Life Science Journal

Websites:
http://www.lifesciencesite.com
http://www.sciencepub.net

Emails:
editor@sciencepub.net
sciencepub@gmail.com

MARSLAND PRESS
Multidisciplinary Academic Journal Publisher

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

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

[1]Research Scholar, Anna University, Coimbatore.
[2]Professor, Department of Electrical and Communication Engineering, Bannari Amman Institute of Technology, Erode, India.
[3]Assistant Professor Senior Grade, Department of Electrical and Communication Engineering, Bannari Amman Institute of Technology, Erode, India

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

[B. Gomathy, A. Shanmugam, S.M. Ramesh. **A Proficient System for Automatic Detection of Risk Level in Disease Detection using Association Rule based DRF Algorithm**. *Life Sci J* 2021;18(9):19-27] ISSN 1097-8135 (print); ISSN 2372-613X (online) http://www.lifesciencesite.com. 4. doi:10.7537/marslsj180921.04.

Index terms - Association Rule Mining, 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 we use a profound technique named data mining. 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. 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 patterns finding 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. We initially perform preprocessing process using normalization technique for 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. Our proposed DRF algorithm has two stages. At stage 1, class-labels are framed from preprocessed data set, based on which base rule for our DRF algorithm is generated. As by this approach, rule formation is based on user need, and it can be adaptable to any kind of medical data set. S-grid takes the base rule and frame heuristics 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 our 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, threshold is set to

filter the items in a data set according to the requirement of risk analysis. Fig. 1 shows the flow of our 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 record. 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 are undertaken to generate the report of 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 vulnerabilities.

In our 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 given data set.

Rest of this paper is structured as follows: Section 2 provides reviews related to our work. Section 3 presents our 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, we presented some works of various authors that were related to our proposal. Analyzing and determining risk factors in medical data set was a mind-numbing work, which consumes time. To overcome this problem, Baxt in [1] used artificial neural network to classify the 356 record of patients with acute coronary occlusion. This detection system used back propagation of neural network to train the data set. This model accurately determined 80% of patients with infraction. [2] followed his work, and used neural networks to predict 1008 clinical data set of breast cancer patients. Review of methodologies for neural networks and logistics regression was presented in [5], and results showed the performance of neural network was not significant as logical regression. Logical regression outperforms neural network, due to its property of interpretability of model parameter and easy to use.

Due to the associated memory characteristics and generalization capacity of neural network, it plays a good role in prediction and classification. However, with massive data sets, its application in handling classification is limited. To deal with classification of large data sets, the neural network was integrated with the multivariate adaptive regression splines (MARS) approach in [6]. Results in [6] demonstrated, MARS with neural network outperformed the results of discriminant analysis, artificial neural networks and multivariate adaptive regression splines separately. Furthermore, authors of [11] 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 [11] proved that decision system with neural network with association rule achieved 95.6% of accuracy in classification.

Fuzzy based method was introduced in [4] to classify medical data with consideration in classification and prediction time. Fuzzy entropy was used for feature selection and partitioned the pattern space into non-overlapping decision regions. Partitioned regions were classified through fuzzy classifiers. Examination of fuzzy filter in [4] reveled that it performed well for pattern classification. It was experimented with Iris database and Wisconsin breast cancer database. Better results were obtained while using a hybrid method based on fuzzy-artificial immune system with k-nearest neighbour algorithm. Hybrid method was evaluated in [8] using Wisconsin Breast Cancer Dataset (WBCD) and the results showed that this method has accurate detection that other methods proposed before the hybrid method. Ex-DBC (Expert system for Diagnosis of Breast Cancer) a diagnostic tool was proposed in [13] for diagnosing breast cancer. It used neuro-fuzzy method and diagnosed with 96% and 81% positive and negative predictive accuracy level respectively. A comparison among Multilayer perceptron neural network (MLPNN), combined neural network (CNN), probabilistic neural network (PNN), recurrent neural network (RNN) and support vector machine (SVM) were made in [9] to find classification accuracies on Wisconsin breast cancer database. Results demonstrated that SVM achieved diagnostic accuracies, which were higher than that of the other systems taken for comparison.

An Automated method for classification of medical data set through the help of quantitative measure 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. A new mathematical approach to uncertainty, imprecision and vagueness was rough set [3]. Its uniqueness made it flexible to real-time applications such as discovery of data dependencies and patterns of data, evaluates the importance of attributes, redundant attribute reduction and extraction of rules from data sets. In [7] rough set data reduction was used to find class labels for classification. Rough set approach based rules were evaluated and compared with the ID3 classifier algorithms. Results showed that classification accuracy of Rough set is much better.

In [12] 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 exact classification for the PCI, and 75% of perfect classification for the CABG events. The profound works of [3, 7, 10, and 12] paved a way for designed a new methods such as in [14, and 15]. Support vector classifier was integrated with rough set and framed RS_SVM [14] model for diagnose 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. 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 large data set. AR is widely applicable in market based analysis to procure frequent item sets. Profitably it can be suitable to 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 we have maximum chance for occurrence of Y. 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, we applied a well known technique named normalization for preprocessing the data set. Preprocessed data set is given as input to our proposed 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 $F = \{f_1, f_2, f_3, \ldots\ldots\ldots, f_n\}$ 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 huge medical data set, which helps in predicting the outcome. Feature section also improves the performance of prediction of 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 our approach we use Wega tool's ChiSquaredAttributeEVal method for 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.

3.1.2 Framing Base Rule:

Base rules (BR) are framed from the class labels that we have selected from above step. The base rule can be represented as $BR(f_1 = \delta_{11} \,||\, f_2 = \delta_{25} \,||\, f_3 = \delta_{32} || \ldots\ldots\ldots || \, f_j = \delta_{j9})$. Each feature has various class labels. Depending on the analysis, criteria for feature differs. Class labels that we have chosen will act as the criteria for prediction of risk level. Therefore, forming base rule entirely depends on the decision of 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 symbolized 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 item I_2's feature with the BR's f_3 class label value δ_{3x}. If 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.

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

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

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

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

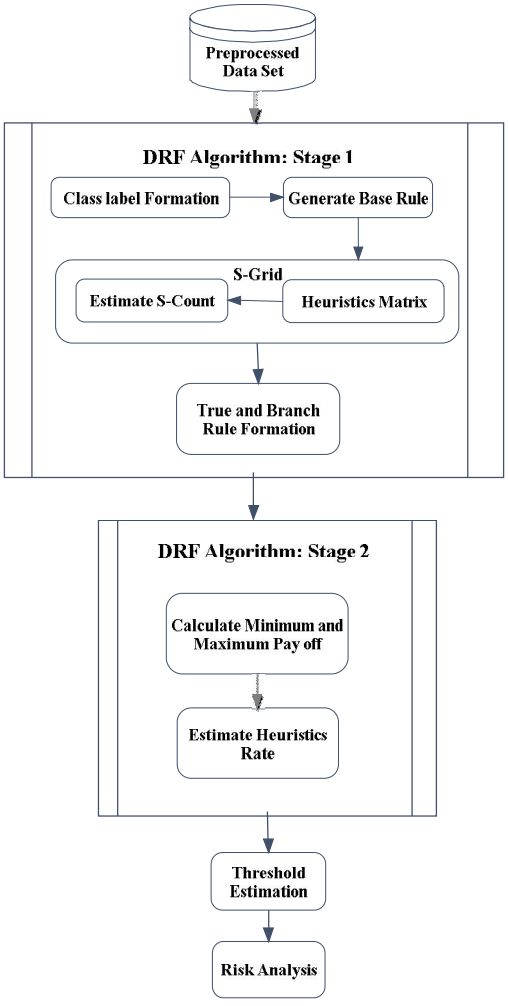

Fig. 1 DRF Flow Diagram

3.1.4 True and Branch Rule formation:

True rule (TR_i) and branch rule (BrR_i) are the most sophisticated rules formed from HM matrix by our DRF algorithm a 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_i has the features that have value equal to one in HM matrix for a item I_i. All other features will 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 $f_1 = \delta_{11} || f_4 = \delta_{49}$. BrR_1 contain 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_1 class label values are not same.

3.2 DRF: Stage 2

3.2.1 Calculate Maximum and Minimum pay off:

Minimum and maximum pay off values are the key values which are used for decision making. To acquire these values, we formulate a new rule named significant rule (SR_i). SR_i is framed by combining the TR_i and BrR_i. From the previous example SR_1 can be framed as $f_1 = \delta_{1.1} || = \delta_{2.0} || f_3 = \delta_{3.5} || f_4 = \delta_{4.9} || f_5 = \delta_{5.1} || f_6 = \delta_{6.2} || f_7 = \delta_{7.6}$. Class label of SR_i's feature is compared with feature's of all item set's class label 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 $\{I_1, I_2, I_3, \ldots\ldots\ldots, I_n\}$ 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 $\{I_1, I_2, I_3, \ldots\ldots\ldots, I_n\}$ in D is calculated as $\{SV_1, SV_2, SV_3, \ldots\ldots\ldots, SV_n\}$.

For better and efficient risk analysis, we check $\{SV_1, SV_2, SV_3, \ldots\ldots\ldots, SV_n\}$ 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. This support value calculation is carried for all $\{SR_1, SR_2, \ldots\ldots\ldots SR_n\}$ and compared with F_{count}. A R_{count_i} value is calculated using equation 3 for $\{SR_1, SR_2, \ldots\ldots\ldots SR_n\}$.

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

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 continuous addition of support values of SR_i, for each item $\{I_1, I_2, I_3, \ldots\ldots\ldots, I_n\}$ 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, we framed two conviction measures confidence and lift. For enumerating these conviction measures, support value for true rule SV_{Ti} and branch rules SV_{BrRi} are calculated independently. SV_{Ti} and SV_{BrRi} are the number of features in 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. $\{TR_1, TR_2, TR_3, \ldots\ldots\ldots, TR_n\}$ and $\{BrR_1, BrR_2, BrR_3, \ldots\ldots\ldots, BrR_n\}$.

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

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

CM_1 and CM_2 values 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 item are calculated and arranges the rules in ascending order according to HR value of each item. Find 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 predict risk factor as moderate, low and higher risks respectively.

4. Analysis Of Drf Algorithm

To examine our projected DRF algorithm, data set of breast cancer is extracted from UCI learning

repository [16]. 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 is given in Table 1.

Before employing out DRF algorithm for the data set D, we apply normalization based preprocessing method. Preprocessing is the most needed step in the decision making process. Since, a data set may hold missing values, noisy data, etc. In order to remove such data, we apply the preprocessing. Feature selection is the most important step in the prediction algorithm because the accuracy of decision making depends upon the features selected. A strong prediction system is generated only through best feature selection.

Table 1. Features & class lables 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

Using the features and class labels, of table 1, we employ a subset of features and class labels to frame BR, 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 we calculate $R_{count}, Min_{payoff}, Max_{payoff}, Sig_{payoff}$ using equations 3, 4, 5, and 6.

To improve the efficiency of our proposed algorithm, we 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, we did a small amount of pruning in 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. We used Hurwicz criterion 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 exits 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 rate is 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 graphical representation, we have taken five different sample set. Each sample set have five thousand records taken from the data set repository of UCI. The value denotes the average value of minimum payoff for various sample set 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.

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

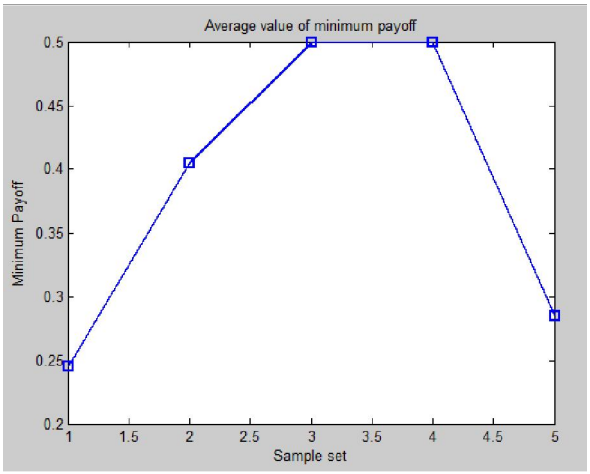

Fig. 2 Average value of Minimum Payoff

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	$f_5 = \delta_{5.0} \| \| f_7 = \delta_{7.9} \Rightarrow f_3 = \delta_{3.1} \| \| f_4 = \delta_{4.1} \| \| f_6 = \delta_{6.9} \| \| f_8 = \delta_{8.0} \| \| f_9 = \delta_{9.9}$

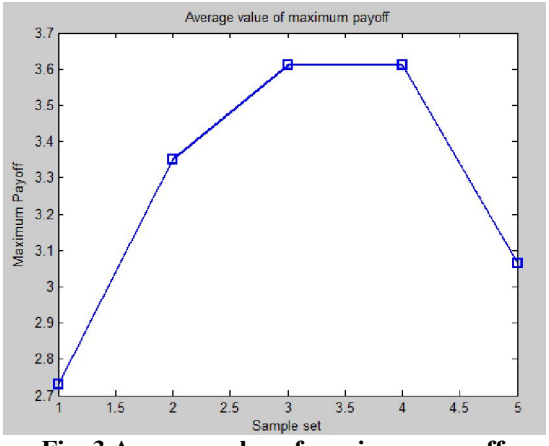

Fig. 3 Average value of maximum payoff

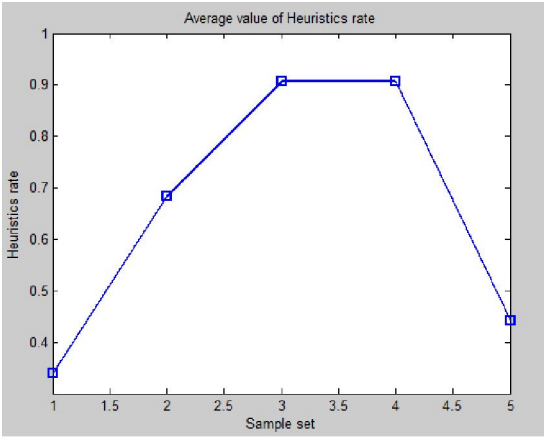

Fig. 4 Heuristic Rate

Life Science Journal

Websites:
http://www.lifesciencesite.com
http://www.sciencepub.net

Emails:
editor@sciencepub.net
sciencepub@gmail.com

MARSLAND PRESS
Multidisciplinary Academic Journal Publisher

Efficiency of our proposed 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 patient's records. Fig. 5 shows, our proposed work outperforms 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 our 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 our algorithm outperform the existing model for decision making in breast cancer analysis using association rule.

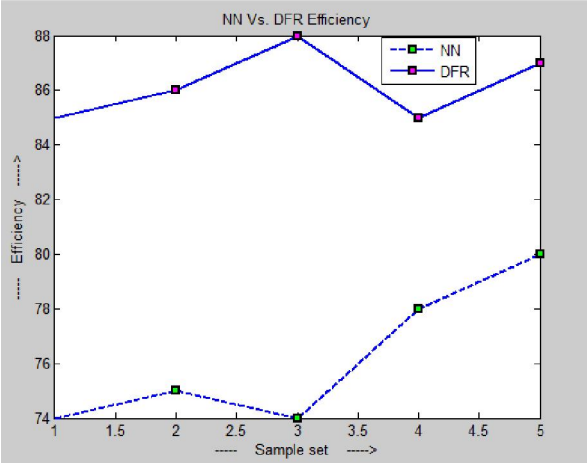

Fig. 5 Efficiency of NN vs. DRF

5. Conclusion And Future Work

In this proposal, an automatic analyzing algorithm for determining daunting diseases for any given data set is proposed using association rules named Discriminative Rule Framing (DRF). 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. Our proposed 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, we 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 proposed DRF achieved the highest prediction accuracy of 91.5% for a given sample set of 5,000 records of patients. Meanwhile, we compared DRF 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 future, we planned to decide on work on reducing the total number of rules generated through association rules.

References

[1] William Baxt, G., "Use of an Artificial Neural Networks for Data Analysis in Clinical Decision-Making: The Diagnosis of Acute Coronary Occulusion," in the Journal of Neural Computation, vol. 2, no. 4, pp. 480-489, 1990

[2] Peter M. Ravdin, Gary M. Clark, Susan G. Hilsenbeck, Marilyn A. Owens, Patricia Vendely, M. R. Pandianand William L. McGuire, "A demonstration that breast cancer recurrence can be predicted by Neural Network analysis," *in the journal of BREAST CANCER RESEARCH AND TREATMENT*, vol. 21, no. 1 pp. 47-53, 1992, DOI: 10.1007/BF01811963

[3] Pawlak, Z, "Rough set approach to knowledge-based decision support," European Journal of Operational Research, 99(1), 48–57, 1997.

[4] Hahn-Ming Lee, Chih-Ming Chen, Jyh-Ming Chen, and Yu-Lu Jou, "An efficient fuzzy classifier with feature selection based on fuzzy entropy," in the journal of IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, vol. 31, issue 3, pp. 426-432, Jun 2001.

[5] Stephan Dreiseitl, and Lucila Ohno-Machado, "Logistic regression and artificial neural network classification models: a methodology review," in the Journal of Biomedical informatics, vol. 35, issue 5-6, pp. 352 -359, Oct 2002.

[6] Shieu-Ming Chou, Tian-Shyug Lee, Yuehjen E. Shao, and I-Fei Chen, "Mining the breast cancer pattern using artificial neural networks and multivariate adaptive regression splines," in the journal of Expert System with Applications, vol. 27, issue 1, pp. 133-142, Jul 2004.

[7] Aboul-Ella Hassanien, "Rough set approach for attribute reduction and rule generation: a case of patients with suspected breast cancer," in the journal of the American Society for Information Science and Technology, vol. 55, issue 11, pp. 954-962, Sep 2004

[8] Seral Sahan, Kemal Polat, Halife Kodaz, and Salih Gunes, "A new hybrid method based on fuzzy-artificial immune system and k-nn algorithm for breast cancer diagnosis," in the journal of Computers in Biology abd Medicine, vol. 37, issue 3, pp. 415-423, Mar 2007

[9] Elif Derys Ubeyli, "Implementing automated diagnostics system for breast cancer detection," vol. 33, issue 4, pp. 1054-1062, Nov 2007

[10] Ilias Maglogiannis, Elias Zafiropoulos, and Ioannis Anaganostopoulos, "An intelligent system for automated breast cancer diagnosis and prognosis using SVM based classifiers," in the journal of Applied Intelligence, vol. 30, issue 1, pp. 24-36, Feb 2009

[11] Murat Karabatak, and Cevdet Ince, M., "An expet system for detection of breast cancer based on association rules and neural network," in the journal of Expert Systems with Applications, vol. 36, issue 2, part 2, pp. 3465-3469, Mar 2009.

[12] Minas A. Karaolis, Joseph A. Moutiris, and Constantinos S. Pattichis, "Assessment of the Risk Factors of Coronary Heart Events Based on Data Mining With Decision Trees," in the journal of IEEE Transactions on Information Technology in Biomedicine, vol. 14, no. 3, pp. 559-566, MAY 2010

[13] Ali Keles, Ayturk Keles, and Ug˘ur Yavuz, "Expert system based on neuro-fuzzy rules for diagnosis breast cancer," in the journal of Expert System with Applications, vol. 38, issue 5, pp. 5719-5726, May 2011.

[14] Hui-Ling Chen, Bo Yang, Jie Liu, and Da-You Liu, "A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis," in the journal of Expert System with Applications, vol. 38, issue 7, pp. 9014-9022, Jul 2011

[15] Chang-Sik Son, Yoon-Nyun Kim, Hyung-Seop Kim, Hyoung-Seob Park, and Min-Soo Kim, "Decision-making model for early diagnosis of congestive heart failure using rough set and decision tree approaches," in the journal of Biomedical Informatics, Apr. 2012 (Accepted but not yet published).

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

5/1/2021