International Journal of Medical, Medicine and Health Sciences
ISSN: 2517-9969
Vol:7, No:2, 2013

Dynamic Features Selection for Heart Disease Classification

Walid MOUDANI

***Abstract*—**The healthcare environment is generally perceived as being information rich yet knowledge poor. However, there is a lack of effective analysis tools to discover hidden relationships and trends in data. In fact, valuable knowledge can be discovered from application of data mining techniques in healthcare system. In this study, a proficient methodology for the extraction of significant patterns from the Coronary Heart Disease warehouses for heart attack prediction, which unfortunately continues to be a leading cause of mortality in the whole world, has been presented. For this purpose, we propose to enumerate dynamically the optimal subsets of the reduced features of high interest by using rough sets technique associated to dynamic programming. Therefore, we propose to validate the classification using Random Forest (RF) decision tree to identify the risky heart disease cases. This work is based on a large amount of data collected from several clinical institutions based on the medical profile of patient. Moreover, the experts' knowledge in this field has been taken into consideration in order to define the disease, its risk factors, and to establish significant knowledge relationships among the medical factors. A computer-aided system is developed for this purpose based on a population of 525 adults. The performance of the proposed model is analyzed and evaluated based on set of benchmark techniques applied in this classification problem.

***Keywords*—**Multi-Classifier Decisions Tree, Features Reduction, Dynamic Programming, Rough Sets.

I. Introduction

MEDICAL diagnosis is an important but complicated task that should be performed accurately and efficiently and its automation would be very useful and advantageous. Unfortunately, all doctors do not possess expertise in every sub specialty and moreover they are in many places a scarce resource. However, appropriate computer-based information and/or decision support systems can aid in enhancing medical care and in achieving clinical tests at a reduced cost. Moreover, efficient and accurate implementation of automated system needs a comparative study of various available techniques. Indeed, most hospitals today employ some kind of hospital information systems to manage their healthcare or patient data. These systems typically generate huge amounts of data which, unfortunately, are rarely used to support clinical decision making. There is a wealth of hidden information in these data that is largely untapped. The main motivation of this research is to process data in order to get useful information that can enable healthcare practitioners to make intelligent clinical decisions. In healthcare domain, data mining has been used intensively and extensively by many organizations. This leads to improve decision-making by discovering patterns and trends in large amounts of complex data. Recently, the data mining techniques were utilized by several authors to present diagnosis approaches for diverse types of heart diseases [3]. However, such analysis has become increasingly essential as financial pressures have heightened the need for healthcare organizations to make decisions based on the analysis of clinical and financial data. Insights gained from data mining can influence cost, revenue, and operating efficiency while maintaining a high level of care. In this paper, we deal with the diagnosis of one of the real health problem, called Coronary Heart Disease (CHD), because of its increasing frequency over the countries. CHD is assessed as the primary cause of mortality among adults in the world, and it is the top concern of healthcare organizations and medical doctors. The diagnosis of this disease is a vital and intricate job in medicine. The recognition of CHD from diverse features or signs is a multi-layered problem that is not free from false assumptions and is frequently accompanied by impulsive effects. Thus, the attempt to exploit knowledge and experience of specialists and clinical screening data of patients composed in databases to assist the diagnosis procedure is regarded as a valuable option. In this study, a reduction features algorithm based on dynamic rough sets technique is proposed. We propose an intelligent multi-classifier decision trees using RF method for efficient data classification and prediction. The paper is organized as follows. In section 2 and 3, we present the considered disease and we discuss the works found in the literature related to this disease and also the computational techniques applied for solving this task. In section 4, the hybrid strategy of proposed model in the data mining setting is presented. It describes the proposed solution approach based on dynamic rough sets in order to prepare the processed CHD database which fed into the multi-classifier decision trees using RF. In Section 6, we describe the proposed solution approach through a numerical example applied to a real medium size of CHD information followed by discussion and analysis of the results obtained. Finally, we ended by a conclusion concerning this new approach and the related new ideas to be tackled in the future.

Walid MOUDANI: Head of Business Computer Department, Lebanese University, Lebanon (wmoudani@ul.edu.lb).

II. Description of Coronary Heart Disease

Despite immense medical progress in the last 3 decades, heart disease continues to be a major health problem in both industrialized and developing nations. Recognizing the symptoms of a heart attack and seeking immediate medical attention may mean the difference between life and death. CHD is caused by sudden loss of blood and oxygen to the heart [9]. The symptoms related to this disease are illustrated such as: (1) CHD is most common condition that predisposes a person to heart attack – (2) Reducing the amount of blood in

the artery. The plaque and resulting blood clots block the artery partially or completely, leading to reduce the amount of blood that can flow through the artery to the heart - (3) Cutting off the oxygen supplied to part of the heart muscle – (4) If the blood supply is cut off long enough, that part of the heart muscle dies. This is a heart attack – (5) If a large enough part of the heart muscle is affected, a dangerous rhythm disorder called ventricular fibrillation may occur. If this happens, the heart may stop. This is called cardiac arrest, and most people who have cardiac arrest die.

III. Literature Overview

Medical diagnosis of CHD is an important but complicated task that should be performed accurately and efficiently and its automation would be very useful. All doctors are unfortunately not equally skilled in every sub specialty and they are in many places a scarce resource. An automated medical diagnosis system would enhance medical care and reduce costs. Indeed, predicting Heart Attack is not limited on clinical factors but may be also based on several machine learning techniques like Artificial Neural Networks (ANN), Decision Trees (DT), and Rough Sets Theory (RST) are used for data classification. These tools have recently fascinated many researchers since they are practical and robust for many real-world problems, and are rapidly developed nowadays. A tool based on hybrid techniques (RST and ANN), has been developed in order to construct an algorithm called "Intelligent Rough Neural Network System" [2]. RST is a tool for reducing quantized data sets by discarding attributes that have no or limited discriminatory power, its basic notions are: Information system, approximation, reduction of attributes, and accuracy. The multi-layer perceptron with back-propagation is used to minimize the error between the desired and computed unit values. This combination resulted in a hybrid algorithm that uses RST for pre-processing of data and ANN for classification or prediction. The experimental results show that the proposed hybrid architecture is very efficient for medical data analysis in significantly lesser processing time. In [6], they proposed to find useful information from heart disease data set in India. Hence, the proposed Decision Support System (DSS) helps to identify a risk score for predicting the Heart risk of a Patient in the subsequent Years. The proposed system has 19 features which have been reduced to most important features using Genetic algorithms. The system is anticipated that data mining could help in the identification of risk subgroups of subjects for developing future events and it might be a decisive factor for the selection of therapy, i.e., angioplasty or surgery. In [7], they have developed a study for diagnosing heart diseases by analyzing some attributes using data mining and ANN. Firstly, the data is extracted, loaded in a data warehouse, cleansed, and changed into data sets with appropriate characteristics; then, the pre-processed data is clustered using the K-mean algorithm with k = 2 (2 clusters) one cluster for data relevant to heart disease and the other for the remaining data. After the clustering, the frequent patterns are mined using the Maximal Frequent Itemsets (MAFIA) algorithm that extracts the association rules from the clustered dataset and prepares the data to calculate the significance weight-age of each pattern. After all this process, the Multi-Layer Perceptron Neural Network is used to map the sets of input onto a set of appropriate output. The designed prediction system helped to create a model to the heart attack prediction with great efficacy.

IV. Presentation of Random Forest

Multi-classifiers are the result of combining several individual classifiers. When individual classifiers are combined appropriately, we usually obtain a better performance in terms of classification precision and/or speed to find a better solution. Different multi-classifiers have different characteristics. In [8], they divided the methods for building multi-classifiers in two groups: ensemble and hybrid methods. Nowadays, numerous attempts to construct ensembles of classifiers towards increasing the performance have been introduced [4]. Examples of such techniques are Adaboost, Bagging and RFs. In [1], Brieman defined RF as a multi-classifier composed by decision trees where every tree h_t had been generated from the set of data training and a vector θ_t of random numbers identically distributed and independent from the vectors vectors θ_1, θ_2,.., θ_{t-1} used to generate the classifiers h_1, h_2, .., h_{t-1}. Each decision tree is built from a random subset of the training dataset. It used a random vector that is generated from some fixed probability distribution, where the probability distribution is varied to focus examples that are hard to classify. A random vector can be incorporated into the tree-growing process in many ways. The leaf nodes of each tree are labelled by estimates of the posterior distribution over the data class labels. Each internal node contains a test that best splits the space of data to be classified. A new, unseen instance is classified by sending it down every tree and aggregating the reached leaf distributions. There are three approaches for RFs such as: Forest-RI (Random Input selection) and Forest-RC (Random combination), mixed of Forest-RI and Forest-RC. Forest-RI consists of randomly selecting F input features to split at each node of the decision tree. As a result, instead of examining all the available features, the decision to split a node is determined from these selected F features. The tree is then grown to its entirety without any pruning. This may help reduce the bias present in the resulting tree. Once the trees have been constructed, the predictions are combined using a majority voting scheme. The strength and correlation of RFs may depend on the size of F. If F is sufficiently small, then the trees tend to become less correlated. On the other hand, the strength of the tree classifier tends to improve with a larger number of features, F. As a trade-off, the number of features is commonly chosen to be $F = \log_2 d + 1$, where d is the number of input features. Since only a subset of the features needs to be examined at each node, this approach helps to reduce significantly the runtime. Forest-RC is used to create a combination of the input features. In case the number of

original features d is too small, then it is difficult to choose an independent set of random features for building the decision trees. One way to increase the features space is by creating linear combination of the input features. Specifically, at each node, a new feature is generated by randomly selecting L of the input features. The input features are linearly combined using coefficients generated from a uniform distribution in the range of [-1, +1]. At each node, F of such randomly combined new features is generated, and the best of them is subsequently selected to split the node.

A third approach for generating the random trees is to randomly select one of the F best splits at each node of the decision tree. This approach may potentially generate trees that are more correlated than Forest-RI and Forest–RC, unless F is sufficiently large. It also does not have the runtime savings of Forest-RI and Forest–RC because the algorithm must examine all the splitting features at each node of the decision tree.

The use of RFs technique has some desirable characteristics shown such as: it is easy to use; it does not require models, or parameters to select except for the number of predictors to choose at random at each node; it is unexcelled in accuracy among current algorithms, it runs efficiently on large databases, it is relatively robust to outliers and noise; it is simple and easily parallelized; it is faster than bagging or boosting; it can handle thousands of input variables without variable deletion; it gives estimates of what variables are important in the classification; it generates an internal unbiased estimate of the generalization error as the forest building progresses, it has an effective method for estimating missing data and maintains accuracy when a large proportion of the data are missing, it has methods for balancing error in class population unbalanced data sets, and it computes proximities between pairs of cases that can be used in clustering. However, the generalization error of RFs classifiers depends on the strength of the individual trees in the forest and the correlation between them. It has theoretically proven that the upper bound for generalization error of RFs converges to the following expression, when the number of trees is sufficiently large.

$$Generalization\ error \le \frac{\bar{\rho}\left(1-s^2\right)}{s^2} \quad (1)$$

where $\bar{\rho}$ is the average correlation among the trees and s is a quantity that measures the strength of the tree classifiers. The strength of a set of classifiers refers to the average performance of the classifiers, where performance is measured probabilistically in terms of the classifier's margin:

$$margin,\ M(X,Y) = P\left(\hat{Y}_\theta = Y\right) - \max_{Z \neq Y} P\left(\hat{Y}_\theta = Z\right) \quad (2)$$

where $\hat{Y}_\theta$ is the predicted class of X according to a classifier built from some random vector θ. The higher the margin is, the more likely it is that the classifier correctly predicts a given example X. As the trees become more correlated or the strength of the ensemble decreases, the generalization error bound tends to increase.

V. Methodology and Solution Approach

We present an embedded hybrid intelligent classification solution approach based on dynamic reduced subsets of features. This approach is validated by using multi-classifier RF classification technique to identify the CHD attack cases.

A. Description of the Methodology

The strategy reported can be described as a KDD (Knowledge discovery in databases) experiment. Following a typical KDD framework, where Data Mining is the core in the overall process, the experiment went through all steps of Fig. 1, starting from the stage of gaining profound knowledge of the domain till the actual use of discovered knowledge. We start by understanding the objectives and requirements from a business perspective, converting this knowledge into a data mining problem definition, and designing a preliminary plan to achieve the objectives. Therefore, we use the raw the data and proceed to understand it, identify its quality, gain preliminary insights, detect interesting subsets to form hypotheses for hidden information, and finally to construct the final dataset that will be fed into the modeling tools. A description of database, source of data, pre-processing steps (cleaning, transformation, and integration) is given here.

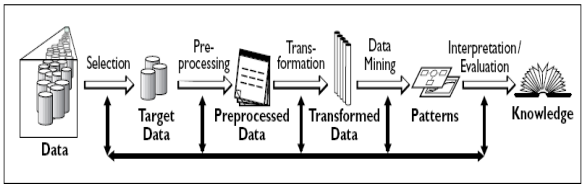

Fig. 1 Methodology roadmap of the KDD process

B. Data Source, Description, and Processing

The aim is to build an intelligent tool for extracting the significant patterns from the CHD data warehouse. The collected data suffers from missing data, inconsistent data, duplicate data, cleaning and filtering of the data, which lead to carry out a processing phase with respect to the data and data mining algorithm employed so as to avoid the creation of deceptive or inappropriate rules or patterns. In this study, the CHD data warehouse which contains the screening clinical data of cardiac patients, is refined by removing duplicate records, normalizing and transforming the values used to represent information in the database, accounting for missing data points, supplying missing values, and removing unneeded data fields. The raw data is changed into data sets with a few appropriate characteristics. Moreover it might be essential to combine the data so as to reduce the number of datasets besides minimizing the memory and processing resources required by the data mining algorithm (Fig. 2). A total of 525 patients with 19 medical factors were gathered from many clinical data sources. The gathered data covers a set of relevant and significant features collected, defined,

preprocessed, and validated based on the knowledge of physician experts. The retained significant patterns are defined by experts in order to have an efficient mining process and predict the appropriate risk level (Table I). The output is a web-based calculation tool that assesses the historical risk of

CHD. For the sake of consistency, categorical and non-categorical attributes were used in this model. The decision attribute was identified as the predictable attribute with a wide range of values that states 5 level of CHD risk (No Risk – Low Risk – Moderate Risk – High Risk – Severe Risk). However, the most classical approaches had dealt with only two values indicating if the case is risky or not. The results provided by the system (decision attribute), called Diagnosis, was identified as the predictable attribute with 5 states that reflect the level of risk of coronary heart risk. They are retained based on experts' knowledge.

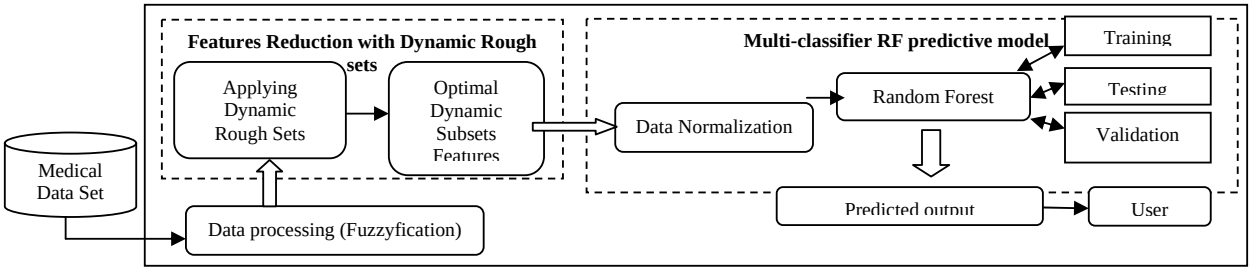

Fig. 2 The model of the proposed solution approach

C. Dynamic Rough Sets Attribute Reduction

We describe the proposed solution approach applied in order to reduce optimally the initial set of features by saving the most significant and relevant ones. This process has been employed in order to remove redundant conditional features from discrete-valued datasets, while retaining their information content. The reduction of the features consists to enumerate dynamically the optimal subsets of the reduced attributes of high interest by reducing the degree of complexity. The criterion for reducing these features is based on the dependency factor among the attributes. By calculating the change in dependency when an attribute is removed from the set of considered conditional attributes, a measure of the significance of the attribute can be obtained. The higher the change in dependency, the more significant the attribute is. However, the reduction of attributes is achieved by comparing equivalence relations generated by sets of attributes. The results help to generate a suitable and non-complex classification related to the considered problem where the constraints are involved in verifying the validity of the developed solution. In fact, the choice of the criterion aims to maximize the dependency degree of the solution while meeting all the constraints level. Indeed, scaling constraints level lead to address every step of the optimization process exponentially growing number of states within the parameters sizing the problem, making it impossible to process numerically the problem of consequent dimensions. The proposed method, called Dynamic Rough Sets Attribute Reduction (DRSAR), shows competitive performance compared with some other computational intelligence tools since it produces optimistic reduced attribute subsets [5].

D. Predicting CHD Cases using RF

The main concept of the proposed approach is the build optimally, step by step; the dynamic features subsets for constructing the effective RF decision tree. The use of DP permits to evaluate the accuracy of the proposed model reached at a given stage with defined features subsets. Indeed, the criterion to be optimized is related to the accuracy deduced from the considered dynamic subsets of features. The accuracy of the prediction can be improved gradually as the size of these subsets may develop. At the end of this process, the highest accuracy associated to an optimal reduced subset(s) is retained as solution (Table II). The choice of RF has been validated while comparing it to the other type of decision trees such as J48 and ID3. The results show an enhanced solution in term of accuracy of classification and error rate (Table III).

TABLE I
CORONARY HEART ATTACK FACTORS

Attribute	Type	Description
AG	Numeric	Age
SE	Categorical	Sex
HTA	Boolean	Blood pressure
CH	Boolean	Cholesterol
BS	Boolean	Blood Sugar
Tabac	Boolean	Smoking
Obesity	Boolean	Overweight
HE	Boolean	Hereditary
AC	Boolean	Alcohol
HR	Numeric	Heart Rate
RF	Numeric	Renal Failure
BPCO	Numeric	Lung disease
PIDM	Numeric	PastIDM
AP	Boolean	Angioplasty
CABG	Boolean	
AF	Boolean	Atrial fibrillation
INA	Boolean	Inactivity
BC	Categorical	Bad Cholesterol
CVA	Boolean	Cerebral vascular accident
CHDRL	Categorical	Diagnosis of CHD Risk Level

TABLE II
RESULTS OF FEATURES REDUCTION ALGORITHMS

Classification Methods	Reduced # of attributes (Initial, Reduced)
SimRSAR	(19 → 16)
GenRSAR	(19 → 12)
AntRSAR	(19 → 14)
DRSAR	(19 → 9)

TABLE III
RESULTS OF CLASSIFICATION TECHNIQUES

Classification Methods	Incorrectly Classified Instances	Error rate	Accuracy
J48	58	0.110476	0.889524
ID3	148	0.281905	0.718095
Forest-RC	12	0.022857	0.977143
Forest-RI	13	0.024762	0.975238

VI. RESULTS ANALYSIS AND DISCUSSION

The results of the experimental analysis in finding significant patterns for CHD prediction are presented. With the help of the dataset, the patterns significant to the CHD prediction are extracted using the approach discussed above. The CHD data set is preprocessed successfully by removing duplicate records and supplying missing values. The refined CHD data set, resultant from preprocessing, is then reduced using dynamic programming to define the most relevant features to CHD. Therefore, multi-classifier RF decision tree method is applied to predict the level of coronary heart risk. The results of the experimental analysis in finding significant patterns for CHD prediction, presented in this paper, are analyzed in order to compare the relative performance followed by an interpretation, validation and discussion. The proposed system for automated medical diagnosis has been performed accurately and efficiently. His performance has been validated by physician experts. It shows a great potential of the proposed in predicting the CHD cases. Moreover, the results are interpreted leading to some features stated as below:

- The "Young" category of the "age" attribute is not present at all in those heart attack risk results; which means that the persons who are under 30 years old are not risky and do not have the chance to get a heart attack.
- The first elaborated classification rule is highly important "if (Patient has Cholesterol)=> classified as 'YES' for MI => risky" because it shades lights on the top cause of heart attacks and treats "Cholesterol" as the main reason behind heart failure.
- It is turn of "Heart Rate" attribute which affects badly the heart work in case of a "high or low" level.
- The presence of "Diabetes", "Hereditary" and "HTA" factors increases also the risk of getting a heart attack.
- In addition, a patient who has got a myocardial infarction in the past "PastMI" is more risky to get another heart attack in the future.
- The "Alcohol", "Tabaco" and "Age" affect so badly the risk of getting a heart attack and their influences differ from patient to another with the combination with other risk factors.
- We observe that only one factor ("Angioplasty") plays a positive role if it exists.

The performance of the proposed prediction system is based on some parameters such as: Attributes reduction, Misclassification rate, and Accuracy. The results issued by applying several decision trees techniques are summarized in Tables II and III. The variation of misclassification rate between the different techniques shows that the rate of the incorrectly classified instance (0.0007%) is the lowest by using Forest-RC comparing to the other set of techniques presented. ID3 decision tree produces the highest rate of misclassification (0.1543). Therefore, it is shown that the accuracy rate of Forest-RI is the best among the different techniques. The classification results using RFs are obtained from ten-fold cross-validation. However, we conclude that, by using RFs multi-classifier technique, the initial number of attributes is reduced from 19 to 9. The relevant features are only taken into consideration which leads to enhance the complexity of the proposed model by focusing the study based on reduced features. This number is great when using ID3 and J48 techniques.

After analyzing the values of different parameter, the performance of the classification process generates a highly precision while using RF multi-classifier decision trees, especially, when using the variant Forest-RC which provides the highest accuracy rate.

VII. CONCLUSIONS

In this study, we have presented an embedded intelligent and effective CHD prediction system using data mining and RF techniques. Moreover, we have provided an efficient dynamic approach for the extraction of significant patterns from the CHD data warehouses for the efficient prediction. This leads to reduce the features by using the rough sets associated to the DP technique in order to obtain the equivalence most relevant subsets of features. These features were mined successfully with the aid of RF decision tree algorithm. The experimental results have illustrated the efficacy of the designed prediction system in predicting the cardiac level of risk. However, we can state the most important steps as:

- The results issued from the proposed system have not only focused on informing about the presence of a risk or not, but also they have provided the level of the CHD risk for the patient
- It contributes in managing CHD by identifying early the patients, assessing accurately the risk, and improving the patient's perception of that risk
- Performing an evaluation of the performance of the proposed model based on a set of benchmark techniques.

REFERENCES

[1] Brieman L. (2001). Random Forests. In Machine Learning, Kluwer Academic Publisher, 45(1).
[2] Durairaj, M. & Meena, K. (2011). A Hybrid Prediction System Using Rough Sets and Artificial Neural Networks, Intl. Journal of Innovative Technology & Creative Engineering, 1 (7).
[3] Guru, N., Dahiya, A. & Rajpal, N. (2007). Decision Support System for Heart Disease Diagnosis Using Neural Network. Delhi Business Review, Vol. 8 (1).

[4] Ho, T.K.. (1998). The random subspace method for constructing decision forests. IEEE Trans. on Pattern Analysis and Machine Intelligence, 20(8).
[5] Moudani, W., Shahin, A., Chakik, F. & Mora-Camino, F. (2011). Dynamic Rough Sets Features Reduction. Intl. Journal of Computer Science and Information Security, Vol. 9(4).
[6] Rajeswari, K., Vaithiyanathan, V. & Amirtharaj, P. (2011). Prediction of Risk Score for Heart Disease in India Using Machine Intelligence. International Conference on Information and Network Technology, Vol. 4.
[7] Patil, S.B. & Kumaraswamy, Y.S. (2009). Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network, European Journal of Scientific Research, Vol. 31 (4), pp.642-656.
[8] Segrera, S. & Moreno, M. (2005). Multiclassifiers: applications, methods and architectures. Proc. of Intl. Workshop on Practical Applications of Agents and Multiagents Systems, 263–271.
[9] Srinivas, K., Kavihta, B. & Govrdhan, A. (2010). Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks. International Journal on Computer Science and Engineering, Vol. 2 (2).