International Journal of Computer and Communication Technology

Volume 7 | Issue 2 Article 1

April 2016

# ANALYSIS OF VARIOUS CLUSTERING ALGORITHMS OF DATA MINING ON HEALTH INFORMATICS

PANKAJ SAXENA
*Deptt. Of Computer Applications, RBS Management Techanical Campus, Agra*, pankaj_rbs@yahoo.com

SUSHMA LEHRI
*Institute of Engineering & Technology , DR. B. R. Ambedkar University,Agra*, sushmalehri@gmail.com

# ANALYSIS OF VARIOUS CLUSTERING ALGORITHMS OF DATA MINING ON HEALTH INFORMATICS

**[1]PANKAJ SAXENA & [2]SUSHMA LEHRI**

[1] Deptt. Of Computer Applications, RBS Management Techanical Campus, Agra
[2] Institute of Engineering & Technology , DR. B. R. Ambedkar University,Agra
Email: pankaj_rbs@yahoo.com

**Abstract:-** There are large quantities of information about patients and their medical conditions. The discovery of trends and patterns hidden within the data could significantly enhance understanding of disease and medicine progression and management by evaluating stored medical documents. Methods are needed to facilitate discovering the trends and patterns within such large quantities of medical documents. Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Weka is a data mining tools. It contains many machine leaning algorithms. It provides the facility to classify our data through various algorithms. In this paper we are studying the various clustering algorithms. Cluster analysis or clustering is the task of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar (in some sense or another) to each other than to those in other clusters. Our main aim to show the comparison of different clustering algorithms of Data Mining and find out which algorithm will be most suitable for the users working on health data.

***Keywords***— *Data mining algorithms, KDD, Clustering methods, K-means algorithms,Weka tool etc.*

## I. INTRODUCTION

Data mining techniques when compared with the earlier methodologies proves to be the effective and reliable one for retrieving information from the database and providing a provable solution to the respective problem domain by eliminating the redundancy that takes place with other techniques. The main concept that contributes to Data mining[5,6] is the Knowledge discovery process, a sequential process which prone to provide a reliable prediction model for the defined problem domain. In simple words it can be defined as the process of extracting the potentially useful, valid and novel patterns from the data warehouse for resolving a problem domain. Data mining is being used in several applications like banking, insurance, hospital and Health Informatics. In case of Health Informatics, Data mining plays a vital role in helping Physicians to identify effective treatments , and Patients to receive better and more afforadable health sevices.

## II. DATA MINING AND HEALTH INFORMATICS

Data Mining in Health Informatics [4] is an emerging discipline, concerned with developing methods for exploring the unique type of data that come from Health Care database management system , and using those methods to better understand human Health [1], and the settings which they learn in . Data mining is extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data. As we know large amount of data is stored in Health database, so in order to get required data & to find the hidden relationship, different data mining techniques are developed & used. There are varieties of popular data mining task[2,3] within the Health data mining e.g. classification, clustering, outlier detection, association rule, prediction etc. We can use the data mining in Health system as: predicting human characteristics whether they are prone to desease or not, relationship between the human disease test reports & their impact on human health, predicting patient's risks to be infected with disease, discovery of strongly related factors causing disease, knowledge discovery on health data, classification of patient's different factors causing liver disease.

## III.CLUSTERIZATION

Partitioning a set of objects into homogeneous clusters[13,17] is a fundamental operation in data mining. The operation is needed in a number of data mining tasks, such as unsupervised classification and data summation, as well as segmentation of large heterogeneous data sets into smaller homogeneous subsets that can be easily managed, separately modeled and analyzed. Clustering is a popular approach used to implement this operation. Clustering methods[11] partition a set of objects into clusters such that objects in the same cluster are more similar to each other than objects in different clusters according to some defined criteria.

Statistical clustering methods use similarity measures to partition objects whereas conceptual clustering methods cluster objects according to the concepts objects carry. Data mining applications frequently involve categorical data. The biggest advantage of

these clustering algorithms is that it is scalable to very large data sets.

In this paper we present a different clustering algorithm used to cluster categorical data[9]. The algorithm, called k- means, is also well known k-means algorithm. Compared to other clustering methods, the k-means algorithm and its variants are efficient in clustering large data sets, thus very suitable for data mining.

## IV METHODOLOGY

This paper compares various clustering algorithms on LIPD (Indian Liver Patient Dataset) data set using weka tool and predicts the result that will be useful for researchers working on health informatics. The data used for this research 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.(available on UCI repository) . Attribute information is as follows

| S.No | Attribute | Description of Attribute |
|---|---|---|
| 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/G Ratio | Albumin and Globulin Ratio |
| 11. | Selector field | used to split the data into two sets (labeled by the experts) |

## V COBWEB CLUSTERING ALGORITHM

The COBWEB algorithm was developed by machine learning researchers in the 1980s for clustering objects in a
object-attribute data set. The COBWEB algorithm yields a clustering dendrogram called classification tree that characterizes each cluster with a probabilistic description. Cobweb generates hierarchical clustering, where clusters are described probabilistically. COBWEB uses a heuristic evaluation measure called category utility to guide construction of the tree. It incrementally incorporates objects into a classification tree in order to get the highest category utility.

## EXPERIMENTAL RESULTS

The LIPD data set is with COBWEB algorithm with evaluation on training set of WEKA toos[8]. The number of merges found were 216, and number of splits were 183. Total time taken to build model on full trainng data was 29.88 seconds. Number of clusters were 671. See fig 1.

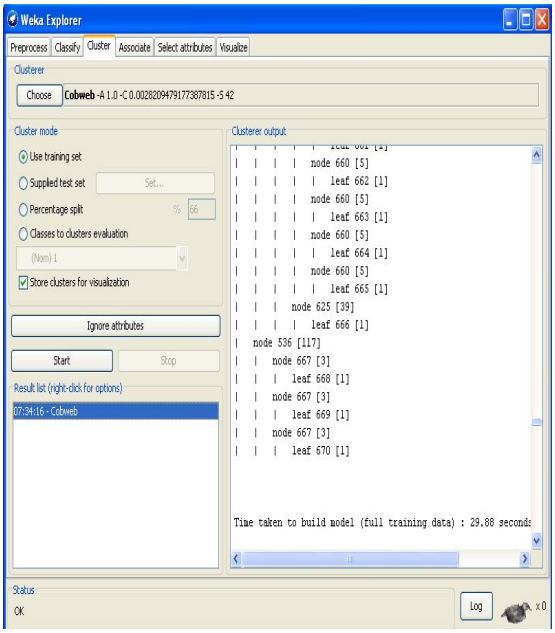


**Figure 1 : Cobweb clustering algorithm**

Visualization of various clusters in different colours can be seen in fig 2.

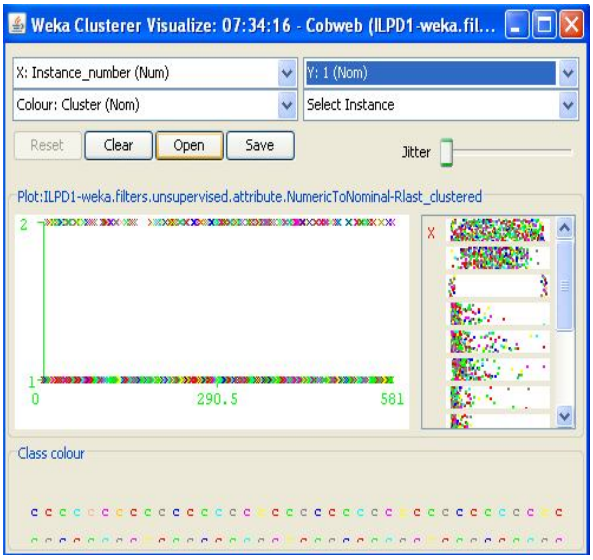


**Figure 2: Result of cobweb cluster in form of graph**

## VI. DBSCAN CLUSTERING ALGORITHM

DBSCAN (for density-based spatial clustering of applications with noise) is a data clustering algorithm proposed by Martin Ester, Hans-Peter Kriegel, Jorge Sander and Xiaowei Xu in 1996. It is a density-based clustering algorithm because it finds a number of clusters starting from the estimated density distribution of corresponding nodes. DBSCAN is one of the most common clustering algorithms and also most cited in scientific literature.OPTICS can be seen as a generalization of DBSCAN to multiple ranges, effectively replacing the parameter with a maximum search radius..

## EXPERIMENTAL RESULTS

The LIPD data set is applied with COBWEB algorithm for evaluation on training set with Epsilon: 0.9, minPoints: 6. Distance-type of clusterers for dDBScan Data Objects is Euclidian DataObject . TheNumber of generated clusters are 4 and elapsed time is 1.14 seconds. see fig 3.

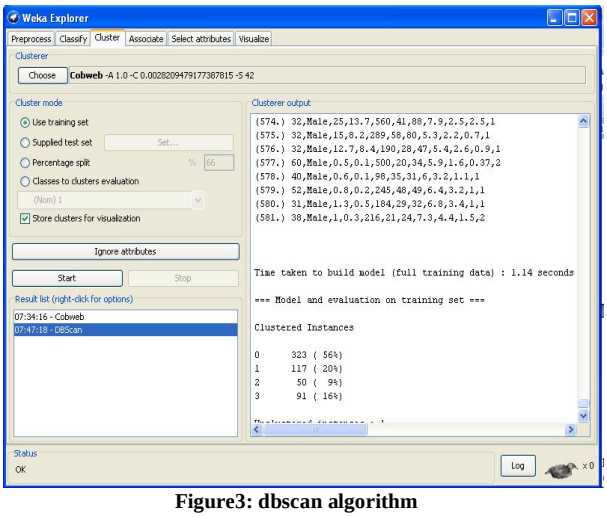


**Figure3: dbscan algorithm**

isualization of various clusters in different colours can be seen in fig 4.

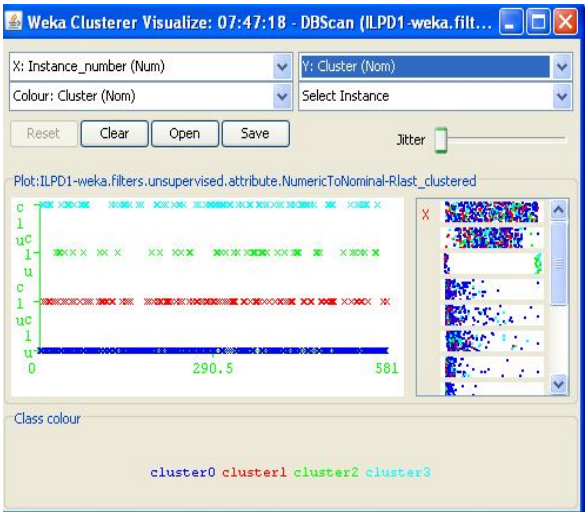


**Figure 4: Result of dbScan cluster in form of graph**

## HIERARCHICAL CLUSTERING

In hierarchical clustering the data are not partitioned into a particular cluster in a single step. Instead, a series of partitions takes place, which may run from a single cluster containing all objects to n clusters each containing a single object. Hierarchical Clustering[12,16] is subdivided into agglomerative methods, which proceed by series of fusions of the n objects into groups, and divisive methods, which separate n objects successively into finer groupings. Hierarchical clustering may be represented by a two dimensional diagram known as dendrogram which illustrates the fusions or divisions made at each successive stage of analysis.

## EXPERIMENTAL RESULTS

The LIPD data set is applied with hierarchical clustering algorithm for evaluation on training set with four clusters.The elapsed time is 1.88 seconds.see fig 5. The tree visualizer can also be analysed (see fig 6).

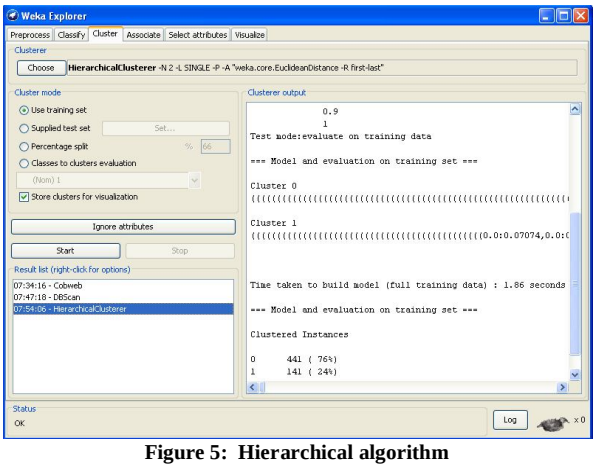


**Figure 5: Hierarchical algorithm**

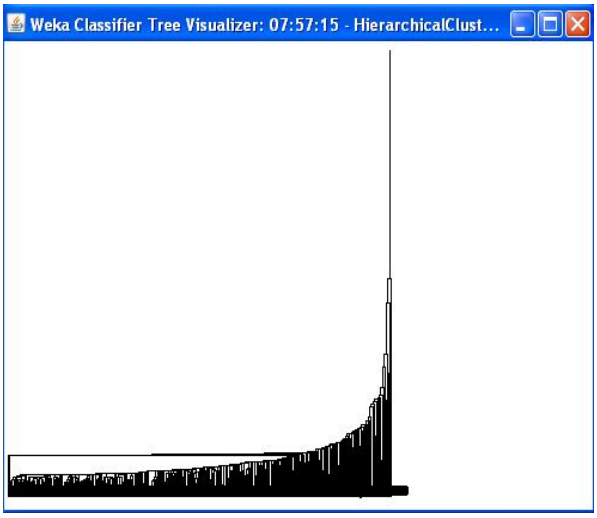


**Figure 6. Tree Visualizer of Hierarchical algorithm**

## THE K MEANS CLUSTERING

The k-means algorithm[7,10] is one of a group of algorithms called partitioning methods. The k -means algorithm is very simple and can be easily implemented in solving many practical problems. The k-means algorithm [15,18]is the best-known squared error-based clustering algorithm.

1. Selection of the initial k means for k clusters
2. Calculation of the dissimilarity between an object and the mean of a cluster
3. Allocation of an object to the cluster whose mean is nearest to the object
4. Re-calculation of the mean of a cluster from the objects allocated to it so that the intra cluster dissimilarity is minimized.

Except for the first operation, the other three operations are repeatedly performed in the algorithm until the algorithm converges. The essence of the algorithm is to minimize the cost function

Where n is the number of objects in a data set X, $X_i$ € X, $Q_l$ is the mean of cluster l, and $Y_{i,l}$ is an element of a partition matrix $Y_{n \times l}$ as in (Hand 1981). d is a dissimilarity measure usually defined by the squared Euclidean distance. There exist a few variants of the k-means algorithm which differ in selection of the initial k means, dissimilarity calculations and strategies to calculate cluster means.The sophisticated variants of the k-means algorithm include the well-known ISODATA algorithm and the fuzzy k-means algorithms .

## EXPERIMENTAL RESULTS

The LIPD data set is applied with k-means algorithm for evaluation on training set. Distance-type of clusterers for k-means Data Objects is Euclidian Distance. The number of iterations is 17,within cluster sum of squared errors:127.02034887051619. Missing values globally replaced with mean/mode The Number of generated clusters are 4 and elapsed time is 0.16 seconds. see fig 7.

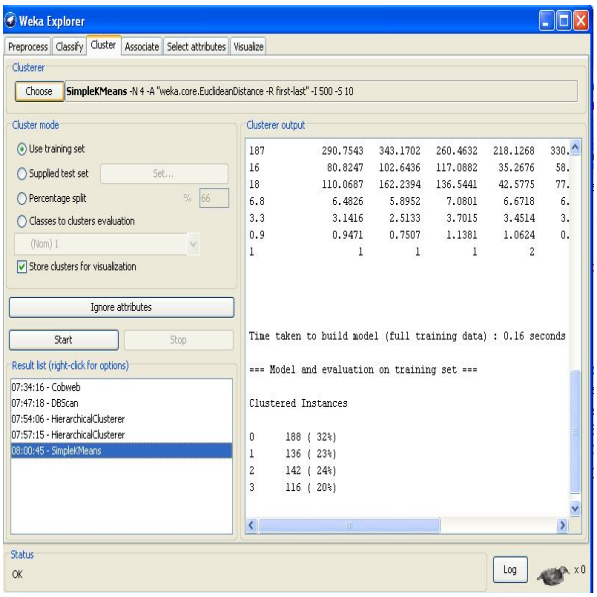


**Figure 7: k-means algorithm**

Visualization of various clusters in different colours can be seen in fig 8.

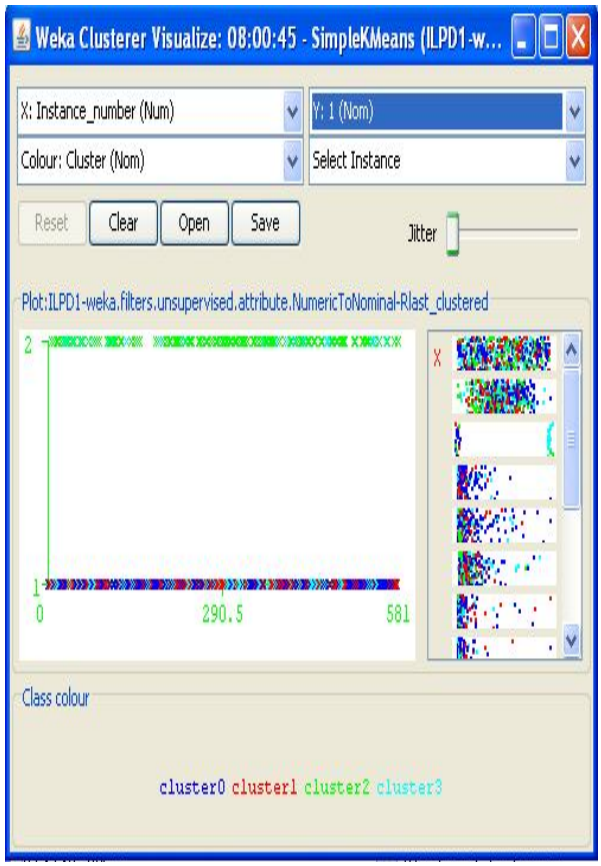


**Figure 8: Result of Simple K-means cluster in form of graph**

## RESULT & CONCLUSION

In the recent few years data mining techniques covers every area in our life. We are using data mining techniques in mainly in the medical, banking, insurances, education etc. Before start working with the data mining models, it is very necessary to have knowledge of available algorithms.The main aim of this paper to provide a detailed introduction of weka clustering algorithms. Weka is the data mining tools. It is the simplest tool for classify the data various types. It is the first model for provide the graphical user interface of the user. For performing the clusterization, we used the UCI data repository. It provide the past medical data for analysis. LIPD dataset was analyesd with different clustering algorithms.With the help of figures we are showing the working of various algorithms used in weka.. Every algorithm has their own importance and we use them on the behavior of the data, but on the basis of this research we found that k-means clustering algorithm is simplest and fastest algorithm as compared to other algorithms. That's why k-means clustering algorithm is more suitable for health datamining applications. This paper shows only the clustering operations in the weka, further research can be performed using other data mining algorithms on health Infomatics.

## REFRENCES:


[1] Hemalatha, M. and Megala, S., 2011 Mining Techaniques in Health Care: A Survey of immunization. *Journal of Theoretical and Applied Information Technology,* Vol. 25, Issue 2, pp. 63–70.

[2] Soni, S. and Vyas,O.P., 2010. Using Associative Classifiers for Predictive Analysis in Health Care Data Mining. *International Journal of Computer Applications,* Vol. 4, Issue 5, pp. 33–37.

[3] Soni,J.,Ansari,U. et al, 2011. Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction, *International Journal of Computer Applications (0975 – 8887), Volume 17– No.8, March 2011.*

[4] Abdul Nazeer, K. A., Sebastian M. P., and. Madhu Kumar S. D., 2011 A Heuristic k-Means Algorithm with Better Accuracy and Efficiency for Clustering Health Informatics Data, *Journal of Medical Imaging and Health Informatics(*American Scientific Publisher) Vol. 1, 66–71.

[5] Han,J. and Kamber, M., "*Data Mining: Concepts andTechniques",* 2nd edition.

[6] Han, J., How can data mining help bio-data analysis? 2002 *Proceedings of the Workshop on Data Mining in Bioinformatics,* BIOKDD .

[7] Z. Huang. 1998 "Extensions to the *k*-means algorithm for clustering large data sets with categorical values". *Data Mining and Knowledge Discovery*, **2**:283–304.

[8] http://www.cs.waikato.ac.nz/ml/weka/

[9] Dempster. A. P., Laird., N. M. and Rubin, D. B. (1977)—"Maximum Likelihood from Incomplete Data via the EM Algorithm "*Journal of the Royal Statistical Society.* Series B (Methodological), Vol. 39, No. 1.(1977), pp.1-38.

[10] Jinxin Gao, David B. Hitchcock .(2009)" James-Stein Shrinkage to Improve K-meansCluster Analysis",University of South Carolina, Department of Statistics November 30, 2009.

[11] N. Ailon, M. Charikar, and A. Newman. (2005)"Aggregating inconsistent information: ranking and clustering". *In Proceedings of the thirtyseventh annual ACM Symposium on Theory of Computing*, pages 684–693, 2005.

[12] E.B Fawlkes and C.L. Mallows.(1983)." A method for comparing two hierarchical clusterings". *Journal of the American Statistical Association*, 78:553–584, 1983.

[13] M. and Heckerman, D. (February, 1998). An experimental comparison of several clustering and intialization methods.*Technical Report MSRTR-98-06*, Microsoft Research, Redmond,WA.

[14] Microsoft academic search: most cited data mining articles: DBSCAN is on rank 24, when accessed on: 4/18/2010

[15] Z. Huang.(1998) "Extensions to the *k*-means algorithm for clustering large data sets with categorical values". *Data Mining and Knowledge Discovery*, **2**:283–304, 1998.

[16] E. B. Fowlkes & C. L. Mallows (1983), "A Method for Comparing Two Hierarchical Clusterings", *Journal of the American Statistical Association* 78, 553–569.

[17] Sharma,N. , Bajpai,A., and Litoriya, R. 2012.Comparison the various clustering algorithms of weka Tools, *International Journal of Emerging Technology and Advanced Engineering*, Volume 2, Issue 5, May 2012.

[18] Abdul Nazeer, K. A., Sebastian M. P., and. Madhu Kumar S. D., 2011 A Heuristic k-Means Algorithm with Better Accuracy and Efficiency for Clustering Health Informatics Data, *Journal of Medical Imaging and Health Informatics(*American Scientific Publisher) Vol. 1, 66–71.