# Approaches to Partition Medical Data using Clustering Algorithms

P.Kalyani,
Research Scholar of Mother Teresa Women's University,
Koadikanal.

## ABSTRACT

The successful application of data mining in fields like e-business, marketing and retail have led to the popularity of its use in knowledge discovery in databases (KDD) in other industries and sectors. Data is a great asset to meet long-term goals of any organization and can help to improve customer relationship management. It can also benefit healthcare providers like hospitals, clinics and physicians, and patients, for example, by identifying effective treatments and best practices popularity of its use in knowledge discovery in databases (KDD) in other industries and sectors. Efficient clustering tools reduce demand on costly healthcare resources. It can help physicians cope with the information overload and can assist in future planning for improved services. Clustering results are used to study independence or correlation between diseases and for better insight into medical survey data. To achieve this, create clustering algorithms that enhances the traditional K-Means, DB-Scan and Fuzzy C-Means algorithms.

## Keywords

Knowledge discovery, cluster, K-means, Density based scan.

## 1. INTRODUCTION

Unsupervised clustering property that takes into account the cluster densities, number of data points in each subset. Several factors have motivated the use of data mining applications in medical field. The existence of medical insurance fraud and abuse, for example, has led many healthcare insurers to attempt to reduce their losses by using data mining tools to help them find and track offenders (Christy, 1997). Fraud detection using data mining applications in healthcare fraud and abuse detection is an emerging field (Milley, 2000). Another factor is that the huge amounts of data generated by healthcare transactions are too complex and voluminous to be processed and analyzed by traditional methods. Data mining can improve decision-making by discovering patterns and trends in large amounts of complex data (Biafore, 1999). Such analysis has become increasingly essential as financial pressures have heightened the need for healthcare organizations to make decisions based on the analysis of clinical and financial data. Insights gained from data mining can influence cost, revenue, and operating efficiency while maintaining a high level of care (Silver *et al*., 2001).

Clustering is an unsupervised learning technique that deals with finding a structure in a collection of unlabeled data. It is the process of organizing objects into groups whose members are similar in some characteristics. The result of clustering represents a data concept, where a cluster represents a collection of objects that are similar between them and are dissimilar to the objects belonging to other clusters. In medical domain, cluster analysis provides a systematic, formalized method for data exploration and defining groups with clinical similarities.

Clustering tools can be used as an evidence-based medicine analysis system that can help in the prevention of hospital errors. Further they can be used to discover patterned diseases from stored patient data, which can help during screening, diagnosis, therapy, prognosis, monitoring, epidemiological studies, biomedical/biological analysis, hospital management, medical instruction and training.

Efficient clustering tools reduce demand on costly healthcare resources. They can help physicians cope with the information overload and can assist in future planning for improved services. Clustering results are used to study independence or correlation between diseases and for better insight into medical survey data. All these benefits motivated the researcher to develop clustering models to group medical data.

Clustering medical data faces a number of new challenges.

- Information overload – Advances in medical equipments combined with high computing ability is increasing the amount of data collected and stored in health care industry. Knowledge discovery and retrieval of information from such huge databases is challenging and are prohibitively expensive.
- Too many disease markers (attributes or dimensions) available for decision making and are heterogeneous in nature.
- The high awareness for quality care among public and increased life expectancy is increasing the demand for quality health services. But with overworked and tired physicians, stressful work conditions, etc., misdiagnosis and imprecise treatment solutions occur.

In order to meet the above challenges, the problem statement of the present research is formulated as :

*Given D = $\{d_1, d_2, \ldots, d_n\}$ set of edical records with C = $\{c_1, c_2, \ldots, c_m\}$ set of disease categories and T = $\{t_1, t_2, \ldots, t_n\}$ features, the research problem is to use a similarity or distance metric along with a partitioning criteria, to group records with similar features into various categories.*

To develop such clustering models, the primary objective was to propose algorithms for clustering that can efficiently partition data set into an optimal number of clusters. Efficiency is defined as an

## 2. ENHANCED CLUSTERING ALGORITHMS

Three clustering algorithms, K-Means, DB-Scan and Fuzzy C Means algorithms are selected and enhanced. The methods used are explained in this section.

### 2.1. Enhanced K-Means Algorithm

K Means clustering generates a specific number of disjoint, flat (non-hierarchical) clusters. It is well suited to generating globular clusters. The K Means method is an unsupervised, non-deterministic and iterative method with the following properties.

- There are always K clusters.
- There is always at least one item in each cluster.

- The clusters are non-hierarchical and they do not overlap.

## 2.2. Enhanced DBSCAN

DBSCAN (Density Based Spatial Clustering of Applications with Noise) finds number of clusters starting from the estimated density distribution of corresponding nodes. DBSCAN's definition of a cluster is based on the notion of density reachability. Basically, a point q is directly density-reachable from a point p if it is not farther away than a given distance ε (i.e., is part of its ε-neighborhood) and if p is surrounded by sufficiently many points such that one may consider p and q be part of a cluster.

## 2.3. Enhanced Fuzzy C Means Clustering Algorithm (EFCM)

The traditional Fuzzy C Means (FCM) clustering algorithm works by assigning membership to each data point corresponding to each cluster center on the basis of distance between the cluster center and the data point. When the data is near to a cluster center, it's membership towards the cluster centre is considered to be correlated. Here, the summation of membership of each data point should be equal to one. After each iteration, membership and cluster centers are updated according to the formula:

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c} (d_{ij} / d_{ik})^{(2/m-1)}}$$

$$(1)\ v_j = \frac{\left(\sum_{i=1}^{n} (\mu_{ij})^m x_i\right)}{\left(\sum_{i=1}^{n} (\mu_{ij})^m\right)}$$

$\forall_j = 1, 2, \ldots c$ (2)

where n is the number of data points, $v_j$ is the $j^{th}$ cluster center, m is the fuzziness index $m \in [1, \infty]$, c is the number of cluster center, $\mu_{ij}$ is the membership of $i^{th}$ data to $j^{th}$ cluster center and $d_{ij}$ is the Euclidean distance between $i^{th}$ data and $j^{th}$ cluster center. The main objective of fuzzy c-means algorithm is to minimize:

$$J(U,V) = \sum_{i=1}^{n} \sum_{j=1}^{c} (\mu_{ij})^m \| x_i - v_i \|^2 \quad (3)$$

where$\|x_i - v_j\|$ is the Euclidean distance between $i^{th}$ data and $j^{th}$ cluster center.

## 3. RESULTS AND DISCUSSION

Four medical datasets, namely, Bupa, Heart, Pima and Thyroid, were used during experimentation. A brief description of the datasets is given below.

- Bupa data : The BUPA dataset contains 345 single male patients with 6 numeric attributes. Five of these attributes are blood tests which are thought to be relevant to liver disorders. The sixth attribute corresponds to the number of alcoholic beverages drunk per day. The dataset has two classes.
- Heart Disease : This 13 attribute data set has 2 classes having a total of 270 records.
- Pima Indians Diabetes Data Set : This 8-dimensional data set has a separate training set, a separate testing set and 2 classes having a total of 768 records.
- Thyroid Disease Data Set : This 21-dimensional data set has a separate training set, a separate testing set and 3 classes. The training set consists of 3772 samples and the testing set consists of 3428 samples.

The performance evaluation was based on three performance metrics, silhouette measure, entropy measure and speed of clustering. Figure 1 shows the results of clustering algorithm with respect to silhouette measure.

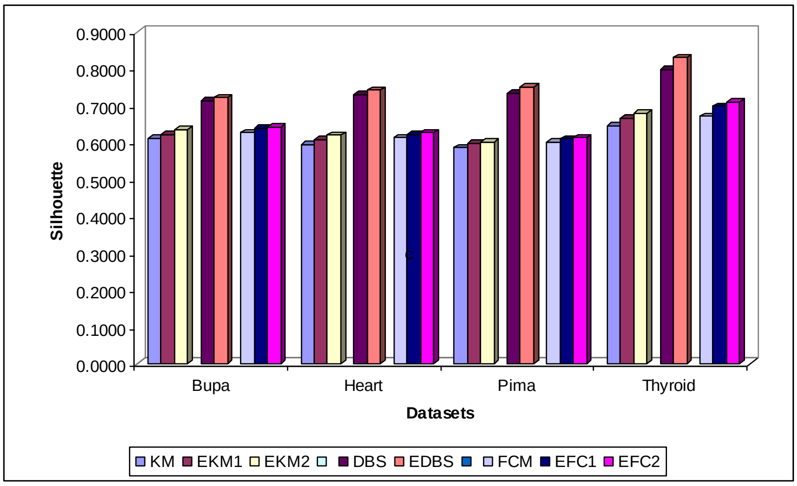

**Figure 1. Silhouette measure**

From the results, it is clear that all the proposed enhanced clustering models have produced quality clusters when compared with the traditional algorithms. While comparing between the EKM1 and EKM2 models, the EKM2 model showed more efficiency (average 1.15% efficiency gain). The EDBS algorithm showed an efficiency gain of more than 2.26% on average. Comparison of EFC1 and EFC2 showed that EFC2 is an improved version over EFC1 and gained 0.81% efficiency on silhouette measure. While comparing all the five proposed models, the EDBS algorithm was the clear winner with an average silhouette measure of 0.7611. Similar trend was observed with all datasets.

Figure 2 shows the performance of the proposed five models with respect to Entropy measure. The analyze the efficiency gained by the inclusion of enhancements, the results are compared with its traditional counterparts.

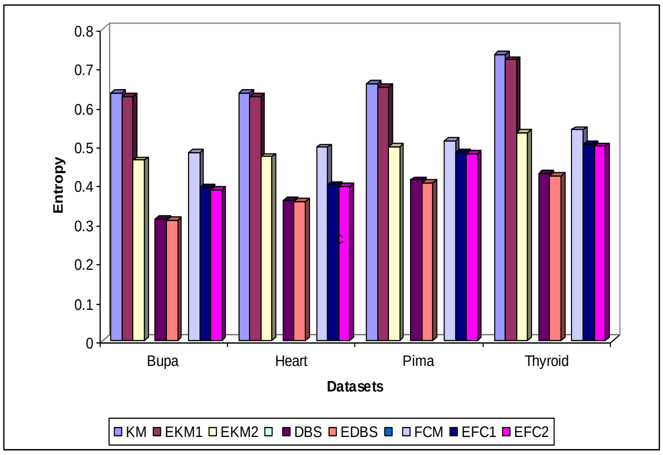

**Figure 2. Entropy measure**

From the results, it could be seen that the again all the five proposed models have outperformed the traditional algorithms with respect to entropy measure. This is evident from the low entropy values obtained by the proposed models. While comparing K-means based method, the EKM2 algorithm performed better than EKM1 algorithm achieving more than 25.05% efficiency gain while entropy measure is considered. The EDBS algorithm produced 1.21% efficiency gain when compared with the traditional DBS algorithm showing that the enhancement has a positive impact over clustering. Comparison of FCM-based algorithms showed that the performance of EFC2 model (average 0.4398) is higher to that of EFC1 (average 0.4447)and traditional FCM algorithms (average 0.5076). Comparison of all the five proposed clustering algorithms again showed that the EDBS with average entropy measure 0.373 produced quality clusters. The same results were envisaged for all the models.Table 1 shows the execution speed of the proposed and traditional clustering models.

**Table 1.EXECUTION SPEED (SECONDS)**

| **Dataset** | **KM** | **EKM1** | **EKM2** | **DBS** | **EDBS** | **FCM** | **EFC1** | **EFC2** |
|---|---|---|---|---|---|---|---|---|
| Bupa | 5.21 | 6.04 | 6.26 | 5.96 | 6.42 | 5.88 | 6.47 | 7.14 |
| Heart | 6.58 | 7.36 | 7.95 | 7.19 | 8.24 | 7.02 | 8.55 | 8.56 |
| Pima | 10.17 | 12.13 | 12.99 | 11.57 | 13.87 | 11.29 | 14.15 | 15.27 |
| Thyroid | 499 | 531 | 538 | 527 | 561 | 519 | 568 | 571 |

From the results, it could be seen that while considering small dataset, the traditional algorithms were faster than the enhanced counterparts. The possible reason for slowness of the proposed algorithms is that the number of algorithms used to enhance the traditional algorithms in the process of automatic parameter estimation and improving the quality of clustering, requires more calculations. However, the time difference is very minimum and since medical industry is more concerned with accuracy than time, the EDBS algorithm is still considered the best among the proposed algorithms. While comparing between dataset, the algorithm was faster with small datasets (Bupa, Heart) and slowed with large dataset (Thyroid).

## 4. CONCLUSION

The study and analyzed the problem of partitioning medical data. The enhanced the existing traditional algorithms (K Means, DBScan and Fuzzy C Means) and the proposed K-Means, DBScan and Fuzzy C-Means Clustering Algorithm. Performance evaluation was based on silhouette measure, entropy and speed of clustering. The experiments were conducted at each stage and all the experimental results proved that while all proposed models performed better than traditional algorithms.

## 5.REFERENCES

1. Christy, T. (1997) Analytical tools help health firms fight fraud. Insurance & Technology, 22(3), 22-26.

2. Ester, M., Kriegel, H., Sander, J. and Xu, X. (1996) A density-based algorithm for discovering clusters in large spatial databases with noise, Proc. KDD 96, Pp. 226–231.

3. Gillespie, G. (2000) There’s gold in them thar’ databases. Health Data Management, 8(11), 40-52.

4. Halkidi, M., Batistakis, Y. and Vazirgiannis, M. (2001) Clustering algorithms and validity measures, Proceedings of SSDBM Conference, Virginia, USA.

5. Hamerly, G. and Elkan, C. (2003) Learning the k in k-means, Proceedings of the 17th Annual Conference on Neural Information Processing Systems, Pp. 281-288.

6. Han, J. and Kamber, M. (2000) Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers.

http://archive.ics.uci.edu/ml/datasets.html