ISSN 2395-695X (Print)
ISSN 2395-695X (Online)
International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)
Vol. 2, Special Issue 15, March 2016

ACUTE MYOCARDIAL INFARCTION RISK PREDICTION USING SVM

[1]Ms.J.JAFRINE JESSILET
Second Year [M.E]
Department of Computer Science and Engineering
Renganayagi Varatharaj College of Engineering
Sivakasi, Tamil Nadu, India jafri2011@gmail.com,
Indian

[2]Mr.R.KARTHIBAN
Assistant Professor
Department of Computer Science and Engineering
Renganayagi Varatharaj College of Engineering
Sivakasi, Tamil Nadu, India
karthibanlink@gmail.com, Indian

Abstract - Acute Myocardial Infarction(AMI) occurs when the flow of blood to the heart becomes blocked. They can cause tissue damage and can even be life-threatening. Acute myocardial infarction is one of the most vital causes of death. Early prediction of AMI is important. The Cardiovascular Health study Dataset(CHS) is used to predict the disease. Computerized information with advanced data mining techniques are used for appropriate results and for predicting the diseases from the CHS datasets. Develop AMI prediction models and investigate the effect of sample age and prediction resolution on elder subjects aged 65 and above. New unsupervised classification system is adopted for heart attack risk prediction at the early stage using the patient's medical records. The information in the patient record are pre-processed initially and then the attributes are classified using a SVM classifier. The cardiovascular health study dataset is used. We proposed to find out the AMI events through data mining, Support Vector Machine(SVM), EDC-AIRS Algorithm, Kennard-Stone (KS) algorithm and Genetic Algorithm(GA). This SVM algorithm shows high performance in predicting disease with less error. SVM algorithm is capable of achieving high accuracy with 90.1%.

Index terms – Acute Myocardial Infarction(AMI), Cardiovascular Health study Dataset(CHS), Support Vector Machine(SVM),Genetic Algorithm(GA).

INTRODUCTION

The heart is a compound body organ that pumps nearly 5l of blood in the body provided that the human body with improved materials [1].Recently Medical research shows that there is much interest from the scientific researchers in implementing the human intelligence, in health care industry. The human heart operation is complex and any failure is very dangerous to human life. Therefore, heart diagnose systems has been a main anxiety to the scientific researchers in the recent days [2]. Various methods have been proven to be effective in the identification systems [3]. Physician perception and experience are not always enough to achieve high quality medical results. So, Medical errors and unwanted results are reasons for a need for innovative computer-based analysis systems, which in turns reduce medical fatal errors, increase patient safety and save life.

The healthcare environment is generally perceived as being ‘information rich’ yet ‘data poor’. The assets are accessible within the healthcare system. However, there is a lack of successful analysis tools to discover hidden interaction and trends in data. Knowledge discovery and data mining have found many applications in business and scientific domain. Valuable data can be exposed from application of data mining techniques in healthcare system. The analysis of heart disease is a important and tedious task in medicine. Good performance of this method comes from the use of the SVM[4] and the relevant genetic algorithm[5]. Using medical profiles such as age, gender, smoke, cholesterol, hypertension and stroke it can predict the likelihood of patients getting a heart disease. For heart attack risk prediction, selected field values of each new record taken are to be processed for finding distance vector of the record. the distance vector is calculated with the minimum dynamic support and maximum dynamic support of frequent item-set combinations.

In general, feature selection aims to identify a cost-conscious subset of useful features (from a large set of features) that (1) does not reduce the classification accuracy, (2) reduces the computational time needed to learn a sufficiently exact classification model, (3) does not acutely changes the class distribution while adequately representative for descriptions of the target concept, and (4) reduces the amount of examples that need to be collected in order to develop a classification model with the desired precision (Dash & Liu, 1997). A GA based wrapper approach using SVM which selects important clinical features capable of dichotomizing patients experiencing a phenotypic appearance from healthy individuals, was implemented. This cross algorithm (called GA-SVM) was used to identify important clinical features.

I. RELATED WORK

Cardiovascular disease (CVD) is a group of diseases related with the heart and/or blood vessels. It includes disorders that cause (1) narrowing of blood vessels supplying blood to the heart (i.e. coronary heart disease), brain (i.e. cerebrovascular disease) and limbs (i.e. peripheral arterial disease), (2) damage to the heart muscle and heart valves from rheumatic fever (i.e. rheumatic heart disease), (3) weakening of heart muscle to pump adequate blood into the blood vessels (i.e. congestive heart failure), (4) abnormal formations of heart structures at birth (i.e. congenital heart disease), and (5) formation of blood clots in leg veins which could give rise to serve pain and disability, or even life threatening complications when the clots dislodge and move to the heart and lungs (i.e. deep venous thrombosis and pulmonary embolism) (World Health Organization, 2013).

In view of the detrimental impact of MI on the society, several epidemiology studies have been carried out to better understand and characterize the disease. This includes the Cardiovascular Health Study (CHS) (Fried et al., 1991), the Honolulu Heart Program (HHP) (Robertson et al., 1977; Marmot et al., 1975; Syme et al., 1975), the Framingham Heart Study (O'Donnella & Elosua, 2008) and the INTERHEART study (Ounpuu et al., 2001). These studies have identified major risk factors associated with CVD which include age, gender, cholesterol, hypertension, obesity, diabetes, smoking, alcohol, psychosocial factors, sedentary lifestyle and unhealthy diet (Yusuf et al., 2004; Hubert et al., 1983; Psaty et al., 2001; Stokes et al., 1989; Yano et al., 1984; Anand et al., 2008).

Broadly, risk factors can be categorized into 2 groups, namely non-modifiable and modifiable risk factors. The non-modifiable risk factors include age, gender, race and family history while the modifiable risk factors include blood pressure, cholesterol, body mass index, diabetes, smoking, diet and physical activities among others. Identification of these risk factors is important as they are measurable elements or characteristics that are causally correlated to an increased risk of a disease (O'Donnella & Elosua, 2008). However, caution need to be taken when analysing risk factors as their degree of impact on individuals' health may change as one ages (Asia Pacific Cohort Studies Collaboration, 2006) (which will be addressed in this thesis). Therefore, careful monitoring, analysis and management of these risk factors could reduce mortality rate.

II. SYSTEM DESIGN

Develop MI prediction models and investigate the effect of sample age and prediction resolution on the performance of MI risk prediction models. The cardiovascular health study dataset was used. One benefit of performing risk prediction using different prediction resolution and sample age is that it allows more refined and progressive risk prediction to be conducted. This provides the advantage of estimating the seriousness of a disease one is experiencing; enabling clinicians to offer a more personalized management and/or therapeutic strategy to the patient. Increase in the amount of clinical and molecular data collected from routine medical examination. To overcome the challenges associated with human scale of thinking and analysis, data mining techniques which have been postulated as a "central feature" for future health- care system became a popular method for extracting in- sights from this data deluge.

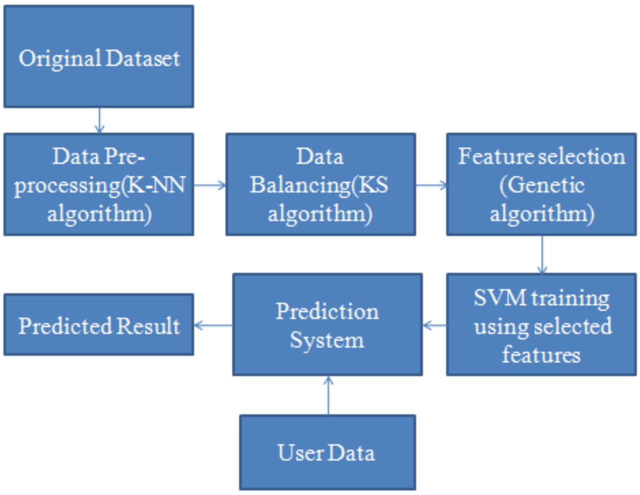

Original Dataset
Data Pre-processing(K-NN algorithm)
Data Balancing(KS algorithm)
Feature selection (Genetic algorithm)
SVM training using selected features
Prediction System
Predicted Result
User Data

Figure 1: System Configuration

A. DATA IMPUTATION

Data imputation is the process of substituting missing entries in a dataset with plausible values and aims to improve the quality of the data. Imputation is the process of replacing missing data with substituted values. When substituting for a data point, it is known as "unit imputation" when substituting for a component of a data point, it is known as "item imputation". Because missing data can create problems for analyzing data, imputation is seen as a way to avoid pitfalls involved with list wise deletion of cases that have missing values. When one or more values are missing for a case, most statistical packages default to discarding any case that has a missing value, which may introduce bias or affect the representativeness of the results. Imputation preserves all cases by replacing missing data with an estimated value based on other available information. Once all missing values have been imputed, the data set can then be analyzed using standard techniques for complete data.

In this stage, KNN imputation method is used to find the missing values. In the KNN method, missing values in a case are imputed using values calculated from the K nearest neighbor, hence the name.

B. DATA BALANCING

A Dataset is imbalanced if the classification categories are not approximately equally represented. Class imbalance problem become greatest issue in data mining. Imbalance problem occur where one of the two classes having more sample than other classes. The most of algorithm are more focusing on classification of major sample while ignoring or misclassifying minority sample. The minority samples are those that rarely occur but very important. In data-preprocessing technique sampling is applied on data in which either new samples are added or existing samples are removed. Process of adding new sample in existing is known as over-sampling and process of removing a sample known as under-sampling. Sampling techniques used to solve the problems with the distribution of a dataset, sampling techniques involve artificially re-sampling the data set, it also known as data preprocessing method. Sampling can be achieved by two ways, Under -sampling the majority class, oversampling the minority class, or by combining over and under -sampling techniques.

a) Under Sampling

The most important method in under -sampling is random under -sampling method which trying to balance the distribution of class by randomly removing majority class sample. In this paper Kennard-stone based under sampling method is used for balancing the dataset. The Kennard–Stone algorithm allows to select samples with a uniform distribution over the predictor space (Kennard and Stone, 1969).

b) Over Sampling

Random Oversampling methods also help to achieve balance class distribution by replication minority class sample. Kennard-Stone Sampling algorithm works as follows to Find the two most separated points in the ExampleSet. For each candidate point, find the smallest distance to any already selected object. Select the point which has the largest of these smallest distances. This algorithm always gives the same result because the two starting points are always the same. This implementation reduces the number of iterations by holding a list with candidates of the largest smallest distances.

C. FEATURE SELECTION

For Feature selection is a term commonly used in data mining to describe the tools and techniques available for reducing inputs to a manageable size for processing and analysis. Feature selection implies not only cardinality reduction, which means imposing an arbitrary or predefined cutoff on the number of attributes that can be considered when building a model, but also the choice of attributes, meaning that either the analyst or the modeling tool actively selects or discards attributes based on their usefulness for analysis. Usually, feature selection is performed automatically in Analysis Services, and each algorithm has a set of default techniques for intelligently applying feature reduction. Feature selection is always performed before the model is trained, to automatically choose the attributes in a dataset that are most likely to be used in the model. However, you can also manually set parameters to influence feature selection behavior. In general, feature selection works by calculating a score for each attribute, and then selecting only the attributes that have the best scores. In this paper genetic algorithm is used for the feature selection.

a) Genetic Algorithm

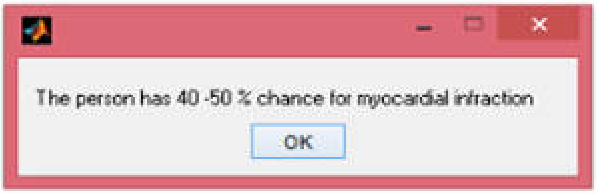
A Genetic Algorithm (GA) is a search heuristic that mimics the process of natural evolution. This heuristic is routinely used to generate useful solutions to optimization and search problems. Genetic algorithms belong to the larger class of Evolutionary Algorithms (EA), which generate solutions to optimization problems using techniques inspired by natural evolution, such as inheritance, mutation, selection, and crossover.

The fitness function is defined over the genetic representation and measures the quality of the represented solution. The fitness function is always problem dependent. For instance, in the knapsack problem one wants to maximize the total value of objects that can be put in a knapsack of some fixed capacity. A representation of a solution might be an array of bits, where each bit represents a different object, and the value of the bit (0 or 1) represents whether or not the object is in the knapsack. Not every such representation is valid, as the size of objects may exceed the capacity of the knapsack. The fitness of the solution is the sum of values of all objects in the knapsack if the representation is valid or 0 otherwise. In some problems, it is hard or even impossible to define the fitness expression; in these cases, interactive genetic algorithms are used.

Once the genetic representation and the fitness function is defined, GA proceeds to initialize a population of solutions randomly, and then improve it through repetitive application of mutation, crossover, inversion and selection operators.

D. SVM

Support Vector Machines map the training data into kernel space. There are many differently used kernel spaces–linear(uses dot product),quadratic, polynomial, Radial Basis, Funtion kernel, Multilayer Perceptron kernel, etc. to name a few. In addition, there are multiple methods of implementing SVM, such as quadratic programming, sequential minimal optimization, and least squares. The challenging aspect of SVM is kernel selection and method selection such that you model is not over optimistic or pessimistic.



The person has 40 -50 % chance for myocardial infraction

Figure 2: Prediction Of Patients Having Acute Myocardial Infraction

III. Conclusion

Early detection of individuals with high risk of experiencing MI is very important clinically, but has proved to be exclusive. we investigated the effect of sample age and prediction resolution in relation to the development of accurate clinical risk prediction model. Our experiments indicate that both sample age and prediction resolution do not have a significant impact on prediction models developed using subjects aged 65 and above. Overall, high validation sensitivity, specificity, and balanced accuracy were achieved by SVM algorithm.

FUTURE WORK

Constructing the predictive models capable of detecting MI early, allowing clinicians to take preventative measures promptly, improving the quality of individuals' life, and reducing avoidable mortality. In view of the observations from this study and the importance of screening since young, we aim to investigate the effect of prediction resolution and sample age on younger subjects as part of our future work.

REFERENCES

[1] Hannan, S.A., V.D. Bhagile, R.R. Manza and R.J. Ramteke, 2010. Diagnosis and medical prescription of heart disease using support vector machine and feedforward backpropagation technique. Int. J. Comput. Sci. Eng., 2: 2150-2159.

[2] Uguz, H," A biomedical system based on artificial neural network and principal component", analysis for diagnosis of the heart valve diseases. J. Med. Syst., 36: 61-72, 2012.

[3] Lomsky, M., P. Gjertsson, L. Johansson, J. Richter and M. Ohlsson et al.,. 2008." Evaluation of a decision support system for interpretation of myocardial perfusion" gated SPECT. Eur. J. Nuclear Med. Mol. Imaging, 35: 1523-1529.

[4] Chih-Wei Hsu, Chih-Chung Chang, and Chih- Jen Lin. "A Practical Guide to Support Vector Classification" . Deptt of Computer Sci. National Taiwan Uni, Taipei, 106, Taiwan http://www.csie.ntu.edu.tw/~cjlin 2007.

[5] Sivanandam S. N. Deepa S. N. Introduction to Genetic Algorithms Springer-Verlag , Berlin, Heidelberg, 2008.

[6] Darwin Tay, Chueh Loo Poh, Eric Van Reeth, and Richard I. Kitney, "The Effect of Sample Age and Prediction Resolution on Myocardial Infarction Risk Prediction," IEEE journal of biomedical and health informatics, vol. 19, no. 3, may 2015

[7] R. J. Hye, A. E. Smith, G. H. Wong, S. S. Vansomphone, R. D. Scott, and M. H. Kanter, "Leveraging the electronic medical record to implement an abdominal aortic aneurysm screening program," J. Vascular Surg., vol.59, no. 6, pp. 1535–1543, 2014.

[8] Akash jarad, Rohit katkar, abdul rehaman shaikh, anup salve, "Intelligent heart disease prediction system with mongodb," vol.4, ISSN 2278-6856, 2014.

[9] R. Chitra and V. Seenivasagam, " Review Of Heart Disease Prediction System Using Data Mining And Hybrid Intelligent Techniques ," Vol.03, ISSN 2229-6956, 2013.

[10] K. Chandra Shekar, K. Ravi Kanth, K. Sree Kanth, "Improved Algorithm for Prediction of Heart Disease on Non-Binary Datasets," Vol. 1, ISSN 2278-5841, 2013.

[11] Jyoti Soni, Uzma Ansari, Dipesh Sharma, "Intelligent and Effective Heart Disease Prediction System using Weighted Associative Classifiers," Vol. 3, ISSN 0975-3397, 2013.

[12] Abhishek Ta, "Heart Disease Prediction System Using Data Mining Techniques," Vol.6, ISSN 0974-6471, 2013.

[13] Song .X, Mitnitski .A, Cox .J, Rockwood .K, "Comparison of machine learning techniques with classical statistical models in predicting health outcomes," pt.1, PMID: 15360910 , 2011.

[14] Wayne C.Levy, Dariush Mozaffarian, santosh.C, "The seattle heart failure model prediction of survival in heart failure," pp.1, PMID 17894510, 2010.

[15] M.Akhil jabbar, B.Priti Chandra, L.Deekshatuluc, "Missing data imputation using statistical and machine learning methods in a real heart problem," pp.1, PMID 12889015, 2010.

Ms.J.Jafrine Jessilet received the B.E Degree in Computer Science and Engineering from Unnamalai Institute of Technology, Kovilpatti, Tamilnadu, India. which is affiliated to Anna University, Chennai in 2012 and she is currently pursuing the M.E degree in Computer Science and Engineering in Renganayagi Varatharaj College of Engineering Affiliated to Anna University, Chennai. Her area of Interest is Data mining.

Mr.R.Karthiban is currently working as assistant professor in computer science and engineering department in Renganayagi Varatharaj College of Engineering. He received his M.E degree from United Institue Of Technology, Coimbatore, Tamilnadu, India which is affiliated to Anna University, Chennai in 2008. He has a experience of above three years in teaching field. His area of interest is Cloud Computing, Data Mining.