Volume 3, Issue 9, September 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Knowledge Discovery from Database using an Integration of Clustering and Association Rule Mining

Ritu Ganda*
Department of Computer Science
J.C.D.M college of Engineering and Technology
Guru Jambheswar University of Science and Technology
India

Abstract—Clustering and Association are two important techniques of data mining. Association rule learning is a well researched method for discovering interesting relations between variables in large databases. It identifies and defines strong rules discovered in databases using different measures of interestingness. While, clustering is an unsupervised learning problem that group objects based upon distance or similarity. Each group formed is known as a cluster. In this paper we make use of a large database 'Kidney Dataset' containing 7 attributes and 154 instances to perform an integration of clustering and association rule mining to determine some essential and interesting rules formed in case of each cluster and can also demonstrate the results as the minimum support changes based on various parameters using WEKA (Waikato Environment for Knowledge Analysis), a Data Mining tool. The results of the experiment show that integration of clustering and Association Rule Mining give some essential and content-related rules.

Keywords— Data Mining; KMEANS; APRIORI; WEKA; Kidney Dataset.

I. Introduction

Data mining is the process of automatic classification of cases based on data patterns obtained from a dataset. A number of algorithms have been developed and implemented to extract information and discover knowledge patterns that may be useful for decision support [8]. Data Mining, also popularly known as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases [2]. Several data mining techniques are pattern recognition, clustering, association, classification and clustering [7]. The proposed work will focus on challenges related to integration of clustering and association rule mining. Association rules were first introduced by Agarwal [12]. Association rules are helpful for analyzing customer behavior in retail trade, banking system, healthcare system etc. Association rule can be defined as {X, Y} => {Z}. It means in retail stores if customer buys X, Y he is likely to by Z. This concept of association rule today used in many application areas like intrusion detection, biometrics, production planning etc. Clustering is the unsupervised classification of patterns into clusters [3].The community of users has played lot emphasis on developing fast algorithms for clustering large datasets [15].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 [16] [10] is used for performing clustering and association algorithms. This paper deals with the use of the integrated clustering-association rule mining on WEKA which results in formation of some interesting rules. The dataset used in this paper is Kidney Dataset, consists of 154 instances and 7 attributes. The paper is organized as follows: Section 2 defines K-Means and Apriori Algorithm. Section 3 defines problem formulation. Section 4 describes an integration of clustering and association rule mining to demonstrate some interesting rules formed and to determine the variation in rules with change in minimum support. Experimental results and performance evaluation are presented in Section 5 and finally, Section 6 concludes the paper and points out some potential future work.

II. K-means and Apriori Algorithm

A. *K-Means Algorithm*

Simple K-Means is one of the simplest clustering algorithms [14]. K-Means algorithm is a classical clustering method that group large datasets in to clusters [1] [13]. The procedure follows a simple way to classify a given data set through a certain number of clusters. It select k points as initial centriods 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.

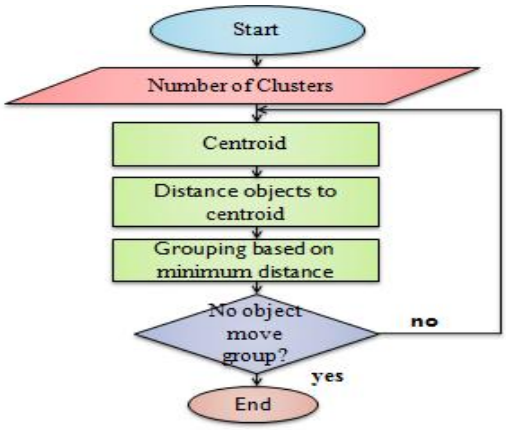

Fig 1 Flowchart of K-means Algorithm

B. Apriori Algorithm

Apriori is the Latin word and its meaning is "from what comes before". Apriori uses bottom up strategy. It is the most famous and classical algorithm for mining frequent patterns. This algorithm works on categorical attributes. Apriori uses breadth first search [4].

1) *Association rule*: Association rule are the statements that find the relationship between data in any database. Association rule has two parts "Antecedent" and "Consequent". Antecedent is the item that found in database, and consequent is the item that found in combination with the first. Association rules are generated during searching for frequent patterns. The problem of finding association rules is divided into two sub problems: first is to find frequent itemsets and second is to find association rules from these itemsets [9]. For important relationships association rule uses the criteria of "Support" and "Confidence" that are explained below:
   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) [5].
   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 [5].

III. Problem Formulation

The problem in particular is to determine and analyze rules formed for each cluster and to demonstrate the changes as the value of minimum support changes for each cluster considering various parameters on integration of clustering(K-means, HAC) and association(Apriori algorithm) using Kidney Dataset consisting of 154 instances and 7 attributes using WEKA, a data mining tool.

IV. Proposed Method

Clustering is the task of segmenting a diverse group into a number of similar subgroups or clusters. Association rules are created by analyzing data for frequent if/then patterns and using the criteria support and confidence to identify the most important relationships. Fig.1 shows a general framework of an integration of clustering and association rule mining. Fig.2 shows the block diagram of formation of rules for each cluster formed. Apply clustering technique on the original data set using WEKA [11] tool and now we are come up with a number of clusters. It also adds an attribute "cluster" to the data set and then the clusters are manually summarized and apply association rule mining on each cluster formed which result in formation of some important and well-defined rules for each cluster for Kidney Dataset. The system involves different consecutive stages communicating with one another in generating rules as the data pre-processing, data partitioning, data transformation, and association rule mining. Before proceeding to the rule mining of datasets, raw data must be pre-processed in order to be useful for knowledge discovery. This approach will result in formation of rules for each cluster formed according to the patient condition which help to determine the necessary steps need to be taken to improve the patient's health.

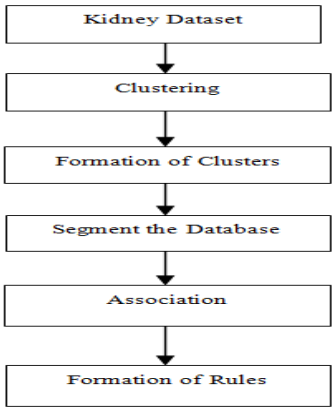

Fig 2 System Architecture

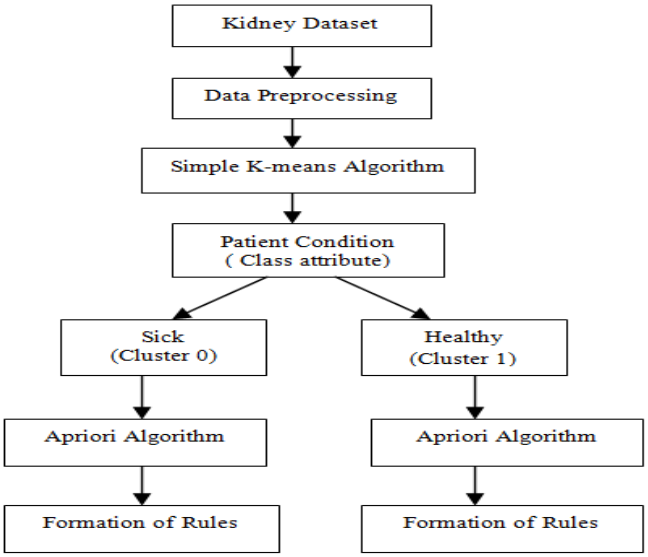

Fig 3 Block Diagram

A. Kidney Dataset Preprocessing

This kidney dataset is collected from Dr. Reema Mehta's Laboratory, Sirsa. It contains 154 instances and 7 attributes. The dataset contains two classes sick and healthy samples. There are 65 samples in the dataset that are assigned to healthy and the other 89 samples are sick. The data collected from the lab consist of 7 attributes as five are continuous attributes i.e. age, bloodurea , creatinine, sodium, potassium and two are discrete attributes i.e. sex and class as shown in table 1 and then the continuous attributes is converted into discrete attributes according to normal range of bloodurea (15-45 mgm/dl), creatinine (0.6-1.4 mgm/dl), sodium (135-155 mEq/Ltr) and potassium (3.3-5.1 mEq/Ltr) as association can be implemented on discrete attribute as shown in table 2 .Complete description of variables is shown in table 3.

TABLE I
SAMPLE INSTANCES FROM KIDNEY DATASET (RAW DATA)

Age	**Sex**	**BloodUrea (mgm/dl)**	**Creatinine (mgm/dl)**	**Sodium (mEq/Ltr)**	**Potassium (mEq/Ltr)**	**class**
60	Female	21	0.9	139	5.3	Healthy
34	Female	70	1.6	143	5.6	Sick
25	Female	24	1.2	143	4.6	Healthy
42	Male	41	0.8	132	5.8	Healthy
54	Female	81	1.6	127	3.4	Sick
27	Male	110	3.9	126	3.1	Sick
53	Male	83	2.3	141	5.3	Sick

TABLE 2
SAMPLE INSTANCES FROM KIDNEY DATASET (TRANSFORM DATA)

Age	**Sex**	**BloodUrea**	**Creatinine**	**Sodium**	**Potassium**	**class**
60-69	Female	15-45	0.6-1.4	135-155	above5.1	Healthy
30-39	Female	above45	above1.4	135-155	above5.1	Sick
20-29	Female	15-45	0.6-1.4	135-155	3.3-5.1	Healthy
40-49	Male	15-45	0.6-1.4	below135	above5.1	Healthy
50-59	Female	above45	above1.4	below135	3.3-5.1	Sick
20-29	Male	above45	above1.4	below135	below3.3	Sick
50-59	Male	above45	above1.4	135-155	above5.1	Sick

TABLE 3
COMPLETE DESCRIPTION OF VARIABLES

Attribute	Category	Information
Age	Discrete	6 values
Sex	Discrete	6 values
Blood Urea	Discrete	6 values
Creatinine	Discrete	6 values
Sodium	Discrete	6 values
Potassium	Discrete	6 values
Class	Discrete	6 values

V. Experiment Results and Performance Evaluation

In this experiment we present an integration of clustering and association rule mining of data mining using Kidney dataset. During simple clustering, the training dataset is given as input to WEKA tool and the clustering algorithm namely K-means was implemented. During an integration of clustering and association rule mining first, Simple K-means clustering algorithm was implemented on the training data set which divides the dataset into two clusters according to class attribute as shown in fig 4. Apriori algorithm was implemented on the resulting clusters formed as shown in fig 5. The results of the experiment show that integration of clustering and association rule mining gives some interesting and well-defined rules for each cluster formed. Table 4 shows the comparison of best association rules with minimum support (0.41 & 0.42) for cluster1 and Table 5 shows the comparison of best association rules with minimum support (0.44 & 0.45) for cluster2.

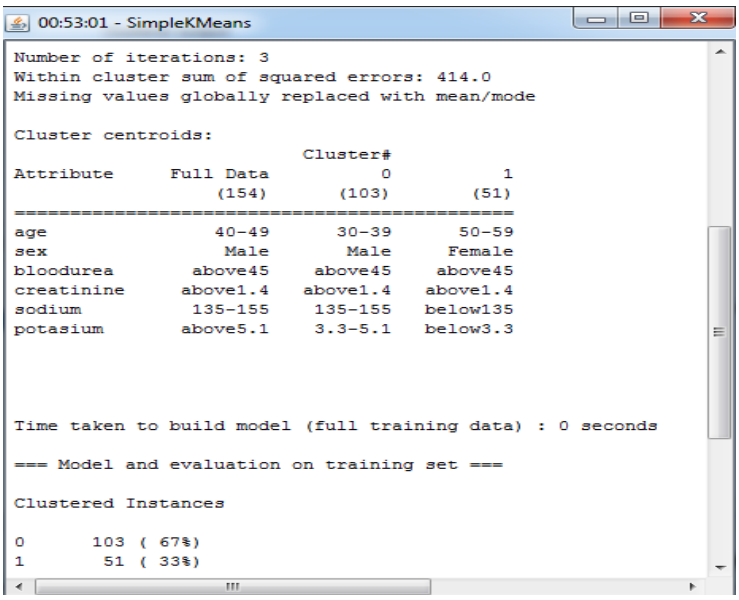

Fig 4 K-means on Weka for Kidney Dataset

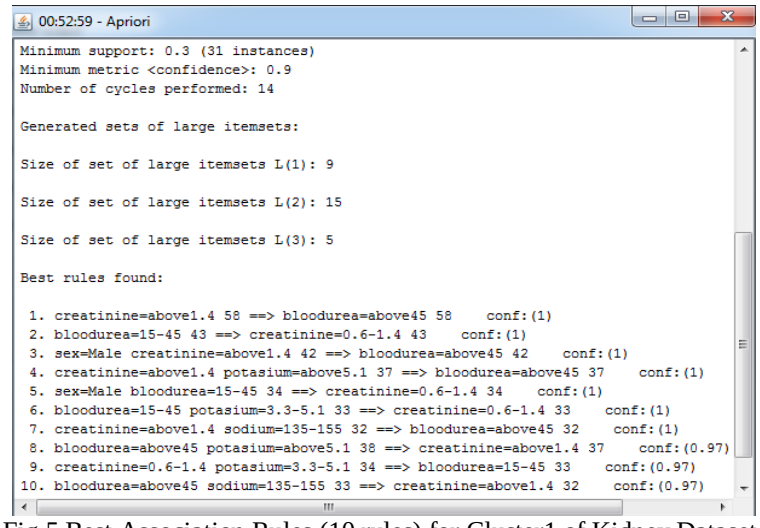

Fig 5 Best Association Rules (10 rules) for Cluster1 of Kidney Dataset

TABLE 4
COMPARISON OF BEST ASSOCIATION RULES WITH MINIMUM SUPPORT (0.41 & 0.42)

Minimum Support = 0.41	**Minimum Support = 0.42**
Minimum support: 0.41 (42 instances) **Minimum metric <confidence>: 0.9** **Number of cycles performed: 12** **Best rules found:** 1. creatinine=above1.4 ==> bloodurea=above45 2. bloodurea=15-45 ==> creatinine=0.6-1.4 3. sex=Male creatinine=above1.4 ==> bloodurea=above45 4. bloodurea=above45 ==> creatinine=above1.4 5. creatinine=0.6-1.4 ==> bloodurea=15-45 6. sex=Male bloodurea=above45 ==> creatinine=above1.4	**Minimum support: 0.42 (43 instances)** **Minimum metric <confidence>: 0.9** **Number of cycles performed: 12** **Best rules found**: 1. creatinine=above1.4 ==> bloodurea=above45 2. bloodurea=15-45 ==> creatinine=0.6-1.4 3. bloodurea=above45 ==> creatinine=above1.4 4. creatinine=0.6-1.4 ==> bloodurea=15-45

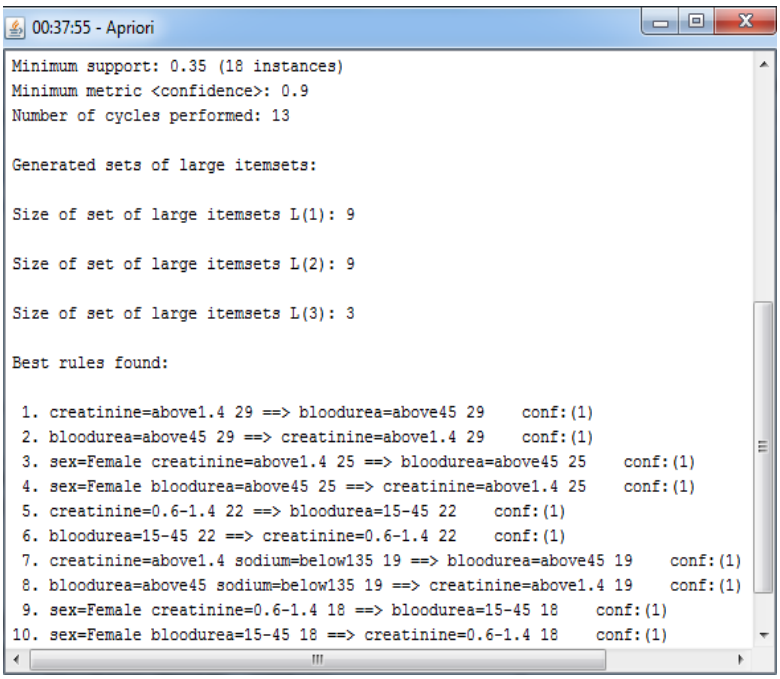

Fig 6 Best Association Rules (10 rules) for Cluster2 of Kidney Dataset

TABLE 5
COMPARISON OF BEST ASSOCIATION RULES WITH MINIMUM SUPPORT (0.44 & 0.45)

Minimum Support : 0.44	**Minimum Support : 0.45**
Minimum support: 0.44 (22 instances) **Minimum metric <confidence>: 0.9** **Number of cycles performed: 12** **Best rules found**: 1. creatinine=above1.4 ==> bloodurea=above45 2. bloodurea=above45 ==> creatinine=above1.4 3. sex=Female creatinine=above1.4 ==> bloodurea=above45 4. sex=Female bloodurea=above45 ==> creatinine=above1.4 5. creatinine=0.6-1.4 ==> bloodurea=15-45 6. bloodurea=15-45 ==> creatinine=0.6-1.4	**Minimum support: 0.45 (23 instances)** **Minimum metric <confidence>: 0.9** **Number of cycles performed: 11** **Best rules found:** 1. creatinine=above1.4 ==> bloodurea=above45 2. bloodurea=above45 ==> creatinine=above1.4 3. sex=Female creatinine=above1.4 ==> bloodurea=above45 4. sex=Female bloodurea=above45 ==> creatinine=above1.4

The proposed integration 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 kidney problem.

VI. Conclusion and Future Scope

The presented experiments shows that integration of clustering and association rule mining give more accurate and well-defined rules in case of each cluster formed for kidney dataset on WEKA. 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 summarization of clusters association is applied to demonstrate the rules formed for each cluster which helps doctors to determine the steps need to be taken and also to determine the symptoms of kidney problem so that the treatment can be given to a group of patients at the same time. This study proposes a new two-stage framework of patient health analysis using clustering technique(K-means) and an association rule mining(Apriori) for analyzing kidney datasets. Furthermore, it was demonstrated that some interesting content-related rules can be discovered from the integrated kidney data, while they could not be discovered using only the standard kidney dataset. This study clearly shows that data mining techniques are promising for clinical datasets. Our future work will be related to missing values and to develop an automatic summarization technique. In addition, the research would be focusing on integration of other clustering algorithm and association algorithm.

References

[1] L.Kaufinan, and P.J Rousseeuw, "Finding groups in Data: an Introduction to Cluster Analysis", John Wiley & Sons 1990.
[2] J. Han and M. Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufmann, 2000.
[3] Jain, A.K., Murty M.N., and Flynn P.J., "Data Clustering: A Review", 1990.
[4] Jaishree Singh, Hari Ram, Dr. J.S. Sodhi, "Improving Efficiency of Apriori Algorithm Using Transaction Reduction", In: proceeding of International Journal of Scientific and Research Publication (IJSRP), ISSN 2250-3153, Vol 3, Issue 1, January 2013.
[5] Jogi.Suresh, T.Ramanjaneyulu, "Mining Frequent Itemsets Using Apriori Algorithm", In: Proceeding of International Journal of Computer Trends and Technology, ISSN 2231-2803, Vol. 4, Issue 4, April 2013.
[6] Paul Agarwal, M.Afsar Alam, Ranjit Biswas. "Analyzing the agglomerative hierarchical Clustering Algorithm for Categorical Attributes*"* ,International Journal of Innovation, Management and Technology, vol.1, No.2, ISSN: 2010- 0248, June 2010.
[7] Ritu Chauhan, Harleen Kaur, M.Afshar Alam, "Data Clustering Method for Discovering Clusters in Spatial Cancer Databases", International Journal of Computer Applications (0975 – 8887) Volume 10– No.6, November 2010.
[8] Desouza, K.C., "Artificial intelligence for healthcare management", In Proceedings of the First International Conference on Management of Healthcare and Medical Technology Enschede, Netherlands Institute for Healthcare Technology Management, 2001.
[9] Han, J., Pei, and Yin, "Mining frequent patterns without candidate generation", In: proceeding of the 2000 ACM SIGMOD International Conference on Management of Data, pp.1-12. ACM Press, New York, 2000.
[10] 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.
[11] Garner S.R. "WEKA: The Waikato Environment for Knowledge Analysis Proc", New Zealand Computer Science Research Students Conference, University of Waikato, Hamilton, New Zealand, pp 57-64, 1995.
[12] 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.
[13] M.S Chen, J.Han, and P.S.Yu., "Data mining: an overview from database perspective", IEEE Trans. On Knowledge and Data Engineering, 5(1): 866-833, Dec 1996.
[14] I. K. Ravichandra Rao, "Data Mining and Clustering Techniques," DRTC Workshop on Semantic Web, DRTC, Bangalore, Paper: K, 8th – 10th December, 2003.
[15] U.M Fayyad and P. Smyth., "Advances in Knowledge Discovery and Data Mining", AAAI/MIT Press, Menlo Park, CA, 1996.
[16] Weka 3- Data Mining with open source machine learning software available from:- http://www.cs.waikato. ac.nz/ml/ weka/
[17] V. Kumar, J.K. Chhabra and D. Kumar, "Effect of Harmony Search Parameters' Variation in Clustering", Procedia Technology, 6, pp. 265-274,2012.
[18] V. Kumar, J.K. Chhabra and D. Kumar, "Initializing Cluster Center for K-Means Using Biogeography Based Optimization", Advances in Computing, Communication and Control, pp. 448-456, 2011.

