Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 6, June 2015, pg.392 – 397

RESEARCH ARTICLE

Analysis of Breast Cancer Recurrence using Combination of Data Mining Techniques

Ashish Jethi[1], Manishi Kalra[2], Niranjan Bhattacharyya[3]

[1]Student, Guru Tegh Bahadur Institute of Technology, India

[2]Student, Guru Tegh Bahadur Institute of Technology, India

[3]Professor, Bhagwan Purshuram Institute of Technology, India

[1] ashishj027@gmail.com; [2] manishikalra@gmail.com; [3] niranjan.bhattacharyya@gmail.com

***Abstract*— Data mining is considered to deal with huge amounts of data which are kept in the database, to locate required information and facts. The goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. In this paper we present a two-step analysis of Breast Cancer database using Weka 3.6 tool. In this paper we make use of a large database 'Breast Cancer Dataset' containing 10 attributes and 286 instances to perform combination of clustering and association in order to analyse the factors responsible for recurrence of Breast Cancer in women.**

***Keywords*— WEKA, Clustering, Association Rule Mining, Breast Cancer Dataset**

I. INTRODUCTION

Data mining is the analysis step of the KDD (Knowledge Discovery and Data Mining) process. It is defined as the process of extracting interesting (non-trivial, implicit, previously unknown and useful) information or patterns from large information repositories such as: relational database, data warehouses etc.[1] The goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Association rules were first introduced by Agarwal [5]. Association Analysis is the detection of hidden pattern or condition that occurs frequently together in a given data. Association Rule mining techniques finds interesting associations and correlations among data set. Association rule are the statements that find the relationship between data in any database. Association rule has two parts "Antecedent" and "Consequent". For example {bread} => {eggs}. Here bread is the antecedent and egg is the consequent. Antecedent is the item that is found in the database, and consequent is the item that is found in combination with the first. Association rules are being used widely in various areas such as telecommunication networks, risk and market management, inventory control, medical diagnosis/drug testing etc. [1]. Clustering is the unsupervised classification of patterns into clusters [3]. It groups similar objects together in a cluster (or clusters) and dissimilar objects in other cluster (or clusters) [6]. In this paper WEKA (Waikato Environment for knowledge analysis) machine learning tool [2] [4] is used for performing clustering and association algorithms. In this paper we use Breast Cancer dataset taken from UCI Cleveland repository, having 10 attributes and 286 instances.

II. PROPOSED METHOD

For more accurate results we are using combination of two data mining techniques i.e. clustering and association. Clustering is the task of segmenting a diverse group into a number of similar subgroups or clusters. While association used for determining interesting patterns from a large database. In our approach we are using Kmeans clustering method and apriori algorithm for finding frequent patterns. Apply Clustering technique on the database using Weka tool [4]. This step will provide us with a number of clusters according to the class attribute in our dataset. The clusters are then partitioned and summarized manually and then association rule mining is applied to obtain important and well defined rules for each cluster of Breast Cancer. This process consist of various consecutive stages communicating with each other as data preprocessing, data partitioning and association rule mining.

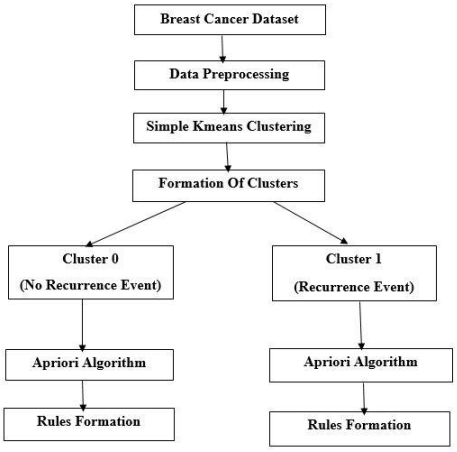

Fig 1. Flowchart of proposed method

A. KMEANS CLUSTERING TECHNIQUE

K-Means is a simple learning algorithm for clustering analysis [6]. The goal of K-Means algorithm is to find the best division of *n* entities in *k* groups, so that the total distance between the group's members and its corresponding centroid, representative of the group, is minimized. It select k points as initial centroids and find K clusters by assigning data instances to nearest centroids. It is the unsupervised classification to find optimal clusters. Distance measure used to find centroids is Euclidean distance.

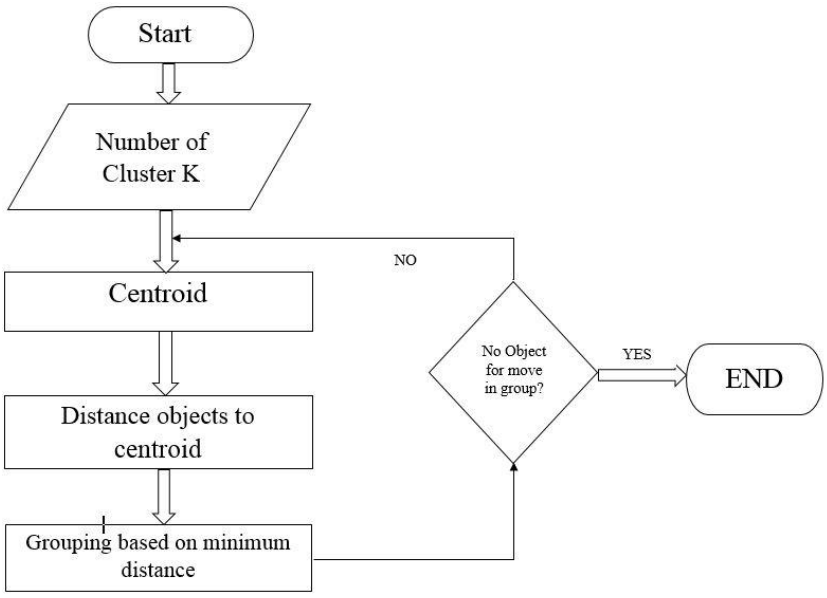

Fig 2. Flowchart of Kmeans Clustering

B. APRIORI ALGORITHM

Association Rule mining techniques finds interesting associations and correlations among data set [9]. An association rule is a rule, which entails certain association relationships with objects or items, for example the interrelationship of the data item as whether they occur simultaneously with other data item and how often. The search for association rules is guided by two parameters: support and confidence. Apriori returns an association rule if its support and confidence values are above user defined threshold values. The output is ordered by confidence. If several rules have the same confidence then they are ordered by support. Thus apriori favors more confident rules and characterises these rules as more interesting. Support (s): it is an indication of item how frequently it occurs in database. For a rule A=> B, its support is the percentage of transaction in database that contain AUB (means both A and B) [7]. Confidence (c): it indicates the no of times the statements found to be true. Confidence of the rule given above is the percentage of transaction in database containing A that also contain B [7].

III. EXPERIMENT RESULT AND ANALYSIS

In our experiment we present a combination of clustering and association techniques to analyse the factors responsible for recurrence of Breast Cancer in women using WEKA [2] tool. The data set is acquired from UCI Cleveland Repository and converted to corresponding arff format. It consist of 10 attributes and 286 instances. Figure 3 provides an overview of attributes taken for analysis.

Attribute Name	Description
Age	Patient's Age in years
Menopause	the period in a woman's life when menstruation ceases
Tumor-size	Patient's tumor-size on her breast
inv-nodes	Node size in main portion of the breast.
Node-caps	Node is present or not in cap of the breast
Deg-malig	Stage of breast cancer
Brest	Left breast or Right breast or both breast
Breast-quad	Portion of the breast for example left-up, left-low, right-up, right-low, central.
Irradiate	Present or not (YES/NO)
Class	no-recurrence-events, recurrence-events (Reduce the risk of breast cancer)

Fig 3. Breast Cancer Dataset

Kmeans clustering algorithm was implemented on the breast cancer training dataset and the resulting clusters were formed as shown in figure 4 and 5. The database was then segmented into corresponding clusters and apriori algorithm was implemented on each cluster to obtain interesting and well defined rules.

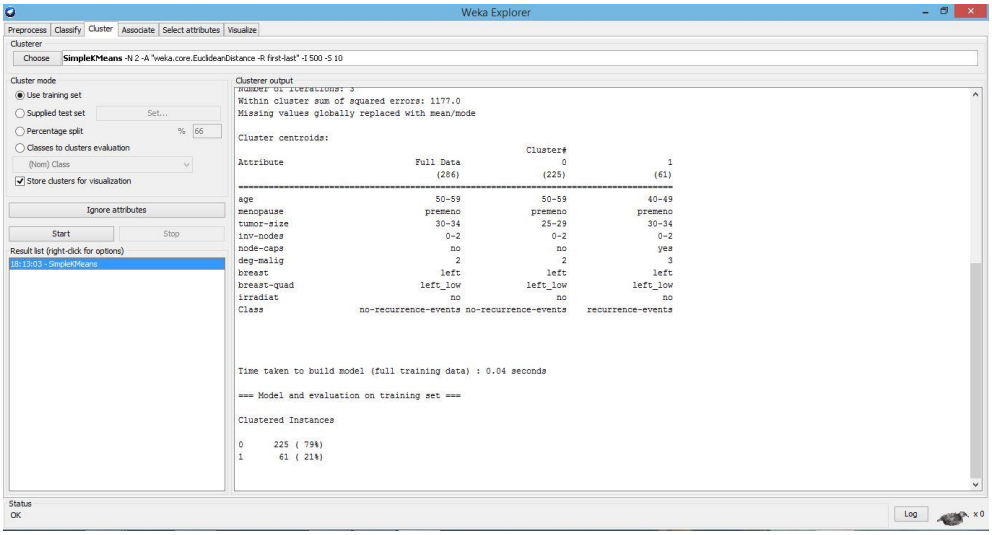

Fig 4. Kmeans Clustering on Dataset

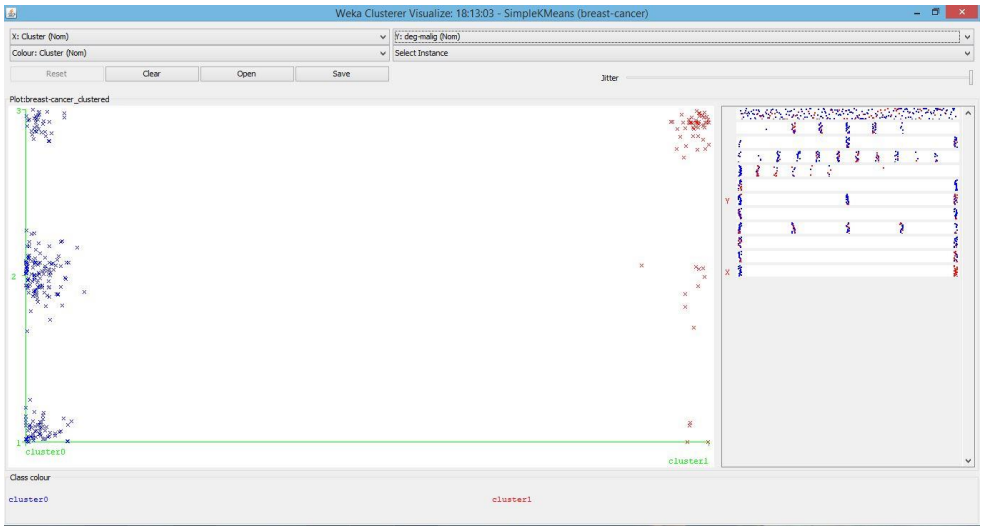

Fig 5. Visualized Cluster Formation

The apriori algorithm was implemented on individual clusters to obtain well defined rules for recurrence and no recurrence of Breast Cancer in women (figure 6 and 7).

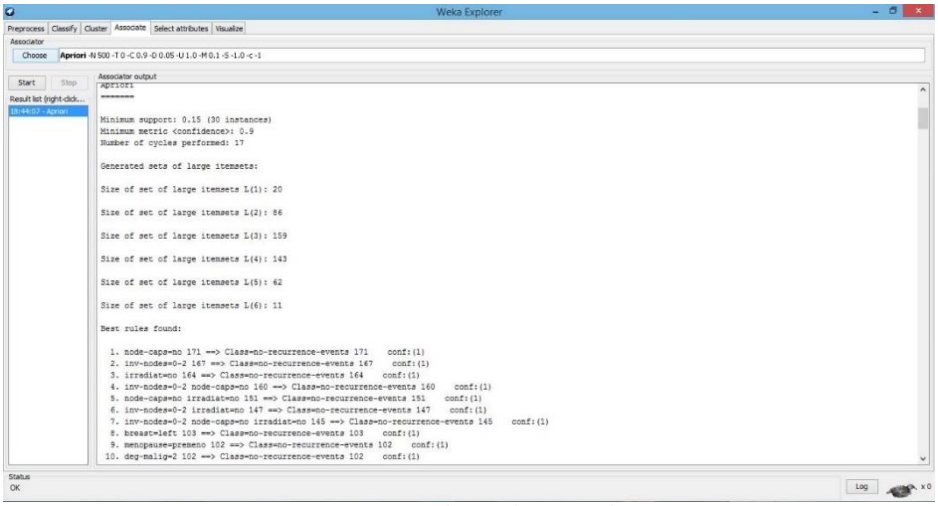

Fig 6. Best association rules (Total of 500) of cluster1 of Breast Cancer Dataset

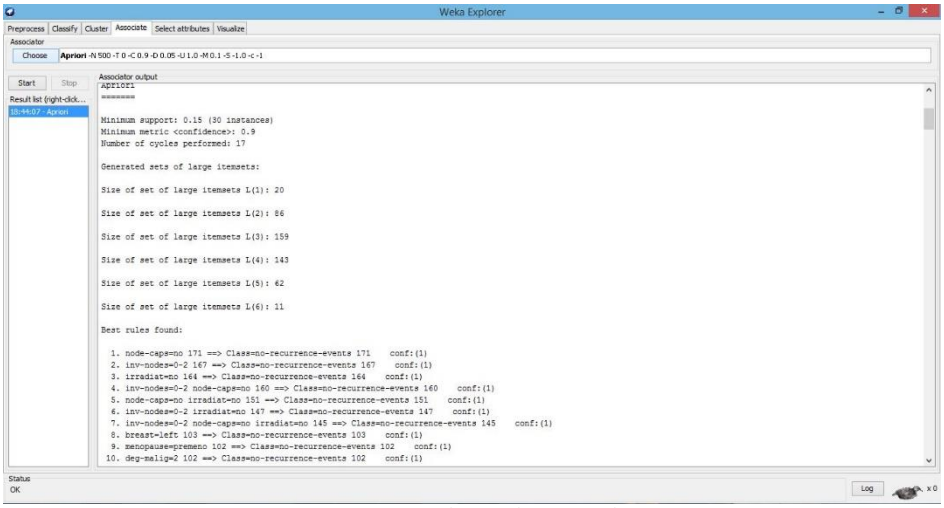

Fig 7. Best association rules (Total of 500) of cluster0 of Breast Cancer Dataset

This approach of segmenting the database provides us association rules for both the clusters and allows us to analyze factors for both the clusters.

IV. CONCLUSIONS

The proposed combination of clustering and association rule mining result in formation of well-defined and content-related rules which provide useful information related to the health of patient and helps to analyze possible measures need to be taken to get cure of Breast Cancer recurrence. As clustering is an unsupervised learning technique therefore, it builds the classes by forming a number of clusters to which instances belongs to, and after manual segmentation of clusters, association is applied to demonstrate the rules formed for each cluster which helps doctors to determine the steps need to be taken so that the treatment can be given to a group of patients at the same time. In our study we observed that the patients with a degree of malignance 3 and with a tumor size greater than 25mm and whose capsular nodes are affected has a higher chance of recurrence of Breast cancer. This combined approach provides us with a two stage health analysis of Breast Cancer patients using Clustering (Kmeans) and association rule mining (Apriori).

REFERENCES

[1] http://en.wikipedia.org/wiki/Data_mining.
[2] Weka 3- Data Mining with open source machine learning software available from:- http://www.cs.waikato. ac.nz/ml/ weka/.
[3] Jain, A.K., Murty M.N., and Flynn P.J., "*Data Clustering: A Review*", 1990.

[4] Holmes, G., Donkin, A., Witten I.H., "*WEKA a machine learning workbench*", In: Proceeding second Australia and New Zeeland Conference on Intelligent Information System, Brisbane, Australia, pp.357-361, 1994.
[5] Rakesh Agrawal, T. Imieliński, A. Swami, "*Mining association rules between sets of items in large databases*", In: Proceedings of the 1993 ACM SIGMOD international conference on Management of data - SIGMOD '93, pp. 207-216, 1993.
[6] http://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R/Clustering/K-Means.
[7] Jeetesh Kumar Jain, Nirupama Tiwari, Manoj Ramaiya / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com Vol. 3, Issue 1, January -February 2013, pp.2065-2069.
[8] Agrawal R., Imielinski, T., and Swami, *"Mining association rules between sets of items in large databases",* In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, 1993.
[9] R. Agrawal and R. Srikant. "*Fast algorithms for mining association rules in large databases*". Research Report RJ 9839, IBM Almaden Research Center, San Jose, California, June 1994.
[10] Han and M. Kamber, Data Mining: Concepts and Techniques, San Francisco: Morgan Kaufmann Publishers,2001.
[11] Jesmin Nahar,Kevin S.Tickle,Shawkat Ali and YI-Ping Phoebe Chen, "*Diagnosis Heart Disease using an Association Rule Discovery Approach*" In Proceedings of the IASTED International Conference Computational Intelligence August 2009.
[12] Shomona Gracia Jacob, R Geetha Ramani *, "Data Mining in Clinical Data Sets: A Review"* International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868 Foundation of Computer Science FCS, New York, USA Volume 4– No.6, December 2012 – www.ijais.org

