Evolving Efficient Clustering Patterns in Liver Patient Data through Data Mining Techniques

Pankaj Saxena
Reader
RBS Management Technical Campus, Agra

Vineeta Singh
Professor,ISS
DR. B. R. Ambedkar University, Agra

Sushma Lehri
Professor, IET
DR. B. R. Ambedkar University, Agra

ABSTRACT

Clustering is one of the most important research areas in the field of data mining. In simple words, clustering is a division of data into different groups. Data are grouped into clusters in such a way that data of the same group are similar and those in other groups are dissimilar. It aims to minimize intra-class similarity while to maximize interclass dissimilarity. Clustering is an unsupervised learning technique. Clustering is useful to obtain interesting patterns and structures from a large set of data. Clustering can be applied in many areas, such as marketing studies, DNA analysis, city planning, text mining, and web documents classification. Large datasets with many attributes make the task of clustering complex. Many methods have been developed to deal with these problems. In this paper, two well known partitioning based methods – k-means and k-medoids are compared over health data. This paper also proposes an improved k-means medoids clustering algorithm. The proposed algorithm is evaluated using the health dataset i.e Liver dataset and compare the results with other previous algorithms. The proposed algorithm is more effective in terms of computation time as compared to K means and K-medoids clustering algorithm. The algorithms under consideration, is evaluated with Rand Index, Jaccard Coefficient, Folkes and Mallows and Run Time as four metrics. Experimental results are obtained on WEKA, a data mining tool.

General Terms

Clusterization, k-means, k-medoids,WEKA Tool

Keywords

Rand index (RI), Jaccard Coefficient, Folkes and Mallows (FM) index, Silhouette Index

1. INTRODUCTION

Clustering is a division of data objects into groups of similar objects. Such groups are called clusters. Objects possessed by same cluster tend to be similar, while dissimilar objects are possessed by different clusters.These clusters represent groups of data and provide simplification by representing many data objects by fewer clusters. And, this helps to model data by its clusters. Clustering is a method of unsupervised learning and a well known technique for statistical data analysis. It is used in many fields such as machine learning, image analysis, pattern recognition, outlier detection, and health informatics. Various researchers have proposed different methods to achieve clustering. Along with managing a very large dataset, a robust clustering method must satisfy some requirements such as scalability, dealing different types of attributes, discovering clusters of arbitrary shape, high dimensionality, ability to deal with noise and outliers, interpretability and usability. With clustering, time complexity increases with large number of dimensions and large set of data objects. Also the effectiveness depends upon the definition of similarity (or dissimilarity) among objects. Along with this, the output of clustering can be interpreted in different ways. 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 (PAM) – which are partitioning based clustering methods and proposes new Efficient Distance Based Improved K- Medoids Clustering algorithm.

The rest of the paper is organised as follows: section 2 gives related work in clustering. Section 3 includes the description of used dataset. Section 4 describes the proposed Improved K-Medoids Clustering Algorithm. Section 5 deals with performance comparison of various previous algorithms with some simulation results.

2. LITERATURE REVIEW

K-Means[2][5][6] is one of the simplest unsupervised non-hierarchical learning methods among all partitioning based clustering methods. It classifies[7][15] a given set of *n* data objects in *k* clusters, where *k* is the number of desired clusters and it is required in advance. But K-Means has some limitations like, it is applicable only when the mean of a cluster is defined, not applicable to categorical data. It is unable to handle noisy data and outliers. To solve this problem, k-medoids clustering method [4][11]is used where representive objects are called medoids instead of centroids because it is based on most centrally located object in a cluster. But it is relatively more costly as its complexity is $O(i k (n-k)2)$, where i is the total number of iterations, *k* is the total number of clusters, and *n* is the total number of objects.

Singh, S.S and Chauhan, N. C., [16] have compared two well known partitioning based methods – k-means and k-medoids. 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.

Partitioning Around Medoids (PAM) proposed by Kaufman and Rousseeuw [8] is known to be the most powerful. PAM replaces all non-representative objects with representative objects randomly, it considers all the objects in the dataset and hence increases the computational time.

Bala Sundar V *et al* [1] proposed technique that classifies the group of the objects based on attributes into K number of groups. The grouping is done by minimizing the sum of squares of distances between data using Euclidean distance formula and the corresponding cluster centroid.

In our proposed algorithm the initial medoids are not chosen randomly, rather distance matrix using Manhattan distance is used. For evaluating the algorithms under consideration, we used Rand Index, Jaccard Coefficient, Folkes - Mallows and Run Time as four metrices.

3. DATA SET USED

ILPD (Indian Liver Patient Dataset) Data Set is used. This data set contains 416 liver patient records and 167 non liver patient records. The data set was collected from north east of Andhra Pradesh, India. Selector is a class label used to divide into groups (liver patient or not). This data set contains 441 male patient records and 142 female patient records. This secondary data is collected from UCI repository [3], available on the web (http://archive.ics.uci.edu/ml/darasets.html), While Primary data is collected from Bisariya Pathological Lab, Etah, having same attributes and 2000 records.

Attribute Information

1. Age → Age of the patient
2. Gender → Gender of the patient
3. TB → Total Bilirubin
4. DB → Direct Bilirubin
5. Alkphos →Alkaline Phosphotase
6. Sgpt → Alamine Aminotransferase
7. Sgot → Aspartate Aminotransferase
8. TP → Total Protiens
9. ALB → Albumin
10. A/GRatio → Albumin and Globulin Ratio
11. Selector field used to split the data into two sets (labeled by the experts)

4 ALGORITHMS AND IMPLEMENTATIONS

4.1 K-MEANS

K-Means [4] 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.In each iteration centroids change their location step by step. In other words, centroids move in each iteration.This process is continued until centroid remains unchanged. As a result, *k* clusters are found representing a set of *n* data objects. An algorithm for k-means method is given below:

Input : '*k*', the number of clusters to be partitioned; '*n*', the number of objects.

Output*:* A set of '*k*' clusters based on given similarity function.

Algorithm

Steps:

i) Arbitrarily choose '*k*' objects as the initial cluster centers;

ii) Repeat,

a. (Re)assign each object to the cluster to which the object is the most similar; based on the given similarity function;

b. Update the centroid (cluster means), i.e., calculate the mean value of the objects for each cluster;

iii) Until no change

Weaknesses of K-Means:

- Applicable only when the mean of a cluster is defined; not applicable to categorical data.
- Need to specify *k*, the total number of clusters in advance.
- Not suitable to discover clusters with non-convex shape, or clusters of very different size.
- Unable to handle noisy data and outliers.
- May terminate at local optimum.
- Result and total run time depends upon initial partition.

4.2 K-MEDOIDS CLUSTERING

The k-means method uses centroid to represent the cluster and it is sensitive to outliers. This means, a data object with an extremely large value may disrupt the distribution of data. K-medoids method [4][12]overcomes this problem by using medoids to represent the cluster rather than centroid. A medoid is the most centrally located data object in a cluster. Here, k data objects are selected randomly as medoids to represent k cluster and remaining all data objects are placed in a cluster having medoid nearest (or most similar) to that data object. After processing all data objects, new medoid is determined which can represent cluster in a better way and the entire process is repeated. Again all data objects are bound to the clusters based on the new medoids. In each iteration, medoids change their location step by step. Or in other words, medoids move in each iteration. This process is continued until no any medoid move. As a result, *k* clusters are found representing a set of n data objects. It follows the principle of minimizing the sum of dissimilarities between each object and its corresponding reference point. An algorithm for K-Medoids method is given below:

Input : k: the number of clusters. D: a data set containing n objects.

***Output* :** A set of k clusters.

Algorithm

Steps

i). Randomly choose k objects in D as the initial representative objects;

ii). for all objects in the data set D

a. Find the cluster C which is nearest to object i by using the dissimilarity measure;

b. assign object i to cluster C;

c. set the member object in cluster C having minimum intra cluster variance as new centroid of C

iii). Display statistics of clusters obtained.

The following weakness of existing algorithm prompted to develop new algorithm.

Weaknesses of K-Medoids

- Relatively more costly; complexity is *O(i k (n-k)2)*, where i is the total number of iterations, k is the total number of clusters, and *n* is the total number of objects.
- Relatively not so much efficient.
- Need to specify *k*, the total number of clusters in advance.

4.3 PROPOSED ALGORITHM

The proposed algorithm is an enhancement of K-Medoids algorithm (Partition Around Medoid). Here initial centroids are not chosen randomly. The distance matrix is found once and its sum values of objects are computed and sorted in ascending order. First k (which is given) sorted objects are chosen as initial objects. Each remaining object is clustered with the representative object to which it is most similar. For the remaining iterations, the object with minimum distance with other object in its cluster is found and assigned as a new object. Here maximum execution time is reduced, as representative objects are reduced by non representative object with in cluster only. This iterative process is continued until there is no change in medoids. This algorithm is tested with different values of k and their maximum Silhouette index cluster is calculated and compared, to find the optimal number of clusters.

Algorithm

Input

D: A dataset containing n objects with p attributes.

Output

A set of K clusters

Method

The proposed algorithm consists of following steps :

Step 1 Initialize k (which is already provided)

Step 2 Calculate distance between every pair of objects by distance matrix using Manhattan measure as follows:

$$dij = \left(\sum_{a=1}^{p} |x_{ia} - x_{ja}|\right) \qquad i = 1, \ldots\ldots n;\ j = 1, \ldots\ldots, n$$

Step 3 Calculate pij to make the initial guess at the centre of clusters

$$p_{ij} = \frac{d_{ij}}{\sum_{i=j}^{n} d_{ij}} \quad i = 1, \ldots\ldots\ldots, n;\ j = 1, \ldots\ldots, n$$

Step 4 Compute the sum value $\sum_{i=1}^{n} p_{ij}$ $(j = 1 \ldots\ldots, n)$ at each object and sort them in ascending order.

Step 5 Select k objects having minimum value as initial representative object.

Step 6 Assign each remaining object to cluster with the nearest representative object

Step 7 Calculate the cost, that is sum the distances of all objects to their medoids.

Step 8 Calculate Silhouette Index for the newly formed cluster.

Step 9

1) Find new medoid, for each clustered objects by finding minimum distance object among cluster objects using Manhattan distance in distance matrix.
2) Assign minimum distance object as a new representative object, cluster remaining objects and compute cluster cost.
3) Calculate the difference between current and previous cost.

Step 10. Stop, if the current cost is same as previous cost otherwise go to step 9.

The algorithm is tested with different values of K and their maximum Silhouette index cluster is calculated and compared. The Silhouette Index value which is higher for k is considered the correct number of clusters for the given dataset.

5 .EXPERIMENT AND DATA ANALYSIS

5.1 SILHOUETTE INDEX

Silhouette Index is a cluster validity index that is used to judge the quality of any clustering solution C. Here, **a** represents the average distance of a point from the other points of its cluster, and **b** represents the minimum of the average distances of the point from the points of the other clusters.

Let $X = \{x_1, \ldots\ldots, x_n\}$ be the dataset and let C = $(C_1, \ldots, C_k)$ be its clustering into k clusters. Let $d(x_k, x_1)$ be the distance between x_k and x_1. Let $C_j = \{x_{j1}, \ldots, x_{jm}\}$ be the j^{th} cluster, j=1, …,k, where $m_j = |C_j|$. The average distance a_{ji} between the i^{th} vector in the cluster C_j and the other vectors in the same cluster is given by Equation (1).

$$a_{ji} = \frac{1}{m_j - 1} \sum_{\substack{k=1 \\ k \neq i}}^{m_j} d(x_{ji}, x_{jk}) \quad i = 1, \ldots, m_j \qquad \ldots(1)$$

The minimum average distance between the i^{th} vector in the cluster C_j and all the vectors clustered in the clusters C_k, k = 1, …, k, k ≠j is given by the Equation (2).

$$b_{ji} = \min\left\{\frac{1}{m_n} \sum_{k=1}^{m_n} d(x_{ji}, x_{jk})\right\}, \quad i = 1, \ldots, m_j \qquad \ldots.(2)$$

n = 1, …, k ; n ≠j

Then the Silhouette width of the i^{th} vector in the cluster C_j is defined in the following way,

$$S_{ji} = \frac{b_{ji} - a_{ji}}{\max\{a_{ji}, b_{ji}\}} \qquad \ldots. (3)$$

From Equation (1), it follows that $-1 \leq S_{ji} \leq 1$. We can now define the Silhouette of the cluster C_j as:

$$S_j = \frac{1}{m_j} \sum_{i=1}^{m_j} S_{ji} \qquad \ldots. (4)$$

Finally, the global Silhouette Index of the clustering is given by:

$$S = \frac{1}{k} \sum_{j=1}^{k} S_j \qquad \ldots.. (5)$$

Silhouette Index S is the average Silhouette width of all the data points and it reflects the compactness and separation of clusters. It was calculated using Equation (5).

The value of Silhouette Index varies from – 1 to 1 and a higher value indicates better clustering result.

Distance based K-medoids clustering was tested with various health datasets with different values of k. For each k value, the received cluster output is validated using Silhouette Index. The Silhouette Index value which is higher for k is considered the correct number of clusters for the given dataset. Table 1 represents the received Silhouette Index for the ILPD (Indian Liver Patient Dataset) dataset for different values of k from 2 to 12. For this dataset, the maximum Silhouette value of 0.4912 was obtained at k = 4. For other values of k, the Silhouette Index is less. So, it clearly concluded that this dataset can be divided into four clusters.

Table 1. Number of Clusters and Silhouette Index for Indian Patient Liver Dataset (IPLD)

Number of Clusters	Mean Silhouette Index
2	.3667
3	.4243
4	.4912
5	.3578
6	.4011
7	.3267
8	.3110
9	.2534
10	.2278
11	.2167
12	.2312

5.2 METRICS USED FOR EVALUATION

In order to measure the performance of a clustering and classification system, a suitable metric will be needed. For evaluating the algorithms under consideration, we used Rand Index, Jaccard Coefficient, Folkes - Mallows and Run Time as four metrics.

5.3 CLUSTER VALIDITY

The procedure of evaluating the results of a clustering algorithm is known as cluster validity. The validation step permits to evaluate the goodness of clustering results using different measures. Supervised measures [7] like Rand Index, Adjusted Rand Index, Jaccard Coefficient and Folkes and Mallows Index are used here for evaluating the cluster output. All these measures evaluate the results according to class labels.

Given a set of n elements $S = \{0_1, \dots, 0_n\}$ and two partitions (U and V) of S to compare,

a – The number of pairs of elements in S that are in the ‘same set’ in U and in the ‘same set’ in V.

b – The number of pairs of elements in S that are in the ‘same set’ in U and in the ‘different sets’ in V.

c- The number of pairs of elements in S that are in the ‘different sets’ in U and in the ‘same set’ in V.

d- The number of pairs of elements in S that are in the ‘different sets’ in U and in the ‘different sets’ in V.

where a, b, c and d are computed for all pairs of data points i and j and their respective cluster assignments.

$$a = |\{i, j | c_U(i) = c_U(j) \wedge c_V(i) = c_V(j)\}|$$

$$b = |\{i,\ j | c_U(i) = c_U(j) \wedge c_V(i) \neq c_V(j)\}|$$

$$c = |\{i, j | c_U(i) \neq c_U(j) \wedge c_V(i) = c_V(j)\}|$$

$$d = |\{i, j | c_U(i) \neq c_U(j) \wedge c_V(i) \neq c_V(j)\}|$$

M = a+b+c+d,

Which is the maximum number of all pairs in the dataset (M = N*(N-1)/2, where N is the total number of points in the dataset.

Following four indices are used to measure the degree of similarity between U and V :

1. Rand Index (R)

   $$R = \frac{(a+d)}{(a+b+c+d)}$$

2. Adjusted Rand Index (ARI)

   $$ARI = \frac{2^*(ad-bc)}{(a+b)(b+d)+(a+c)(c+d)}$$

3. Jaccard Coefficient (J)

   $$J = \frac{a}{(a+b+c)}$$

4. Folkes and Mallows Index (FM)

   $$FM = \frac{a}{\sqrt{M_1 M_2}}$$

   Where $m1 = \frac{a}{(a+b)}$ and $m2 = \frac{a}{(a+c)}$

The four indices have proven that high values of indices great similarity between U and V.

Cluster validity using distance-based K-medoids algorithm for primary & secondary datasets is represented in Table 2. Considering Rand Index (R), the obtained accuracy is 89% for primary dataset and 92% for secondary dataset. Performance analysis of various methods is performed by using WEKA tool.

Table 2: Cluster Validity using Improved Algorithm for Primary & Secondary Liver Dataset

S. No.	Datasets	ARI	R	J	FM Index
1.	Indian Liver Patient Dataset (Primary)	0.76	0.89	0.72	0.84
2.	Indian Liver Patient Dataset (Secondary)	0.91	0.92	0.91	0.91

For evaluating the algorithms under consideration, we used Rand Index, Jaccard Coefficient, Folkes - Mallows and Run Time as four measures.

The performance of improved algorithm is simulated over 2000 samples of primary data by using Rand Index, Jaccard Coefficient, Folkes - Mallows and Run Time as four cluster validity measures. Table 3 summarizes the results, where the adjusted Rand indices were reported in a), the Jaccard Coefficient in b),the Folkes & Mallows in c) and computation time in d).

Table 3 (a) Adjusted Rand Index(R)

N	K-Means	PAM	Proposed Method
800	0.657	0.691	0.764
1000	0.689	0.691	0.772
1200	0.591	0.662	0.756
1400	0.615	0.621	0.762
1600	0.666	0.674	0.774
1800	0.561	0.567	0.662
2000	0.598	0.591	0.771

Table 3 (b) Jaccard Coefficient (J)

N	K-Means	PAM	Proposed Method
800	0.524	0.672	0.772
1000	0.612	0.711	0.792
1200	0.667	0.724	0.782
1400	0.692	0.798	0.811
1600	0.598	0.675	0.724
1800	0.612	0.678	0.728
2000	0.698	0.715	0.756

Table 3 (c) Folkes and Mallows Index (FM)

N	K-Means	PAM	Proposed Method
800	0.774	0.792	0.852
1000	0.715	0.778	0.877
1200	0.784	0.811	0.862
1400	0.712	0.794	0.897
1600	0.811	0.842	0.882
1800	0.722	0.794	0.875
2000	0.677	0.716	0.859

Table 3 (d) Computation Time (in seconds)

N	K-Means	PAM	Proposed Method
800	5.124	4.912	4.711
1000	5.675	5.475	5.321
1200	6.167	6.001	5.881
1400	6.723	6.565	6.115
1600	7.356	7.189	6.567
1800	7.992	7.764	7.234
2000	8.452	8.189	7.992

An efficient distance based improved K- Medoids Clustering Algorithm is reported to be better than previous algorithms. The complexity of proposed algorithm is O(nk) which is better than of PAM.

6. Conclusion

In this paper , an efficient distance – based K –medoids algorithm has been used for clustering. This improved algorithm was applied on primary and secondary Liver dataset. The result of proposed algorithm are more accurate and easily found with less computation time. Compared to other algorithms like k-means algorithm and Parition Around Medoids (PAM), it uses less number of iterations to produce more accurate results. In this method , initial medoids are selected from distance matrix , using Manhattan distance. It avoids scanning of large database every time as it updates the Medoids, using the Manhattan distance matrix. Optimal number of clusters are chosen from Silhouette index.This improved algorithm is less effected by outliers. The experimental results exhibits that by using this algorithm, one can obtain compact clusters very quickly and efficiently.

4. ACKNOWLEDGMENTS

Our thanks to Bisaria Patholgy Lab. Etah, and UCI repository for providing primary, secodary Liver dataset respectively.

5. REFERENCES

[1] Bala Suder, V., Devi, T. and Saravanan N. 2012 "Development of a Data Clustering Algorithm for Predicting Heart" International Journal of Computer Applications" Vol 48, Issue 7, pp 0975-888.

[2] Eisten, M., Spellman, P., Brown, P. and Botstein, D. 1998, "Cluster Analysis and Display of Genome-Wide Expression Patterns", in Proc. Natl. Acad. Science USA, Vol. 95, No. 25, pp. 14863 – 14868.

[3] ftp://ftp.ics.uci.edu/pub/machine-learning-databases

[4] Han, J. W. and Kamber, M., 2001 Data Mining Concepts and Techniques, Higher Education Press, Beijing.

[5] Han, J., Kamber, M. and Tung, A. 2001. Spatial clustering methods in data mining: A survey. In Miller, H., and Han, J., eds., Geographic Data Mining and Knowledge Discovery. Taylor & Francis.

[6] Hartigan, J., A. and Wong, M., A. 1979, " A K-Means Clustering Algorithm", Applied Statistics, Vol. 28, No. 1, pp. 100-108.

[7] Jiang, D., Tang, C. and Zhang, A. 2004, "Clustering Analysis for Gene Expression Data: A Survey", IEEE Transactions on Knowledge and Data Engineering, Vol. 16, No. 11, pp. 1370–1386.

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

[9] Li, S, Wu X and Tan M 2008, "Gene Selection Using Hybrid Particle Swarm Optimization and Genetic Algorithm", Soft Computing, Vol. 12, No. 11, pp. 1039–1048.

[10] Lu, Y, Lu, S, Fotoulhi F, Denf Y. and Brown, S. 2004, "Incremental Genetics K-Means Algorithm and Its Applications in Gene Expression Data Analysis", BMC Bioinformatics, Vol. 5, pp. 172– 180.

[11] Park, H-S and Jun, C-H 2009, "A Simple and Fast Algorithm for K-Medoids Clustering", Expert Systems with Applications, Vol. 36, No. 2, pp. 3336 – 3341.

[12] Raghuvira, P. A., Vani, K. S. and Rao, K. N. 2011. An Efficient Density Based Improved k- medoids Clustering Algorithm. International Journal of Advanced Computer Science & Application. Vol 02, No.6,49-54.

[13] Ranga Raj, R. Punithavalli . 2012. " Evaluation of Enhanced K-means Algorithm to Student Dataset.International Journal of Advanced Networking & Application". Vol 04,Issue 02, pp 1578-80.

[14] Raymond, T. Ng and Jiawei Han 2002, "CLARANS: A Method for Clustering Obejects for Spatial Data Mining", IEE Transactions on Knowledge and Data Engineering, Vol. 14, No. 5, pp. 1003–1016.

[15] Selim, S., Z. and Ismail, M., A. 1984, "K-Means Type Algorithms: A Generalized Convergence Theorem and Characterization of Local Optimality", IEEE Trans. Pattern Anal. Mach. Intel., Vol. 6, No. 1, pp. 81–87.

[16] Singh, S., S. and Chauhan,N. C. 2011. " K-means v/s K-medoids: A Comparitive study ". National Conference on recent trends in Engineering & Technology.