Int. J. Advance. Soft Comput. Appl., Vol. 5, No. 3, November 2013
ISSN 2074-8523; Copyright © SCRG Publication, 2013

A Proficient System for Automatic Detection of Risk Level in Disease Detection using Association Rule Based DRF Algorithm

B. Gomathy[1], S. M. Ramesh[2], A. Shanmugam[3]

[1]Research Scholar, Anna University, Chennai:
[2]Associate Professor, Department of Electronics and Communication Engineering, Bannari Amman Institute of Technology, Erode, India:
[3]Professor, Department of Electronics and Communication Engineering, Bannari Amman Institute of Technology, Erode, India.

Abstract

A challenging research problem for researchers is predicting heart problem, breast cancer, tumor, and the most daunting diseases. Current research in this area is struggling to provide accurate and better solution for the prediction of such deadly diseases. In this paper, Discriminative Rule Framing (DRF) algorithm is proposed to analyze and predict the survivability of disease in a patient. Association rule of data mining is used to reveal the biological hidden patterns and derive association rules from a huge medical data set. Initial rules generated through association rule mining along with subset attributes of the data set are given as input to the DRF risk analysis system to predict the risk level of a given data set. The significance of the DRF is evaluated using confidence, support and lift metrics. Experimental result shows that, prediction level of the DRF is more accurate than other existing algorithms.

Index terms *-Association Rule Mining, Feature Matching, Risk Analysis, Convictional Measures, CART, Machine Learning.*

1. Introduction

To discover useful information from huge data set is a tedious job. To assist discovering useful information profound technique named data mining is used. It retrieves ideas from plenty of disciplines namely statistics, database system, etc. Digitized format of storing information gives a hand for medical department to store and maintain patient's information in a database. An electronic method of storing information is economically feasible. This characteristic of information storage, simulate modern medicine to generate the enormous amount of health care data. The information contained in medical data set is interesting and useful for diagnosis of diseases and patient care.

Models can be designed using data mining for finding patterns in data. There exists a need for a classifier in order to predict serious human disease. Nowadays, physician uses the classifier model to diagnose the diseases. Therefore, to analyze huge data sets, association rules of data mining, is used to refine interesting associations, casual constructions, correlations, frequent patterns, etc indicating the relationship between procedures performed on patients and generated report for diagnose. Most threatening diseases such as brain tumor, breast cancer, etc., detection in earlier stage will increase the survival of patients. Massive data analysis research work is carried out in detecting such diseases using different data mining algorithms. Sensitivity and specificity are improved to increase the survival rate of patients and also decreases the workload of a radiologist.

In this paper in order to determine the existence of disease and its risk level, an algorithm named DRF is introduced. Initially preprocessing process is performed using normalization techniques for the data set. This preprocessing work will enhance the association rule mining to discover medically significant rules by assigning weights from a huge set of medical data sets. The DRF algorithm has two stages. At stage 1, class-labels are framed from the preprocessed data set, based on which base rule for DRF algorithm is generated. As with this approach, rule formation is based on user need, and it can be adaptable to any kind of medical data set. The S - grid takes the base rule and a frame heuristic matrix. This matrix is easily accessible, and it plays a vital role in framing true and branch rules. For each item in the data set s-count value is calculated depending on the values in heuristic matrix.

In the second stage of the algorithm, true and base rules are framed. Based on these rules maximum and minimum values are calculated and heuristic rate is estimated. Along with these values, the threshold is set to filter the items in a data set according to the requirement of risk analysis. Fig. 1 shows the flow of the DRF algorithm. Risk analysis is the most widely used tool by many data mining methods for defining and analyzing of the undesirable events. Medical data set usually holds millions and millions of records. To analyze a collection of records manually consumes more time and also difficult to process all such types of data. Therefore, a

concept in data mining named risk analysis helps to analyze a huge amount of data in an easy manner.

Three interrelated components are encompassed in risk analysis. (1) Risk assessment (2) Risk perception (3) Risk management is undertaken to generate the report of a given data set. The generated data set can be either quantitative or qualitative. Probability determination of the different adverse events and/or the extent of the losses as an effect of a particular event that takes place is referred as quantitative risk analysis. Unlike quantitative risk analysis deals with numerical probabilities; quantitative risk analysis does not consider numerical probabilities. Instead, it involves in defining various threats and/or the extent of the vulnerabilities.

In this method, DRF algorithm provides a processed data set to risk analysis. Depending on the threshold value assigned, the process of finding risk level differs. Therefore, additional care should be given to assign the threshold value. This process will highly reduce the work of physicians and radiologist from manually determining the massive data set and also reduces the time required to process the items of a given data set.

The rest of this paper is structured as follows: Section 2 provides reviews related to the work. Section 3 presents the proposed work, including the DRF algorithm description. Section 5 experimentally evaluates the proposed work with the existing work. Finally, section 6 concludes the paper.

2. Related Works

Many works have been carried out for investigation and analysis of most daunting diseases. Here, some works of various authors were presented related to the proposal. Analyzing and determining risk factors in the medical data set was a mind-blowing work, which consumes time. To overcome this problem, Ince [1] defined over feed-forward, fully-connected artificial neural networks, which have been broadly used in computer-aided decision support systems in medical domain. Moreover, two popular neural network training methods are discovered: conventional backpropagation and particle swarm optimization. Habashy [2] focused on (GEP) Global gene expression profiling of breast cancer have recognized separate biological classes with different clinical and therapeutic inferences. Review of methodologies for artificial neural networks and multivariable logistic regression for differentiating between malignant and benign lung nodules on CT scans was presented in Chen [3].Palaniappan[4] probed an automatic diagnosis system for predicting breast cancer based on association rules and neural network. Moreover, AR1 and AR2 are used for reducing the dimension of breast cancer dataset and NN is used for intelligent classification, which were trained on the attributes of each record in the Wisconsin breast cancer database. Furthermore, the authors of [5] have taken the advantage of association

rules to reduce the dimension of given data set, and neural network was used for effective classification. The experimental results in [5] proved that decision system with neural network with association rule achieved 95.6% of accuracy in classification. A fuzzy rough set method for prototype selection, focused on optimizing the behavior of this classifier, was presented in [6]. This data reduction process is specifically designed to improve the performance of the 1-NN classifier, mutually regarding test accuracy and computational complexity. Nguyen [7]summarized diagnosing and prognosticating breast cancer with a machine learning method based on random forest classifier and feature selection technique. By using Wisconsin Breast Cancer Dataset (WBDC) results showed that this method has accurate detection than other methods. Ex-DBC (Expert system for Diagnosis of Breast Cancer) a diagnostic tool was shown in [8] for diagnosing breast cancer. It used Neuro-fuzzy method and diagnosed with 96% and 81% positive and negative predictive accuracy level respectively. A swarm intelligence technique based maintain vector machine classifier (PSO-SVM) is projected for breast cancer diagnosis [9]. In the devised PSO-SVM, the concern of representation selection and feature selection in SVM is concurrently solved under particle swarm framework. In addition, (TVAC) time variable acceleration coefficients and (TVIW) inertia weight is employed to efficiently control the local and global investigate in PSO algorithm. An automated method for classification of medical data set through the help of quantitative measurement and machine learning technique were emerged hugely. Support Vector Machine (SVM) was such a method, used in [10] for classification. Corresponding results stipulate the superiority of SVM in terms of sensitivity, specificity and accuracy. In [11] wasp swarm optimization algorithm was put in forth for attribute reduction based on rough set and the significance of feature. The significance of feature is constructed based on the mutual information between selected decisional attributes and conditional attributes. The algorithm dynamically computes heuristic information based on the significance of feature to guide search. Results demonstrate that, in terms of computational effort and solution quality, the algorithm can get better results than other intelligent swarm algorithms for attribute reduction. A strategy based on Rough Set Theory (RST) with Particle Swarm Optimization (PSO) was presented in [12]. Rough Set Theory has been predictable to be powerful tools in the medical feature selection. Consequently, a rough set feature selection algorithm based on a search method called Particle Swarm Optimization (PSO) is presented. It is used to select feature subsets to portray the decisions as well as the unique feature set and superfluous the redundant features leading to better prediction accuracy. Results showed that a hybrid approach can help improving classification accuracy, and also finding more robust features to improve classifier performance. In [13] author found a new data mining system for classification of myocardial infarction (MI), percutaneous coronary intervention (PCI), and coronary

artery bypass graft surgery (CABG) events based on decision trees. Experiments were carried out, and the study showed that the system achieved 66% of accurate classification for the MI, 75% of the exact classification for the PCI, and 75% of perfect classification for the CABG events. The profound works of [13] paved a way for designed a new methods such as in [14, 15]. The support vector classifier was integrated with rough set and framed RS_SVM [14] model for diagnosis of breast cancer. Rough set was used in [14] for feature selection. Redundant attributes were removed through rough set and classification accuracy was improved through SVM. The experimental study demonstrated RS_SVM achieves higher accuracy along with detection information about five different features. Chang-Sik Son, et al. In[15] combined decision tree along with rough set techniques for early diagnose of congestive heart failure. Rough set based decision tree achieved classification accuracy of 97.5%.

3. Proposed Methodology

Association Rules (AR) are employed to discover interesting relations among variables in a large data set. AR is widely applicable in the market based analysis to procure frequent item sets. Profitably it can be suitable for medical data set, where clinical data can be maintained electronically as a huge data set. Statements in association rules are expressed as $\{X_1, X_2, X_3, X_4 \cdots\cdots\} \Rightarrow Y$, meaning that if LHS exist, then maximum chance of occurrence of Y is present. Applying AR in medical data set requires preprocessing of data set to remove missing attributes. Preprocessing steps are pertained for easy and accurate prediction of risk factors. In this paper, a well-known technique named normalization is applied for preprocessing the data set. The preprocessed data set is given as input to the DRF, a two stage algorithm. DRF algorithm at stage one uses AR for rule formation. Based on rules generated DRF determines the risk factor at its second stage.

3.1 DRF: Stage 1

3.1.1 Class label formation:

A preprocessed medical data set contains features set such that $F = \{f_1, f_2, f_3, \ldots\ldots\ldots\ldots, f_n\}$ representing 'n' number of features in the data set D. Each feature contains class labels represented as $\delta_{i1,}\delta_{i2,}\delta_{i3,} \ldots \neq \infty$ where, i denotes i^{th} feature in F. To build robust decision making tool subset feature selection is required. It is more important set in analyzing a huge medical data set, which helps in predicting the outcome. Feature section also improves the performance of

prediction of the risk level of each item$\{I_1, I_2, I_3, \ldots\ldots\ldots\ldots, I_n\}$ in data set D, since prediction may not scale up to the full feature set F. Such features acted as the major role in prediction of diseases. In this approach Weka tool's ChiSquaredAttributeEVal method [16] is used for the best feature section. Depending on this method we obtain $DF = \sum_{i=1}^{j} f_i$,a subset of features. Along with features, single class label is chosen for each feature in DF.

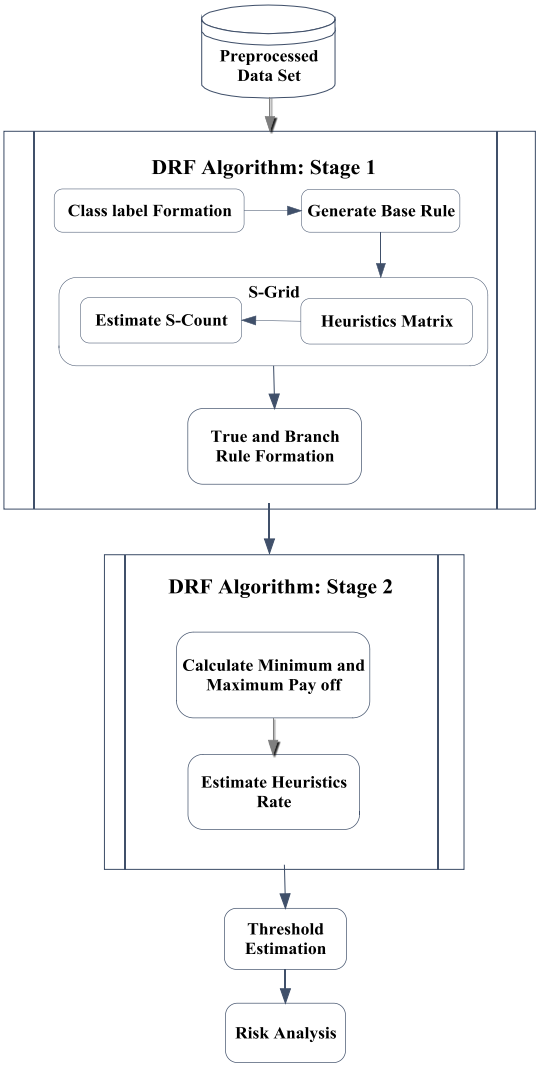

Fig. 1 DRF Flow Diagram

3.1.2 Framing Base Rule:

Base rules (BR) are framed from the class labels that are selected from the above step. The base rule can be represented

as $BR(f_1 = \delta_{11} \,||\, f_2 = \delta_{25} \,||\, f_3 = \delta_{32} || \ldots\ldots || \, f_j = \delta_{j9})$. Each feature has various class labels. Depending on the analysis, criteria for feature differs. Class labels that are chosen will act as the criteria for prediction of risk level. Therefore, forming base rule entirely depends on the decision of the class label. These are the most preliminary rule framed by DRF. Further, well refined decision making rules are generated later in the process but all which depends on this preliminary rule. So it is important to correctly frame this rule.

3.1.3 Support-Grid (S-Grid):

S-Grid is created in order to generate highly refined rule for prediction. Having BR as basic rule, S-Grid frames a matrix named heuristics matrix (HM) and S_{count} value is calculated. For a given data set D with 'n' number of item set (IS), such that $IS = \{I_1, I_2, I_3, \ldots\ldots\ldots\ldots, I_n,\}$, S-grid frame $n\ x\ j$ HM matrix where, *j*depicts the number of features in BR and *n*symbolizedthe number of items in D. Values for the matrix are calculated by comparing the BR's δ_{ix}. For example, the value of HM_{2x3} is computed through comparing the class label of itemI_2's feature with the BR's $f_3$class label value δ_{3x}. If the class label of I_2 is equal to f_3's class label of BR then HM_{2x3} contains the value 1. Otherwise HM_{2x3} entry has 0 as its value. Likewise, equation 1 is used to calculate the values of HM matrix for all items in D and features in BR.

$$HM_{X\,x\,Y} = \begin{cases} 1, & \text{if CL of } I_X = \text{CL of } f_Y \\ 0, & \text{if CL of } I_X \neq \text{CL of } f_Y \end{cases} \qquad (1)$$

In equation 1, I_X represents the x[th] item set in D, $f_Y$denotes the y[th]feature in BR and CL is the class label.

Using the HM matrix value, S_{count} value is determined for every item in D using the equation (2). S_{count}value for an item is the summation value HM entry to all the features in BR for that particular item.

$$S_{count_i} = \sum_{x=1}^{j} HM_{ix} \qquad (2)$$

Here, $S_{count_i}$is the S_{count} value of $I_i$record in D.

3.1.4True and Branch Rule formation:

True rule (TR_i) and branch rule (BrR_i) are the most sophisticated rules formed from HM matrix by DRF algorithm with an item I_i. $TR_i$and$BrR_i$are the subset of BR,

mathematically it is represented as $TR_i \subset BR$ and$BrR_i \subset BR$. Furthermore, $TR_i \cup BrR_i = BR$. TR_ihas the features that have value equal to one in the HM matrix for an item I_i. All other features will be in attendance in BrR_i. For example, consider $f_1 = \delta_{1.1} || f_2 = \delta_{2.5} || f_3 = \delta_{3.2} || f_4 = \delta_{4.9} || f_5 = \delta_{5.0} || f_6 = \delta_{6.1} || f_7 = \delta_{7.3}$ as BR and an item I_1 has $HM_{11} = 1, HM_{12} = 0, HM_{13} = 0, HM_{14} = 1, HM_{15} = 0, HM_{16} = 0, HM_{17} = 0$ values in HM matrix then TR_1 is constructed as$f_1 = \delta_{1.1} || f_4 = \delta_{4.9}$. BrR_1contain features that are there in BR but not in TR_1 i.e. $BrR_1 = BR - TR_1$. $f_2 = \delta_{2.0} || f_3 = \delta_{3.5} || f_5 = \delta_{5.1} || f_6 = \delta_{6.2} || f_7 = \delta_{7.6}$will be the BrR_1. Class labels in BR and TR_1 are same, whereas BR and BrR_1class label values are not same.

3.2DRF: Stage 2

3.2.1 Calculate Maximum and Minimum payoff:

Minimum and maximum pay off values are the key values which are used for decision making. To acquire these values, a new rule named significant rule (SR_i) is formulated. $SR_i$is framed by combining the TR_i and BrR_i. From the previous example SR_1 can be framed as $f_1 = \delta_{1.1} || = \delta_{2.0} || f_3 = \delta_{3.5} || f_4 = \delta_{4.9} || f_5 = \delta_{5.1} || f_6 = \delta_{6.2} || f_7 = \delta_{7.6}$.

$TR_1 \rightarrow f_1 = \delta_{1.1} \parallel f_4 = \delta_{4.9}$

$BrR_1 \rightarrow f_2 = \delta_{2.0} \parallel f_3 = \delta_{3.5} \parallel f_5 = \delta_{5.1} \parallel f_6 = \delta_{6.2} \parallel f_7 = \delta_{7.6}$

$\therefore$ $SR_1 \rightarrow f_1 = \delta_{1.1} \parallel \delta_{2.0} \parallel f_3 = \delta_{3.5} \parallel f_4 = \delta_{4.9} \parallel f_5 = \delta_{5.1} \parallel f_6 = \delta_{6.2} \parallel f_7 = \delta_{7.6}$

It is to be noted that the second feature (f_2) is represented by its class label ($\delta_{2.0}$) while forming SR_1. The class label of SR_i's feature is compared tofeatures of all item set's class labels in original data set D. Total number of feature's class label equal among significant rule and original data set is said to be support value. For each item in D $\{I_1, I_2, I_3, \ldots\ldots\ldots\ldots, I_n\}$has a support value SV_i. For an example, consider a significant rule SR_1's features and corresponding class labels as $f_1 = \delta_{1.1} || = \delta_{2.0} || f_3 = \delta_{3.5} || f_4 = \delta_{4.9} || f_5 = \delta_{5.1} || f_6 = \delta_{6.2} || f_7 = \delta_{7.6}$and I_1's original data set features and class labels as $f_1 = \delta_{1.1} || = \delta_{2.6} || f_3 = \delta_{3.1} || f_4 = \delta_{4.4} || f_5 = \delta_{5.1} || f_6 = \delta_{6.2} || f_7 = \delta_{7.9}$. In this case, SV_1 value is 3. Likewise for SR_1, support values for $\{I_1, I_2, I_3, \ldots\ldots\ldots\ldots, I_n\}$in D is calculated as $\{SV_1, SV_2, SV_3, \ldots\ldots\ldots\ldots, SV_n\}$.

For better and efficient risk analysis, check $\{SV_1, SV_2, SV_3, \dots\dots\dots, SV_n\}$with a value, F_{count} for SR_1, where F_{count}value is the one and the same to total number of features in BR divided by 2. The significant rules are benchmarked based on support values and F_{count}. This support value calculation is carried for all$\{SR_1, SR_2, \dots\dots\dots SR_n\}$ and compared with F_{count}. A R_{count_i} value is calculated using equation 3 for$\{SR_1, SR_2, \dots\dots\dots SR_n\}$.

$$R_{count_i} = \begin{cases} \sum_{i=1}^{n} R_{count_i} + 1, & \text{if } F_{count} \leq SV_i \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

Likewise, another variable max_i is estimated. It holds the highest value of support value for a significant rule SR_i. (i=1, 2… ……n i.e. 'n' number of items present in D).

Similarly Sum_i a variable value is obtained through the continuous addition of support values of SR_i, for each item $\{I_1, I_2, I_3, \dots\dots\dots, I_n\}$in D whose values are greater than F_{count}. From the above computed values minimum, maximum, and significant pay off SR_i can be carried out through equation 4, 5, and 6 respectively.

$$Min_{payoff_i} = \frac{R_{count_i}}{T_{item}} \quad (4)$$

$$Max_{payoff_i} = \frac{Sum_i}{T_{item}} \quad (5)$$

$$Sig_{payoff_i} = \frac{max_i}{T_{item}} \quad (6)$$

In foresaid equations T_{item}represents total number of transactions.

Association rules are best in data mining. However, there exist disadvantages such that it generates a mountainous amount of rules. Furthermore, association rules sometimes do not take sequential information that is available in some data set. In interest to trim down the number of rules, two conviction measures confidence and lift are framed. For enumerating these conviction measures, support value for true rule $SV_{Ti}$and branch rules SV_{BrRi} are calculated independently. SV_{Ti}andSV_{BrRi} are the number of features in the original data set that support $TR_i$and BrR_i features respectively. Consider an example, Let, $f_1 = \delta_{1.1}||= \delta_{2.6}||f_3 = \delta_{3.1}||f_4 = \delta_{4.4}||f_5 = \delta_{5.1}||f_6 = \delta_{6.2}||f_7 = \delta_{7.9}$ be the

features and class labels of original data set, $f_1 = \delta_{1.1} || f_4 = \delta_{4.9}$ and $f_2 = \delta_{2.0} || f_3 = \delta_{3.5} || f_5 = \delta_{5.1} || f_6 = \delta_{6.2} || f_7 = \delta_{7.6}$ are the features and class labels of TR_1 and BrR_1 respectively. Then the $SV_{T1} = 1$ and $SV_{BrR1} = 2$. Depends on this $R_{Tcount_i}$and $R_{BrRcount_i}$values are determined as from the equation 7 and 8 by comparing $SV_{Ti} = 1$ and F_{Tcount} for $rcnt$ and SV_{BrRi} and $F_{BrRcount}$. Where, $F_{Tcount}$is the total number of features in TR divided by 2 and $F_{BrRcount}$ is the total number of features in BrR divided by 2. This process is carried out for all true rules and base rules i.e. $\{TR_{1,}TR_{2,}TR_{3,} \dots \dots \dots \dots, TR_{n,}\}$and $\{BrR_{1,}BrR_{2,}BrR_{3,} \dots \dots \dots \dots, BrR_{n,}\}$.

$$R_{Tcount_i} = \begin{cases} \sum_{i=1}^{n} R_{Tcount_i} + 1, & \text{if } F_{Tcount} \leq SV_{Ti} \\ 0, & \text{otherwise} \end{cases} \quad (7)$$

$$R_{BrRcount_i} = \begin{cases} \sum_{i=1}^{n} R_{BrRcount_i} + 1, & \text{if } F_{BrRcount} \leq SV_{BrRi} \\ 0, & \text{otherwise} \end{cases} \quad (8)$$

CM_1 and CM_2values can be defined as in equation 9 and 10 respectively.

$$confidence_i = \frac{f_x + f_y}{f_x} \quad (9)$$

$$lift_i = \frac{f_{xi} + f_{yi}}{f_{xi} * f_{yi}} \quad (10)$$

Where, f_x and $f_y$can be obtained from equation 11 and 12

$$f_{xi} = \frac{R_{Tcount_i}}{T_{item}} \quad (11)$$

$$f_{yi} = \frac{R_{BrRcount_i}}{T_{item}} \quad (12)$$

Confidence and lift are the two conviction measures that are used to indicate how reliable and important the rules are. Based on confidence measure little pruning is accomplished in the rule set.

3.2.2 Estimate Heuristics Rate:

Computed minimum, maximum and significant rules that are derived from 3, 4, and 5 are taken to determine Heuristics Rate (HR) derived through the equations 13 and 14.

$$ds_i = 0.5 * Min_{payoff_i} * Max_{payoff_i} + (1 - 0.85) * Sig_{payoff_i} \quad (13)$$

$$HR_i = ds_i + Min_{payoff_i}\left(1 - Min_{payoff_i}\right) \quad (14)$$

HR values obtained for all the data items are calculated and arranges the rules in ascending order according to the HR value of each item. Find the median value among the HR values of all items. The median value of HR will act as the analyzing parameter for predicting risk factor. Items, whose value equal to, or lower than or higher than the median value of HR then it is predicted risk factor as moderate, low and higher risks respectively.

4. Analysis of DRF Algorithm

To examine the projected DRF algorithm, dataset of breast cancer is extracted from UCI learning repository [17]. The data set contains 280,660 records of different patients with 16 features ($\{f_1, f_2, f_3, \ldots\ldots\ldots, f_{16}\}$). This study is carried out for multiple times with 5,000 dissimilar records every time. Each time results obtained is similar in predictions. Features and class labels used in this dataset of UCI are given in Table 1.

Before employing DRF algorithm for the data set D, normalization based preprocessing method is applied. Preprocessing is the most needed step in the decision making process. Since, a dataset may hold missing values, noisy data, etc. In order to remove such data, the preprocessing is applied. Feature selection is the most important step in the prediction algorithm because the accuracy of decision making depends

Table 1 Features & class labels of UCI repository

Feature Representation	Features	Class Label Representation	Class Labels
f_1	menopaus	$\delta_{1.0}, \delta_{1.1}$	0,1
f_2	agegrp	$\delta_{2.1}, \delta_{2.2} \ldots\ldots \delta_{21.0}$	1, to 10
f_3	density	$\delta_{3.1}, \delta_{3.2}, \delta_{3.3}, \delta_{3.4}, \delta_{3.9}$	1, 2, 3, 4, 9
f_4	race	$\delta_{4.1}, \delta_{4.2}, \delta_{4.3}, \delta_{4.4}, \delta_{4.5}, \delta_{4.9}$	1, 2, 3, 4, 5, 9
f_5	Hispanic	$\delta_{5.0}, \delta_{5.1}, \delta_{5.9}$	0, 1, 9
f_6	bmi	$\delta_{6.1}, \delta_{6.2}, \delta_{6.3}, \delta_{6.4}, \delta_{6.9}$	1, 2, 3, 4, 9
f_7	agefirst	$\delta_{7.0}, \delta_{7.1}, \delta_{7.2}, \delta_{7.9}$	0, 1, 2, 9
f_8	nrelbc	$\delta_{8.0}, \delta_{8.1}, \delta_{8.2}, \delta_{8.9}$	0, 1, 2, 9
f_9	brstproc	$\delta_{9.0}, \delta_{9.1}, \delta_{9.9}$	0, 1, 9
f_{10}	lastmamm	$\delta_{10.0}, \delta_{10.1}, \delta_{10.9}$	0, 1, 9
f_{11}	surgmeno	$\delta_{11.0}, \delta_{11.1}, \delta_{11.9}$	0, 1, 9
f_{12}	hrt	$\delta_{12.0}, \delta_{12.1}, \delta_{12.9}$	0, 1, 9
f_{13}	invasive	$\delta_{13.0}, \delta_{13.1}$	0, 1
f_{14}	cancer	$\delta_{14.0}, \delta_{14.1}$	0, 1
f_{15}	training	$\delta_{15.0}, \delta_{15.1}$	0, 1
f_{16}	count	$\delta_{15.1} \ldots\ldots \delta_{15.15}$	1 to 15

upon the features selected. A strong prediction system is generated only through best feature selection.

Using the features and class labels, of table 1, a subset of features and class labels to frame BR is employed, which acts as the pedestal for framing further rules and analyzing. Let, Base rule is constructed as$f_3 = \delta_{3.5} || f_4 = \delta_{4.4} || f_5 = \delta_{5.9} || f_6 = \delta_{6.2} || f_7 = \delta_{7.9} || f_8 = \delta_{8.2} || f_9 = \delta_{9.0}$. This BR is compared with
each item (i.e. record of patients) and frame TR and BrR as shown in table 2 for a sample of 5 records for$\{I_1, I_2, I_3, \ldots\ldots\ldots\ldots, I_n\}$. Before framing TR and BrR, S-Grid is constructed as explained in section 3.1.3.

The LHS and RHS of second column data in table 2 represent TR and BrR respectively. These rules are merged together to frame a significant rule (SR). SR rule of an item's class label let (I_1) is compared with all item's class label in dataset D (i.e. $\{I_1, I_2, I_3, \ldots\ldots\ldots\ldots, I_n\}$) and generates support values such as $\{SV_1, SV_2, SV_3, \ldots\ldots\ldots\ldots, SV_n\}$. Based on these values calculate R_{count} , Min_{payoff} , Max_{payoff} , Sig_{payoff} using equations 3, 4, 5, and 6.

To improve the efficiency of the algorithm, reduce the number of rules generated by association rules. In general, AR produces more rules, which are not similar to each other. This reduction can be carried out through the conviction measures such as confidence and lift. These values are derived using equation 9 and 10. Based on confidence, a small amount of pruning is done in the generated rule set. This highly

helps us to reduce the time and memory required for decision making by reducing the number of rules.

Decision under uncertainty is a critical task, since only minimum and maximum payoffs are known as the likelihood of each items risk level. Combination of features and risk level is associated with payoff values. Minimum payoff represents the minimum features that assured for a rule to exist to predict the risk level. Similarly, the maximum payoff represents the maximum features that guaranteed for a rule to subsist for the prediction of risk level. Hurwicz criterion is used for decision making, which stipulates a decision making tool's to balance between the minimum and maximum risk levels. This is calculated using the equation (14).

The graph shows the minimum number of features that support the rule to exist in order to predict the decision making. Minimum payoff value and maximum payoff value are the only known values for predicting the risk level of the given item. Heuristic rateis the value based on which the risk levels are justified. Therefore, this rate is more accurately calculated in order to derive the exact detection of breast cancer.

For a graphical representation, five different sample sets are taken. Each sample set has five thousand records taken from the dataset repository of UCI. The value denotes the average value of minimum payoff for various sample sets and in Fig. 2, average value of maximum payoff value for a range of sample set in Fig.3 and Fig.4 denotes heuristic rate of different sample set.

The figures portray that heuristic rates are maximum if the minimum and maximum values of payoff are maximum. HR isdirectly proportional to payoff values. Depending on the HR value the threshold values are set. The threshold valueis the deciding factor for risk level of a particular patient (item). Therefore, prediction of HR values should be accurate to have higher prediction rates.

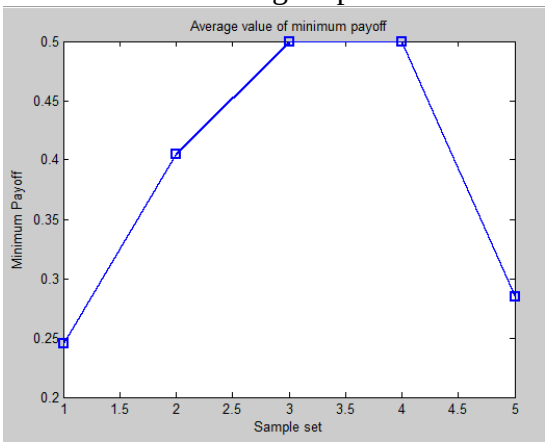

Fig. 2 Average value of Minimum Payoff

Table 2 True and Branch Rule

Item	TR & BrR Rule
I_1	$f_7 = \delta_{7.9} \| f_9 = \delta_{9.0} \Rightarrow f_3 = \delta_{3.1} \| f_4 = \delta_{4.1} \| f_5 = \delta_{5.1} \| f_6 = \delta_{6.9} \| f_8 = \delta_{8.0}$
I_2	$f_5 = \delta_{5.9} \| f_6 = \delta_{6.2} \| f_9 = \delta_{9.0} \Rightarrow f_3 = \delta_{3.1} \| f_4 = \delta_{4.1} \| f_7 = \delta_{7.0} \| f_8 = \delta_{8.1}$
I_3	$f_5 = \delta_{5.9} \| f_9 = \delta_{9.0} \Rightarrow f_3 = \delta_{3.1} \| f_4 = \delta_{4.1} \| f_6 = \delta_{6.3} \| f_7 = \delta_{7.2} \| f_8 = \delta_{8.0}$
I_4	$f_5 = \delta_{5.9} \| f_7 = \delta_{7.9} \| f_9 = \delta_{9.0} \Rightarrow f_3 = \delta_{3.1} \| f_4 = \delta_{4.1} \| f_6 = \delta_{6.4} \| f_8 = \delta_{8.0}$
I_5	$f_5 = \delta_{5.0} \| f_7 = \delta_{7.9} \Rightarrow f_3 = \delta_{3.1} \| f_4 = \delta_{4.1} \| f_6 = \delta_{6.9} \| f_8 = \delta_{8.0} \| f_9 = \delta_{9.9}$

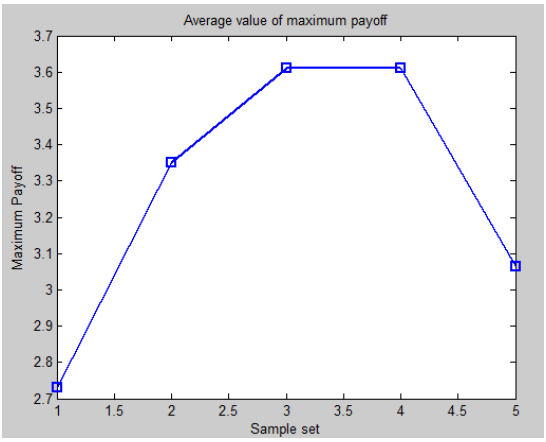

Fig. 3 Average value of maximum payoff

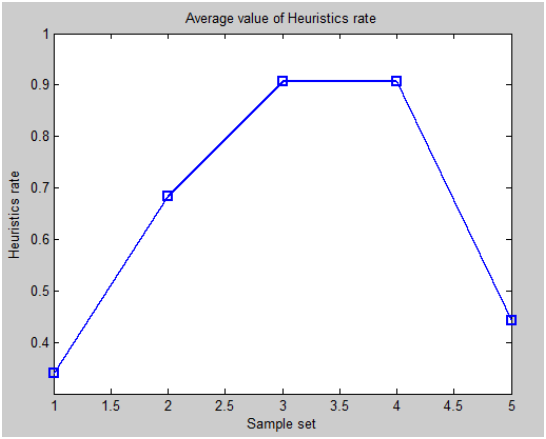

Fig. 4 Heuristic Rate

The efficiency of the DRF algorithm is compared to the detection system in [11], which uses the association rules for rule generation neural networks for classification of breast cancer data set and predicting risk factor of the patient's

records. Fig. 5 shows, the DRF outperform the existing detection system in terms of memory and time required to predict the risk level of given data set D.

The prediction accuracy of the algorithm is tested as 91.5% for a sample set of 5000 records. DRF's prediction accuracy is also higher than the existing system, which has the prediction rate of 97.4% for eight inputs. This experimental study stipulates that the algorithm outperform the existing model for decision making in breast cancer analysis using association rule.

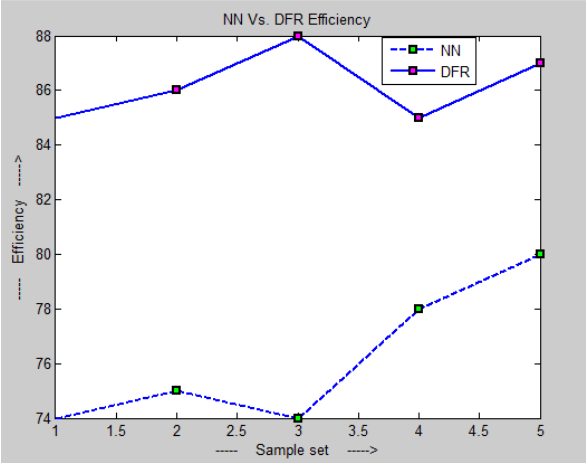

Fig. 5 Efficiency of NN vs. DRF

5. Conclusion and Future Work

In this research, an automatic analyzing algorithm for determining daunting diseases for any given data set is done using association rules named Discriminative Rule Framing. Feature selection is the most imperative part of prediction and pattern recognition. Prediction of risk factor highly depends upon the features extracted. A poor selection of features will implicitly result in worthless prediction. An efficient feature selection also reduces feature vector that contains fruitful information from the original vector. The DRF algorithm provides more attention to feature selection for betterment of risk level prediction in a given data set. DRF uses association rules, the more eminent technique of data mining for dimensionality reduction. With selected and reduced features set, apply DRF algorithm to determine the risk level of an item in a given input data set D. Experimental results in section 4, anticipated that the DRF achieved the highest prediction accuracy of 91.5%for a given sample set of 5,000 records of patients. Meanwhile, DRF is compared with an existent decision making system in [11], which uses the association rule for generation of rules and

neural networks for classification, and analysis explores that DRF outperforms the existing system. Though effective rules are framed through association rules, number of generated rules is enormous. In the future, it is planned to decide to work on reducing the total number of rules generated through association rules.

References

[1] T. Ince, S. Kiranyaz, J. Pulkkinen, and M. Gabbouj, "Evaluation of global and local training techniques over feed-forward neural network architecture spaces for computer-aided medical diagnosis," *Expert Systems with Applications*, vol. 37, no. 12, pp. 8450-8461, 2010.

[2] H. O. Habashy, D. G. Powe, T. M. Abdel-Fatah, J. M. Gee, R. I. Nicholson, A. R. Green, E. A. Rakha, and I. O. Ellis, "A review of the biological and clinical characteristics of luminal-like oestrogen receptor-positive breast cancer," *Histopathology*, vol. 60, no. 6, pp. 854-863, 2012.

[3] H. Chen, J. Zhang, Y. Xu, B. Chen, and K. Zhang, "Performance comparison of artificial neural network and logistic regression model for differentiating lung nodules on CT scans," *Expert Systems with Applications*, vol. 39, no. 13, pp. 11503-11509, 2012.

[4] S. Palaniappan and T. Pushparaj, "A Novel Prediction on Breast Cancer from the Basis of Association rules and Neural Network," *International Journal of Computer Science and Mobile Computing – IJCSMC*, vol.2, no. 4, pp. 269-277, 2013.

[5] M. Karabatak and M. C. Ince, "An expert system for detection of breast cancer based on association rules and neural network," *Expert Systems with Applications*, vol. 36, no. 2, pp. 3465-3469, 2009.

[6] J. Derrac, N. Verbiest, S. García, C. Cornelis, and F. Herrera, "On the use of evolutionary feature selection for improving fuzzy rough set based prototype selection," *Soft Computing*, vol. 17, no. 2, pp. 223-238, 2013.

[7] C. Nguyen, Y. Wang, and H. N. Nguyen, "Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic," *Journal of Biomedical Science and Engineering*, vol. 6, no. 5, pp. 551-560, 2013.

[8] A. Keleş, A. Keleş, and U. Yavuz, "Expert system based on neuro-fuzzy rules for diagnosis breast cancer," *Expert Systems with Applications*, vol. 38, no. 5, pp. 5719-5726, 2011.

[9] H.-L. Chen, B. Yang, G. Wang, S.-J. Wang, J. Liu, and D.-Y. Liu, "Support vector machine based diagnostic system for breast cancer using swarm intelligence," *Journal of medical systems*, vol. 36, no. 4, pp. 2505-2519, 2012.

[10] I. Maglogiannis, E. Zafiropoulos, and I. Anagnostopoulos, "An intelligent system for automated breast cancer diagnosis and prognosis using SVM based classifiers," *Applied Intelligence*, vol. 30, no. 1, pp. 24-36, 2009.

[11] H. Fan and Y. Zhong, "A rough set approach to feature selection based on wasp swarm optimization," *Journal of Computational Information Systems*, vol. 8, no. 3, pp. 1037-1045, 2012.

[12] H. Arafat, S. Barakat, and A. F. Goweda, "Using Intelligent Techniques for Breast Cancer Classification," *International Journal of Emerging Trends & Technology in Computer Science(IJETTCS)*, vol. 1, no. 3, pp. 26-36, 2012.

[13] M. A. Karaolis, J. A. Moutiris, D. Hadjipanayi, and C. S. Pattichis, "Assessment of the risk factors of coronary heart events based on data mining with decision trees," *Information Technology in Biomedicine, IEEE Transactions on*, vol. 14, no. 3, pp. 559-566, 2010.

[14] H.-L. Chen, B. Yang, J. Liu, and D.-Y. Liu, "A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis," *Expert Systems with Applications*, vol. 38, no. 7, pp. 9014-9022, 2011.

[15] C.-S. Son, Y.-N. Kim, H.-S. Kim, H.-S. Park, and M.-S. Kim, "Decision-making model for early diagnosis of congestive heart failure using rough set and decision tree approaches," *Journal of Biomedical Informatics*, vol. 45, no. 5, pp. 999-1008, 2012.

[16] (2001, 16 November 2013). Available: http://weka.sourceforge.net/doc.stable/weka/attributeSelection/ChiSquaredAttributeEval.html

[17] http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)