ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

Functional Efficiency of Hybrid Data Clustering Algorithms for Various Medical Datasets

Dr.M.Vanitha
Assistant Professor
Department of Computer Science and Engineering
Alagappa University
Karaikudi.- 630 003

Abstract: The past few years have seen a tremendous increase of data acquisition in different scientific fields such as molecular biology, bioinformatics or biomedicine. Therefore, novel methods are needed for automatic data processing and analysis of this large amount of data. "Data mining" is the process of applying methods like clustering or classification to large databases in order to uncover hidden patterns. Clustering is the task of partitioning points of a data set into distinct groups in order to minimize the intra cluster similarity and to maximize the inter cluster similarity. In contrast to unsupervised learning like clustering, the classification problem is known as supervised learning that aims at the prediction of group membership of data objects on the basis of rules learned from a training set where the group membership is known. In this paper an attempt is tried to using the familiar Partition clustering methods like K-Means, K-Medoids and Fuzzy C-Means for finding a Hybrid algorithm which could analyze the performance of the medical data. The suggested Hybrid algorithm may be utilized for analyzing the performance evaluation of different UCI Repository datasets like Tumor, Liver Disorder and Lung Cancer data on the basis of Validity Measure and Cost Measure.

Keywords: K-Means, K-Medoids, Fuzzy C-Means, Euclidean Distance, UCI Machine Repository Dataset, Validity Indices

1. Introduction

Data Mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Clustering is an important area of application for a variety of fields including data mining, knowledge discovery, statistical data analysis, data compression and vector quantization. Clustering has been formulated in various ways in machine learning, pattern recognition, optimization and statistics literature. Clustering is the most common form of unsupervised learning. According to the rule of the unsupervised learning, clustering does not require supervision. No supervision means that there is no human expert who has assigned documents to classes. In clustering, it is the distribution and makeup of the data that will determine cluster membership. The notion of what constitutes a good cluster depends on the application and there are many methods for finding clusters subject to various criteria.

In a partitioned algorithm, given a set of n data points in real d-dimensional space, and an integer k, the problem is to determine a set of k points in Rd, called centers, so as to minimize the mean squared distance from each data point to its nearest center. This measure is often called the squared-error distortion and this type of clustering falls into the general category of variance based clustering. The *k*-means problem is to find cluster centers that minimize the intra-class variance, i.e. the sum of squared distances from each data point being clustered to its cluster center (the center that is closest to it). Although finding an exact solution to the *k*-means problem for arbitrary input is NP-hard the standard approach to finding an approximate solution (often called Lloyd's algorithm or the *k*-means algorithm) is used widely and frequently finds reasonable solutions quickly.

The basic strategy of K-Mediods clustering algorithms is to find k clusters in n objects by first arbitrarily finding a representative object (the Medoids) for each cluster. Each remaining object is clustered with the Medoid to which it is the most similar. K-Medoids method uses representative objects as reference points instead of taking the mean value of the objects in each cluster. The algorithm takes the input parameter k, the number of clusters to be partitioned among a set of n objects.

Fuzzy c-means [1, 2] is an extension of k-means clustering. The major difference between the fuzzy c-means and k-means is that the later discovers hard clusters where a particular sample can belong to only one cluster while the former discovers soft clusters where a particular sample can belong to more than one cluster with certain probability. This belongingness of a data sample to the cluster is represented using membership values.

2. Pertinent Earlier Works

Mahendra Tiwari1 Randhir Singh Cluster to analysis is the organization of a collection of patterns in to clusters based on similarity. Intuitively, patterns within a valid

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

cluster are more similar to each other than they are to a pattern belonging to a different cluster [1].

Pavel Berkhin Clustering [2] is a division of data into groups of similar objects. Representing the data by fewer clusters necessarily loses certain fine details, but achieves simplification. It models data by its clusters. Data modeling puts clustering in a historical perspective rooted in mathematics, statistics, and numerical analysis. From a machine learning perspective clusters correspond to hidden patterns, the search for clusters is unsupervised learning, and the resulting system represents a data concept.

Anoop Kumar Jain and Satyam Maheswari [3] In centroid-based clustering, clusters are represented by a central vector, which may not necessarily be a member of the data set. When the number of clusters is fixed to k, k-means clustering gives a formal definition as an optimization problem: find the k cluster centers and assign the objects to the nearest cluster center, such that the squared distances from the cluster are minimized.

L. Wanner, [4] Different clustering methods can be classified into various categories such as partitioning based methods, hierarchical methods, grid-based methods, density-based methods, model-based methods, methods for high dimensional data and constraint-based clustering. Among all these methods, this paper is aimed to explore two methods – k-means and k-medoids – which are partitioning based clustering methods. These methods are discussed along with their algorithms, strength and limitations.

Bradley. P.S, Fayyad. U. M and Reina. C. A, Scaling, [5] Partitional clustering algorithms divide the whole data set into a set of disjoint clusters directly. These algorithms attempt to determine an integer number of clusters that optimise a certain objective function. The process for optimization of objective function is an iterative procedure to get local or global optimizes value. To classify the various types of cancer into its different subcategories, different data mining techniques have been used over gene expression data.

Han J. and M. Kamber, [6] K-medoids is a clustering algorithm that is related to the k-means algorithm . The k-medoids is a partitioning algorithm that divides the data set up into separate clusters. The algorithm attempts to minimize the squared error, which is the distance between points in the cluster and a point that is designated as the center (medoid) of the cluster. A medoid is considered as an object of a cluster whose average dissimilarity to all the objects in a cluster is minimal.

T.Velmurugan and T.Santhanam [7] K-Means is one of the simplest unsupervised learning methods among all partitioning based clustering methods. It classifies a given set of *n* data objects in *k* clusters, where *k* is the number of desired clusters and it is required in advance. A centroid is defined for each cluster. All the data objects are placed in a cluster having centroid nearest (or most similar) to that data object. After processing all data objects, k-means, or centroids, are recalculated, and the entire process is repeated. All data objects are bound to the clusters based on the new centroids.

A. K. Jain, M. N. Murty, [8] new algorithm for K-medoids clustering which runs like the K-means clustering. The algorithm has excellent feature that it requires the distance between every pairs of objects only once and uses this distance at every iterative step. The result from various simulations shows that the proposed method has better performance than K-means clustering and it takes the less computation time than PAM.

3. Proposed work

3.1 Cluster Analysis

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters).

3.2 K-Means Algorithm

K-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. [3]

$$E = \sum_{i=0}^{k} \sum_{p \in C_i} |p - m_i|^2$$

Input: k is the number of clusters, D is input data set

Output: k clusters.

1. Randomly choose k objects from D as the initial cluster centers.
2. Repeat
3. Assign each object from D to one of k clusters to which the object is most similar based on the mean value of the objects in the cluster.
4. Update the cluster means by taking the mean value of the objects for each of k cluster.
5. Until no change in cluster means/ min error E is reached.

3.3 K-Medoids Algorithm

The very popular K-Means algorithm is sensitive to outliers since an object with an extremely large value may substantially distort the distribution of data. How might the algorithm be modified to diminish such sensitivity? Instead

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

of taking the mean value of the objects in a cluster as a reference point, a Medoid can be used, which is the most centrally located object in a cluster [4][5].

A typical K-Mediods algorithm for partitioning based on Medoid or central objects is as follows:

Input: k: The number of clusters
D: A data set containing n objects
Output: A set of k clusters that minimizes the sum of the dissimilarities of all the objects to their nearest medoid.
Method: Arbitrarily choose k objects in D as the initial representative objects;
Repeat assign each remaining object to the cluster with the nearest medoid;
Randomly select a non medoid object O_{random};
Compute the total points S of swaping object O_j with O_{random};
if S < 0 then swap O_j with O_{random} to form the new set of k medoid;
Until no change;

3.4 Fuzzy C-Means Algorithm

In hard clustering, data is divided into distinct clusters, where each data element belongs to exactly one cluster. In **fuzzy clustering** (also referred to as **soft clustering**), data elements can belong to more than one cluster, and associated with each element is a set of membership levels. These indicate the strength of the association between that data element and a particular cluster. Fuzzy clustering is a process of assigning these membership levels, and then using them to assign data elements to one or more clusters.

The FCM aims to minimize an objective function. The standard function is:

$$u_k(x) = \frac{1}{\sum_j \left(\frac{d(center_k, x)}{d(center_k, x)} \right) 2/(m-1)}$$

Input: K-the Number of Clusters to be partitioned;
N- the Number of points

Output: Find Similar and Dissimilar cluster groups with objective function.

1. Choose a number of clusters in a given image.
2. Assign randomly to each point coefficients for being in a cluster.
3. Repeat until convergence criterion is met.
4. Compute the center of each cluster.
5. For each point, compute its coefficients of being in the cluster [6-7].

3.5 Hybrid Clustering Algorithm

Input: 'K' the Number of Clusters to be partitioned; 'N', the Number of points

Output: A Set of 'K' Clusters that Minimize the Sum of Similarities and Dissimilarities of all the points to the Nearest Medoids.

Given the data set X, choose the number of clusters $1 < c < N$.

Initialize with random cluster centers chosen from the data set *X*.

Repeat for $l = 1; 2; : : :$

Step 1 Compute the distances

$$D_{ik}^2 = \left(x_k - v_i\right)^T \left(x_k - v_i\right), \quad 1 \le i \le c, \; 1 \le k \le N.$$

Step 2 Select the points for a cluster with the minimal distances, they belong to that cluster.

Step 3 Calculate fake cluster centers.

$$v_i^{(l)*} = \frac{\sum_{j=1}^{N_i} x_i}{N_i}$$

Step 4 Choose the nearest data point to be the cluster center

$$D_{ik}^{2*} = \sqrt{\left(\left(x_k - v_i^*\right)^T \left(x_k - v_i^*\right)\right)^2},$$

and

$$x_i^* = argmin_i \left(D_{ik}^{2*}\right); v_i^{(l)} = \min\left(x_i^*\right)$$

Until

$$\prod_{k=1}^{n} max \left| v^{(l)} - v^{(l-1)} \right| \neq 0$$

Ending Calculate the partition matrix

4. Methodology

4.1 Euclidean Distance Measure

In mathematics [8], the Euclidean distance or Euclidean metric is the "ordinary" distance between two

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

points that one would measure with a ruler, and is given by the Pythagorean formula. By using this formula as distance, Euclidean space (or even any inner product space) becomes a metric space. The associated norm is called the Euclidean norm. Older literature refers to the metric as Pythagorean metric.

The Euclidean distance, data vector p and centroid q is computed as

$$d(p,q) = \sqrt{\sum_{k=1}^{n} (q_{ik} - p_{ik})^2}$$

4.2 Validity Indices

Cluster validity refers to the problem whether a given fuzzy partition fits to the data all. The clustering algorithm always tries to find the best fit for a fixed number of clusters and the parameterized cluster shapes. However this does not mean that even the best it is meaningful at all. Either the number of clusters might be wrong or the cluster shapes might not correspond to the groups in the data, if the data can be grouped in a meaningful way at all. [9]

4.2.1. Partition Index (SC):

Partition Index is the ratio of the sum of compactness and separation of the clusters. It is a sum of individual cluster validity measures normalized through division by the fuzzy cardinality of each cluster.

$$SC(c) = \sum_{i=1}^{c} \frac{\sum_{j=1}^{N} \left(\mu_{ij}\right)^m \left\| x_j - v_i \right\|^2}{N_i \sum_{k=1}^{c} \left\| v_k - v_i \right\|^2}$$

SC is useful when comparing different partitions having equal number of clusters. A lower value of *SC* indicates a better partition.

4.2.2. Separation Index (S):

Separation Index on the contrary of partition index (SC), the separation index uses a minimum-distance separation for partition validity.

$$S(c) = \frac{\sum_{i=1}^{c} \sum_{j=1}^{N} \left(\mu_{ij}\right)^2 \left\| x_j - v_i \right\|^2}{N \min_{i.j} \left\| x_j - v_i \right\|^2}$$

4.2.3. Xie and Beni's Index (XB):

It aims to quantify the ratio of the total variation within clusters and the separation of clusters.

$$XB(c) = \frac{\sum_{i=1}^{c} \sum_{j=1}^{N} \left(\mu_{ij}\right)^m \left\| x_j - v_i \right\|^2}{N \min_{i.j} \left\| x_j - v_i \right\|^2}$$

The optimal number of clusters should minimize the value of the index.

4.3 Cost Measure

Definition of cost function: The cost function is defined for all possible output vectors and all positive input price vectors w = (w1, w2, . . ., wn). An output vector, y, is producible if y belongs to the effective domain of V(y), i.e,

$$DomV = \{y \in R_{+}^{m} : V(y) \neq \emptyset\}$$

The cost function does not exist it there is no technical way to produce the output in question. The cost function is defined by

$$C(y,w) = \min_{x} \{wx : X \in V(y)\}, y \in DomV, w > 0,$$

or in the case of a single output

$$C(w,y) = \min_{x} \{wx : f(x) \geq y\}$$

The cost function exists because a continuous function on a nonempty closed bounded set achieves a minimum in the set.

5. Experiments and Results

5.1 Datasets

To analysis the performance of K-Means, K-Medoids and FCM, the following familiar UCI Datasets have been processed:

Table 1: Dataset with Description

Dataset	Number of Objects	Number of Attributes
Pima Indian Diabetes	768	8
Lung Cancer	25	9
Liver Disorder	345	7

5.2 Results

Dataset 1: Pima Indian Diabetes

Table 2: Validity Indices for Pima Indian Diabetes

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

Algorithm	SC	S	XB
K-Means	0.7265	0.0014	3.1784
K-Medoids	0.7237	0.0014	Inf
FCM	0.9548	0.0018	2.6115
Hybrid	0.9658	0.0019	2.6321

Dataset 2: Lung Cancer

Table 3: Validity Indices for Lung Cancer

Algorithm	SC	S	XB
K-Means	0.7260	0.0348	1.4609
K-Medoids	1.0597	0.0424	Inf
FCM	2.3320	0.0933	1.0467
Hybrid	2.3419	0.0933	1.0718

Dataset 3: Liver Disorders

Table 4: Validity Indices for Liver Disorders

Algorithm	SC	S	XB
K-Means	0.5784	0.0021	3.2074
K-Medoids	0.6210	0.0018	Inf
FCM	0.7407	0.0021	2.9915
Hybrid	0.7418	0.0021	3.1871

Cost Measure

The solution to the cost minimization problem 2 is a vector x which depends on output vector y and the input vector w. To denote this solution by x(y,w). This demand for inputs at for a fixed level of output and input prices is often called a Hicksian demand curve.

Table 4: Cost measure for Medical Datasets

	Pima Indian Diabetes		Lung Cancer		Liver Disorder	
	Cost		Cost		Cost	
	Min	Max	Min	Max	Min	Max
K-Means	12.02	29.74	10.07	20.34	40.47	54.63
K-Medoids	11.30	14.04	17.98	19.03	42.45	56.86
FCM	10.87	14.36	16.99	20.13	42.45	55.36
Hybrid	9.99	14.54	16.88	19.74	42.45	53.95

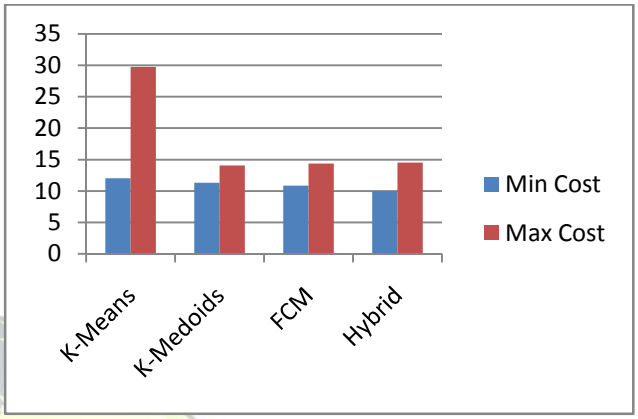

Figure 5.1 Cost Measurements for Pima Indian Diabetes Dataset

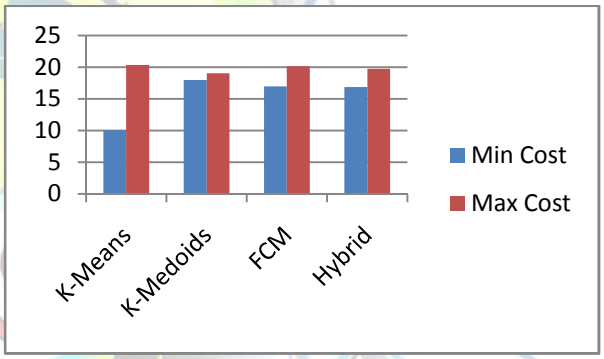

Figure 5.2 Cost Measurements for Lung Cancer Dataset

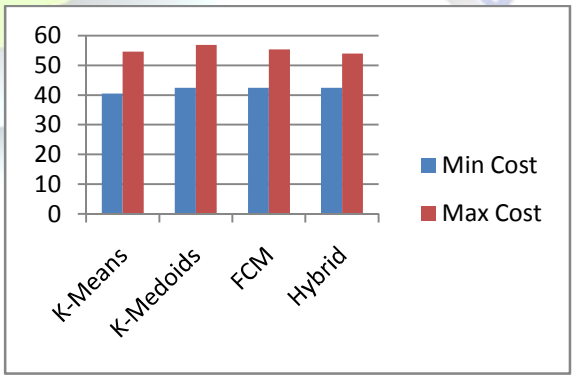

Figure 5.3 Cost Measurements for Liver Disorder Dataset

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

6. Conclusion and Future work

Cluster analysis is one of the major tasks in various research areas. The clustering aims at identifying and extract significant groups in underlying data. In the proposed system to hybrid three major clustering algorithms such as K-Means, K-Medoids, Fuzzy C-Means with Euclidean distance measure by using UCI dataset. It illustrates the efficiency of each clustering algorithm with its validity measures and cost measure. It shows the Hybrid clustering algorithm had better than other clustering algorithms. The experimental result shows the performance of the Hybrid algorithm is improved significantly. The advantage of k-means is its low computation cost, while drawback is sensitivity to noisy data and outliers. Compared to this, k-medoid is not sensitive to noisy data and outliers, but it has high computation cost. Both FCM algorithm and Hybrid approach will produce slighter differences but hybrid method produce high performance and low computation cost when using the large unsupervised data sets.

It is desirable to overcome the drawback of sensitive to noisy data in future studies, however in this present proposed K-Medoid and hybrid algorithms, there is a low computation cost and not sensitive to noisy data .

References

[1]Mahendra Tiwari1 Randhir Singh2, "Comparative Investigation of K-Means and K-Medoid Algorithm on Iris Data".

[2] Pavel Berkhin ,"Survey of Clustering Data Mining Techniques" ,Accrue Software, Inc

[3] Anoop Kumar Jain1 and Satyam Maheswari,"Survey of Recent Clustering Techniques in Data Mining", International Archive of Applied Sciences and Technology,Volume 3 [2] June 2012: 68 – 75

[4] L. Wanner, "Introduction to Clustering Techniques", International Union of Local Authorities, July, 2004.

[5] Bradley. P.S, Fayyad. U. M and Reina. C. A, Scaling, "Clustering Algorithms to Large Databases*", Proc. of the 4th International Conference on Knowledge Discovery & Data Mining (KDD98), AAAI Press, Menlo Park,* CA, 1998, pp. 9-15.

[6] Han J. and M. Kamber, Data Mining: Concepts and Techniques, *Morgan Kaufmann Publishers*, Second Edition, New Delhi, 2006. ISBN : 978-81-312-0535-8

[7] T.Velmuruganand T.Santhanam, "Computational complexity between k-means and k-mediods clustering algorithms for narmal and uniform distributions of data points".

[8] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review". ACM Computing Surveys, Vol .31No 3,pp.264–323, 1999.

[9] Susana. A, Leiva-Valdebenito, Francisco. J and Torres-Aviles, "A Review of the Most Common Partition Algorithms in Cluster Analysis: A Comparative Study", *Colombian Journal of Statistics*, ISSN: 0120-1751, Vol. 33, No. 2, Dec. 2010, pp. 321- 339.

[10] Park, H.S., J.S. Lee and C.H., "A k-Means-Like Algorithm for k-Medoids Clustering and Its Performance", *Department of Industrial and Management Engineering, POSTECH*, South Korea, Jun, 2009.

[11] Matheus C.J.; Chan P.K.; and Piatetsky-Shapiro G. 1993. Systems for Knowledge Discovery in Databases, IEEE Transactions on Knowledge and Data Engineering 5(6): 903-913.

[12] K. Alsabti, S. Ranka, and V. Singh, ªAn Efficient k-means Clustering Algorithm,º Proc. First Workshop High Performance Data Mining, Mar. 1998

[13] Jiawei and Micheline Kamber, "Data Mining Concepts and Techniques" University of Illinois at Urbana-Champaign 2000© Morgan Kaufmann Publishers.

[14] T. Velmurugan,and T. Santhanam, "A Survey of Partition based Clustering Algorithms in Data Mining: An Experimental Approach" An experimental approach.

[15] L. Kaufman and P. J. Rousseeuw, "Finding Groups in Data: An Introduction to Cluster Analysis," Wiley, New York, December 1990.

[16] J. Han, M. Kamber and A. Tun, "Spatial clustering methods in data mining: A survey," In H. J. Miller & J. Han (Eds.), *Geographic data mining and knowledge discovery.* Taylor and Francis, 2001.