A Guided clustering Technique for Knowledge Discovery – A Case Study of Liver Disorder Dataset

Vikram Singh
Professor and Chairperson,
Department of Computer Science and Applications,
Chaudhary Devi Lal University, Sirsa, Haryana, INDIA.
Ph- 91-946609055,
E-mail- vikramsinghkuk@yahoo.com

Sapna Nagpal
Assistant Professor
Department of Computer Science
Pt. J.L.N. Govt. College, Faridabad, Haryana, INDIA.
Ph- 91-9212730974
E-mail- sapu_mehta@yahoo.com

ABSTRACT

The problem of extracting hidden patterns in medial domain is becoming increasingly relevant today, as the records of patient's clinical trials and other attributes are available electronically. The purpose of finding patterns in medical databases is to identify those patients which share common attributes and hence constitute same risk group. This paper presents an experiment based on clustering data mining technique to discover hidden patterns in the dataset of liver disorder patients. The whole effort was directed to find crisp clusters of same risk group of patients; for which the intermediate results during the experiment have been discussed with a domain expert for feedback.

Keywords: Data Mining, Clustering, Guided Clustering, Liver Disorder Dataset.

1. INTRODUCTION

Development of effective and efficient data mining tools has profound influence on interpretation of the final results for optimal experimental design in many fields. Availability of several high-quality health care databases offers high potential for knowledge management in healthcare using data mining tools. Traditionally, decision making in health care was based on ground information, past experience and fund constraints. But with the use of various data mining and knowledge discovery techniques, a knowledge rich health environment can be created [9]. Knowledge discovery can be effective in working with large volume of data to find meaningful relationship and to develop strategic solutions. Data mining tools and techniques assume importance for health care professionals in medical diagnosis as well as other fields of health care management.

Clustering, a data mining technique uses unsupervised learning method for data analysis [11]. Clustering adds to the value of existing databases by revealing hidden relationships in the data, which are useful for understanding trends, making predictions of future events from historical data, or synthesising data records into meaningful clusters [8]. Objectives of clustering in the field of health databases include 1) *data segmentation* - partition of large sets into groups according to similarity and 2) *outlier detection* - values that are “far away” from any cluster. Sometimes, clinicians want to know which patients share common attributes so that they can be treated as a group, or in what attributes a particular patient differs from others.

A clinical syndrome is a set or a cluster of concurrent symptoms which indicate the presence and nature of a disease. Therefore, looking for concurrent symptoms is therefore one of the main tasks in medical diagnosis. This paper targets to answer such questions in medical domain.

The study in this chapter uses the database records pertaining to liver disorder. The liver is a vital organ present in vertebrates; there is currently no way to compensate for the absence of liver function. Many disorders of the liver may occur such as alcohol-induced liver disease, fatty liver, alcoholic hepatitis, alcoholic cirrhosis and many more. In addition to complete medical history and physical examination, diagnostic procedures for liver disease may include: 1) laboratory tests, 2) liver function tests - a series of special blood tests that can determine if the liver is functioning properly, and 3) liver biopsy. The experiment conducted herein the present work has used the database records of male individual having the problem of liver disorder, donated by Richard S. Forsyth BUPA, Medical Research Ltd., and has been accessed from UCI Machine Learning Repository website http://archive.ics.uci.edu:80/ml/datasets.html. The dataset contains 7 attributes and 345 rows out of which 5 variables are the attributes of blood test report which are thought to be sensitive to liver disorders that might arise from excessive alcohol consumption. The sixth variable is 'Drinks' which provides the measurement of alcohol consumption, and the last variable is the 'Selector' variable (with two values 1 & 2) that determines the class of the patient. Table 1 presents a brief description of the attributes of the dataset.

Variable Name	**Description**
Mcv	mean corpuscular volume
Alkphos	alkaline phosphotase
Sgpt	alamine aminotransferase

Sgot	aspartate aminotransferase
Gammagt	gamma-glutamyl transpeptidase
Drinks	number of half-pint equivalents of alcoholic beverages drunk per day.
Selector	field used to split data into two sets

Table1: Description of the dataset attributes.

Clustering based data mining techniques have been implemented to find clusters in the liver disorder patient's database (Table1). SOM (Self Organising Map) toolbox kit available in Matlab5 has been used at the first step of the experiment to extract initial information about the cluster boundaries in the database [6]. SOM toolbox is a powerful toolbox in Matlab that can visualise high-dimensional data in a low-dimensional view. It is an automated procedure that yields two dimensional output layers of the input data to get a view of some inherent clusters in the database. A domain expert can get useful knowledge about the data after visualising the clusters in U-matrix of SOM, which can provide certain useful parameters in further analysing the nature of clusters. With the completion of the first stage clustering analysis, the *k-means* clustering algorithm has been implemented to divide the dataset in meaningful clusters. The clustering results can assist the medical practitioners to study the characteristics of individual clusters and personalise therapy for the patients.

2. LITERATURE REVIEW

Fast retrieval of relevant information from the databases has always been a significant issue. The application of data clustering technique for similarity searching in medical databases lends itself into many different perspectives.

In [4] Shing-Hong Liu, et.al has built an automatic disease classification system using pulse waveform analysis, based on a Fuzzy *c-means* (FCM) clustering algorithm. A self designed three-axis mechanism was used to detect the optimal site to accurately measure the pressure pulse waveform (PPW). A fuzzy *c-means* algorithm was used to identify myocardial ischemia symptoms in 35 elderly subjects with the PPW of the radial artery. The harmonic feature vector (H_2, H_3, H_4) performed at the level of 69% for sensitivity and 100% for specificity while the form factor feature vector (LFF, RFF) performed at the level of 100% for sensitivity and 53% for specificity. A negative group was used to determine the tolerance of the FCM algorithm. The harmonic feature vector (H_2, H_3, H_4) predicted results at 57% accuracy level.

Hirano Shoji and Tsumoto Shusaku has presented modified multiscale matching method that enables the multiscale structural comparison of irregularly-sampled, different-length time series medical data[7]. The conventional multiscale matching algorithm has been revised to produce sequence dissimilarity which includes 1) introduction of a new segment representation that elude the problem of shrinkage at high scales and 2) introduction of a new dissimilarity measure that directly reflects the dissimilarity of sequence values. The usefulness of the method on cylinder-bell-funnel dataset and chronic hepatitis dataset was examined and the results demonstrated that the dissimilarity matrix produced by the proposed method, combined with conventional clustering techniques, lead to the successful clustering for both synthetic and real-world data.

Georg Berks, et.al. in [12] have used fuzzy *c-means* clustering to assign symptoms to the different types of aphasia categories. The authors state that the ambiguities inherent in the definition of the aphasic syndromes [10], cannot be resolved completely by the Aachen Aphasia Test or any other applied algorithm. Definitions of syndromes are probabilistic rather than crisply defined and a symptom may belong to more than one class. As these ambiguities exist, the application of fuzzy methods seems to be an adequate means for an exploration of the results of the patients' clinical investigations. The results were compared with those in some subtests of the Aachen Aphasia Test (AAT) which shows that the clustering procedure leads to clearly distinguishable classes of symptoms.

3. CLUSTERING TECHNIQUES FOR MEDICAL DATA ANALYSIS

The criterion for similarity searching in medical databases does not differ from similarity searching in any other database. For example, in order to detect many diseases like Tumor, the scanned pictures or the x-rays are compared with the existing ones and the dissimilarities are recognised. Therefore, generalised clustering tools and techniques can be applied to medical databases, with little or no modification. The dataset used in the experiment contains records of blood tests of liver-disorder patients. The target is to cluster the patient's records into different groups with respect to the test report attributes which may help the clinicians to diagnose the patient's disease in efficient and effective manner.

SOM toolbox (Matlab5) and *k-means* clustering algorithm have been used to group the database into different clusters. The Self-Organising-Map (SOM) is a powerful visualisation tool based on vector quantisation method which places the prototype vectors on a regular low-dimensional grid in an ordered fashion [5]. The SOM Toolbox is an implementation of the Self-Organising-Map and its visualisation in the Matlab5 computing environment. The Toolbox is available free of charge under the GNU General Public License from http://www.cis.hut.fi/projects/somtoolbox. Since SOM Toolbox provides the visual view of the dataset and some obvious clusters can be immediately pointed out by a domain user, who can specify the input parameters for another powerful clustering algorithm named *k-means* to group the dataset in the found clusters. Figure 1 shows the scenario for a typical run of the clustering experiment. It represents the major activities, their sequential order and their interdependencies.

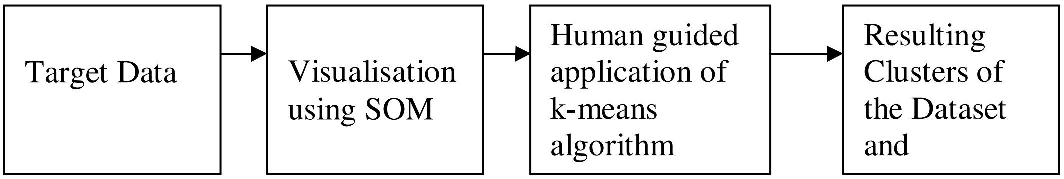

Target Data
Visualisation using SOM
Human guided application of k-means algorithm
Resulting Clusters of the Dataset and

Figure1: Flow of research for the experiment.

4. EXPERIMENTATION AND RESULTS

A case study of liver disorder patient dataset by applying SOM Toolbox and *k-means* clustering algorithm is presented in this section. Firstly the liver disorder dataset was processed using SOM Toolbox to create a learned SOM map. This was done to visualise the structure of the data. The Kohonen self-organising map (SOM) has several important properties that can be used within the data mining/knowledge discovery and exploratory data analysis process. A key characteristic of the SOM is its topology preserving ability to map a multi-dimensional input into a two-dimensional form. This feature is used for classification and clustering of data. However, a great deal of effort is still required to interpret the cluster boundaries.

Figure2 shows the SOM visualisation of the liver disorder data. The U-matrix (unified distance matrix) is an important clue to determine natural clusters in the learned SOM. The U-matrix visualises distances between neighbouring map units, and helps to see the cluster structure of the map. The high values of the U-matrix indicate a cluster border. The elements in the same clusters are indicated by uniform areas of low values. The dark areas are where the distance between the nodes is large - it's a gap. The light areas are where the nodes are close to each other. It is evident in Figure2 U-matrix that there are two clusters (of not properly known inner structure). Top cluster is a big representing a large part of the dataset whereas the bottom cluster contains comparatively few records.

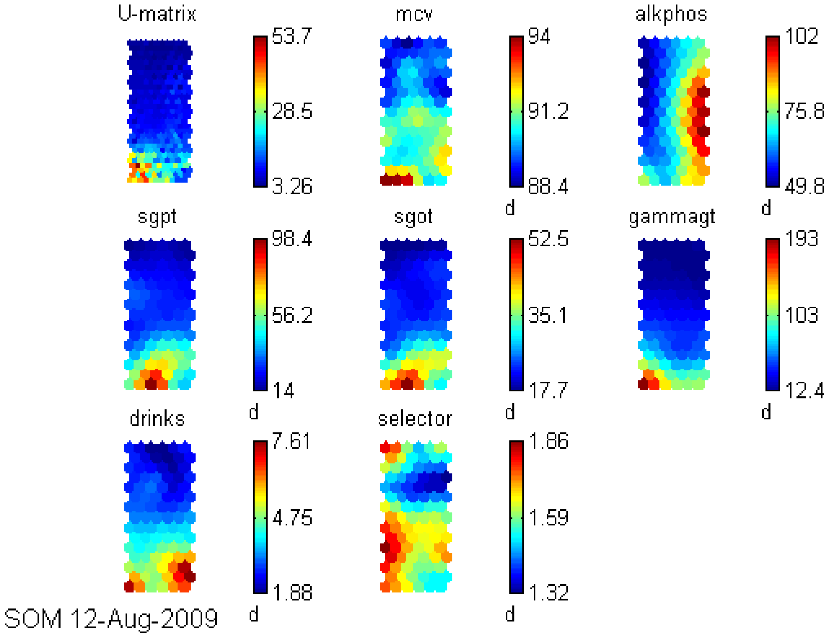


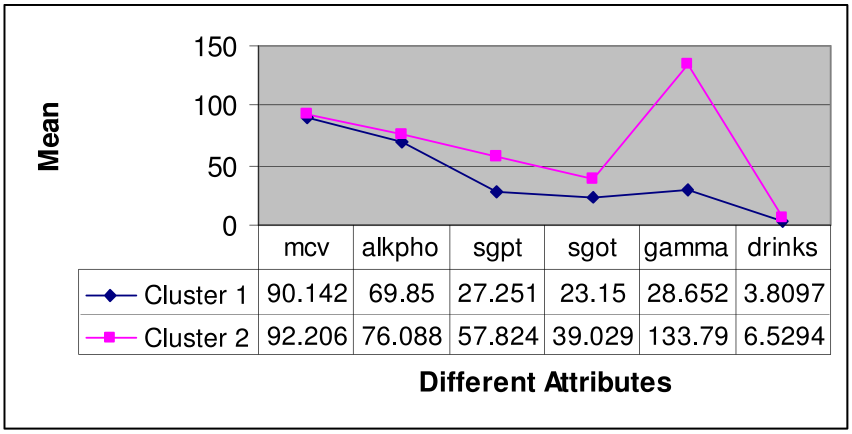
Figure2: Learned SOM using the Liver Disorder Dataset.

Applying SOM toolbox to liver disorder dataset allowed the domain experts to 1) get an idea of rough clusters in the dataset and 2) get the inner structure of the different attributes in the dataset. The endeavour in the experiment is not to bug the problem with over optimistic results and therefore, *k-means* clustering algorithm is applied three time with input criterion two, three and four clusters respectively. *k-means* clustering algorithm in batch mode has been used for data classification. Figure 3, 4, 5 show two, three and four clusters partition of the database respectively. The clusters are represented by the mean or center of the group. By closely examining the three figures, the following observations can be made:

i) The two or three cluster partitions of the dataset has somewhat crisp boundaries.

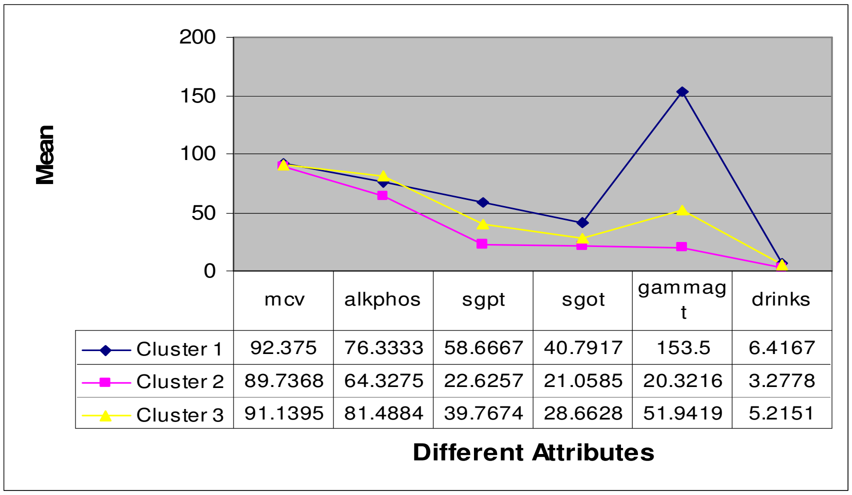
ii) An attempt to divide the dataset into four clusters does not have crisp boundaries and the clusters overlap with each other.

iii) This experiment has best segmented the database into three cohesive sub-groups in which each subgroup has patients with similar behaviour.



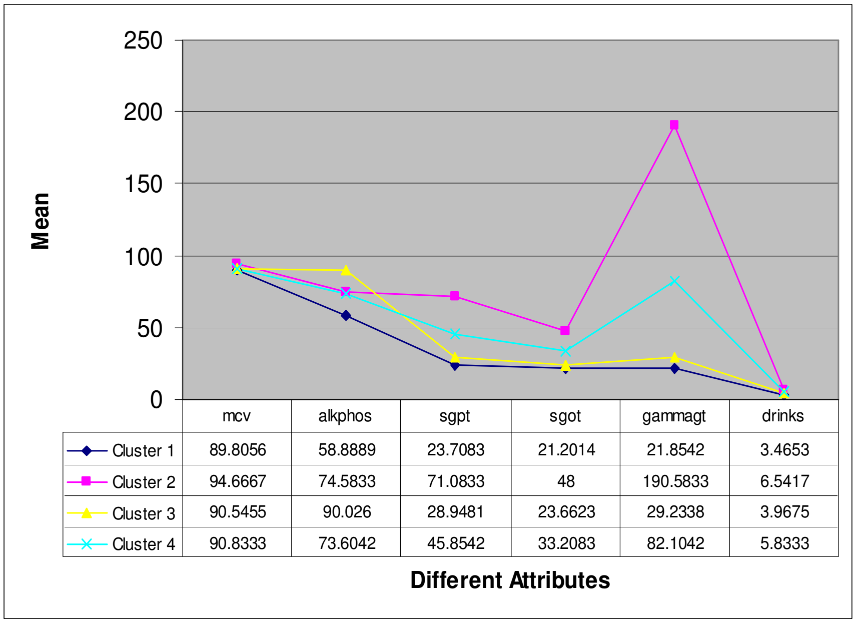
	mcv	alkpho	sgpt	sgot	gamma	drinks
◆ Cluster 1	90.142	69.85	27.251	23.15	28.652	3.8097
■ Cluster 2	92.206	76.088	57.824	39.029	133.79	6.5294

Figure 3: Two cluster view of the database using *k-means* Algorithm.



	mcv	alkphos	sgpt	sgot	gammagt	drinks
◆ Cluster 1	92.375	76.3333	58.6667	40.7917	153.5	6.4167
■ Cluster 2	89.7368	64.3275	22.6257	21.0585	20.3216	3.2778
▲ Cluster 3	91.1395	81.4884	39.7674	28.6628	51.9419	5.2151

Figure 4: Three cluster view of the database using *k-means* Algorithm.



	mcv	alkphos	sgpt	sgot	gammagt	drinks
◆ Cluster 1	89.8056	58.8889	23.7083	21.2014	21.8542	3.4653
■ Cluster 2	94.6667	74.5833	71.0833	48	190.5833	6.5417
▲ Cluster 3	90.5455	90.026	28.9481	23.6623	29.2338	3.9675
× Cluster 4	90.8333	73.6042	45.8542	33.2083	82.1042	5.8333

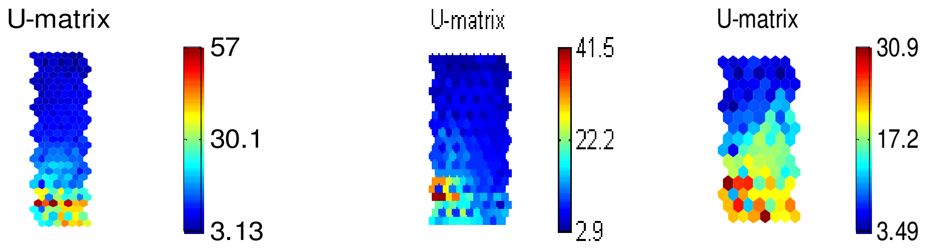
Figure 5: Four cluster view of the database using *k-means* Algorithm.

After dividing the dataset into three clusters, the task is to analyse and explore the inner structure of each of these clusters in terms of simplicity or complexity. The partitioned datasets were again given to the SOM learning to visualise the inner structure of the individual clusters. The results are shown in Figure 6 that yields some even more detailed information about the three clusters.

The partitioned datasets were again given to the SOM learning to visualise the inner structure of the individual clusters. The results shown in Figure 8.6 yield even more detailed information about the three clusters. Visualisation of computed results yielded additional information as described below.

The top-left corner of map visualisation is most important for finding influence of manipulation. It is obvious that influence of manipulation is rather marginal because the values of change in this area vary only around small increase or decrease. The bottom-right corner of map visualisation is most important. It is in direct contrast to the top-left corner of map visualisation and this area can be divided into two parts - darker area representing lower values of the attributes and lighter area in extreme corner representing high levels of the attributes. The first two clusters have a solid area which could be interpreted as similar behaviour (and possibly predictable) of the attributes. On the other hand, the third cluster has more complicated structure and could be divided into several smaller clusters. There are many other indices of cluster validity such as the Bezdek's partition coefficient, the Dunn's separation index, the Xie-Beni's separation index, Davies-Bouldin's index, and the Gath-Geva's index, etc. which can also be used to check variability in the clusters. The clusters can be separated easily, as it is indicated in the small areas of overlap between the respective features. It can be observed that the clustering procedure leads to clearly distinguishable classes of symptoms.

Clustering can prove out to be a pragmatic tool in clinical medicine diagnostics and hence play a vital role in therapeutic decisions. Thus, overall severity of the disease and various factors involved in liver disorder can be distinguished much better. For practical reasons, these points seem to be of greater importance in enhancing the value of existing databases by revealing rules in the data. These rules are useful for understanding trends, making predictions of future events from historical data, and synthesising data records into meaningful clusters.



(a) Cluster 1 (92.375) (b) Cluster2 (89.7368) (c)Cluster3 (91.1395)

Figure 6: Inner Structure of three Clusters discovered.

3. CONCLUSION

The application of clustering technique requires finding a natural association among specific data in the domain. In this work, the clustering technique has been used to discover clusters in the liver disorder dataset using the SOM network's internal parameters and *k-means* algorithm. The research has shown that meaningful results can be discovered from clustering techniques by letting a domain expert specify the input constraints to the algorithm. In addition, the communication between medical scientists and computer engineers may lead to an interdisciplinary advance in the analysis of inconsistencies in medical classifications.

REFERENCES

[1] Kanungo T., Mount D.M., Netanyahu N.S., Piatko C.D., Silverman R. and Wu A.Y. (2002). An Efficient *k-means* Clustering Algorithm: Analysis and Implementation. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, 881-892.

[2] Cheung Y.M. (2003). *k-means*: A New Generalised *k-means* Clustering Algorithm. N-H Elsevier Pattern Recognition Letters 24, Vol 24(15), 2883–2893.

[3] Ding C. and He X. (2004). k-means Clustering via Principal Components Analysis. ACM Proceedings of the 21st International Conference on Machine Learning, Vol. 69, page 29.

[4] Liu S.H., Chang K.M. and Tyan C.C. (2008). Fuzzy *c-means* Clustering for Myocardial Ischemia Identification with Pulse Waveform Analysis. Proceedings of the 13th International Conference on Biomedical Engineering, Singapore, Vol. 23, 485-489.

[5] Vesanto J., Himberg J., Alhoniemi E., Parhankangas J. (1999). Self-organising map in Matlab: the SOM Toolbox. Proceedings of the Matlab DSP Conference, Finland, 35–40.

[6] Malone J., McGarry K., Wermter S. and Bowerman C. (2006). Data mining using rule extraction from Kohonen self-organising maps. Journal of Neural Computing and Applications, Springer London, Vol. 15(1), 9-17.

[7] Hirano S. and Tsumoto S. (2005). Clustering Time-Series Medical Databases Based on the Improved Multiscale Matching. Foundations of Intelligent Systems, Springer Berlin / Heidelberg, Vol. 3488, 612-621.

[8] Suh S.C., Saffer S. and Adla N.K. (2008). Extraction of Meaningful Rules in a Medical Database. Proceedings of the 7th International Conference on Machine Learning and Applications, 450-456.

[9] Krishan W.S. and Kaur H. (2006). Empirical Study on Application of Data Mining Techniques in Healthcare. Journal of Computer Science 2(2), 194-200.

[10] Marshall, J.C., (1986). The description and interpretation of aphasic language disorder. Neuropsychologia 24(1), 5-24.

[11] Wagstaff K., Cardie C., Rogers S. and Schrödl S. (2001). Constrained *k-means* Clustering with Background Knowledge. Proceedings of the 18th International Conference on Machine Learning, 577-584.

[12] Berks G., Keyserlingk D.G.V., Jantzen J., Dotoli M. and Axer H. (2000). Fuzzy Clustering - A Versatile Mean to Explore Medical Databases. ESIT, Aachen, Germany, 453-457.