Mandal, Dilip Dey, S.R. Bhadra Choudhury, A Framework to Predict Size of Different Types Of Mango Considering Effect of Different Parameters Using Factor and Principal Component Analysis , Proceedings of International Journal IJITKM, Department of Computer and System Sciences, Kurukshetra university, Vol-1, Number-2,Page No.303-309,December 2008.

7. AUTHOR'S PROFILE

Bhaskar Mondal was born in a country side village of West Bengal, India in 1986. He received B. Tech. degree in Computer Science and Engineering from West Bengal University of Technology in 2008 and M. Tech. degree in Computer Science and Engineering from Kalyani Government Engineering College, West Bengal, India in the year of 2010.

He is working at National Institute of Technology, Jamshedpur as Assistant Professor in the department of Computer Science and Engineering since January 2011. His research interest includes Secret Image Sharing, Security and Data Mining.

Dr. J Paul Choudhury (Jagannibas Paul Choudhury) completed Bachelor of Electronics and Tele-Communication Engineering (Hons.) from Jadavpur University, Kolkata, M. Tech. in Electronics and Electrical Engineering with specialization of Control and Automation Engineering from Indian Institute of Technology Kharagpur and thereafter completed Ph. D.(Engg.) from Jadavpur University, Kolkata. At present Dr. J Paul Choudhury is with the Departmrent of Information Technology, Kalyani Government Engineering College, Kayani, West Bengal, India, and he has more than 70 publications in different National and International Journals and Conference Proceedings. His field of interest is Soft Computing, Data Base and Data Mining, Object Oriented Methodology.