International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 8, August - 2014

Early Detection of Coronary Vascular Disease using Data Mining: Literature Survey

Sikander Singh Khurl
Student
Dept of Computer Engineering
Punjabi University,Patiala

Gurpreet Singh
Assistant Professor
Dept of Computer Engineering
Punjabi University,Patiala

***Abstract*-In spite of advancements in data mining, the real-time problem remains uncovered and this leads to the wastage of time and economy. Coronary heart disease is the leading cause of mortality in modern society. Although science has made a significance progress in medical diagnosis but the improvement is still needed. And as a human it is our duty to serve patients with proper treatments before they breathe their last,because death is sure. 80% of patients died due to lack of care and care is the child of knowledge. If one could gave us a proper knowledge about the disease it would save thousands of lives. Here we take the support of Data mining and knowledge discovery,because medical science is the best field where these two have proven successful result.We need to again look back and see which technique has proven better results.**

Keywords:Data mining, CVD as cardiovascular disease, Naïve Bayes, clustering

I.INTRODUCTION

The National Heart Lung and Blood Institute reported that 872,000 deaths or 36% of all deaths in the United States were due to cardiovascular disease in 2004[1]. Approximately 50% of heart attack related deaths occur in people with no prior symptoms. Hence, sudden heart attack remains the number one cause of death in the US[9]. Unpredicted heart attacks account for the majority of the $280 billion burden of cardiovascular diseases. The field of cardiology has witnessed a major paradigm shift in its determination of a patient's risk of coronary artery disease. Today, cardiovascular specialists know that heart attacks are caused by inflammation of the coronary arteries and by formation of vulnerable plaques. As a result, the discovery of vulnerable plaque has recently evolved into the definition of "vulnerable patient". A vulnerable patient is defined as a person with more than a 10% likelihood of having a heart attack in the next 12 months [2].

Clinical decisions are often made based on doctor`s advice and experience rather than on the knowledge rich data hidden in the databases[3].

The Advantages of early prediction of CVD are:

- It can reduces the number of deaths due to heart attack and increase patient`s safety.
- It can reduce cost, errors and biases which affects the quality of service provided to the patients.
- It can be a patient`s nursing assistant in remote locations.
- It can create better databases for future references, when found useful.

A. Causes and impact of heart diseases

According to WHO report Global atlas on cardiovascular disease prevention and control states that cardiovascular disease (CVDs) are the leading causes of death and disability in the world. Although a large proportion of CVDs is preventable, they continue to rise mainly because preventive measures are inadequate.

B. Cardiovascular Diseases Key Facts

1. CVDs are the number one cause of death globally: more people die annually from CVDs than from any other cause.
2. An estimated 17.3 million people died from CVDs in 2008, representing 30% of all global deaths. Of these deaths, an estimated 7.3 million were due to coronary heart disease and 6.2 million were due to stroke.
3. Low and middle income countries are disproportionally affected: over 80% of CVD deaths take place in low and middle income countries and occur almost equally in men and women.
4. By 2030, almost 23.6 million people will die from CVDs, mainly from heart disease and stroke. These are projected to remain the single leading causes of death.

II. DETECTING WAYS AND PARAMETERS

Watching CVD the reasons behind the cause are the set of Parameters which can lead a human to corpse. Obviously if we can able to detect and understand these parameters then one can insure the safety of the patients. All the authors in their papers discussed about the set of parameters and some discuss about angiography images. So there are only two ways to detect the presence of the CVD:

1.Using parameters and detect the presence of CVD.
2.Using angiography images, by properly scanning and figure out the point of problem.

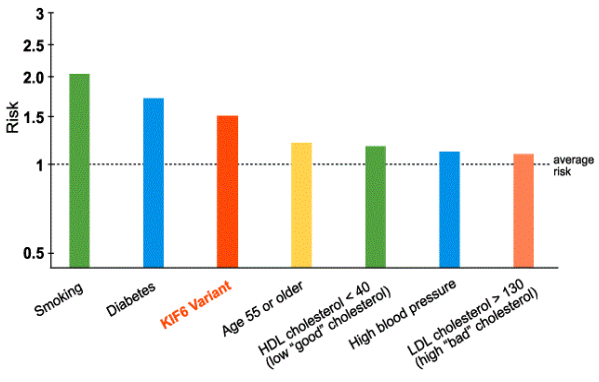

Fig 1.Parameters with risk factors

CVD is caused by disorders of the heart and blood vessels and cerebrovascular disease(stroke), raising Blood pressure(hypertension), Chest pain,congenital heart disease and hear failure. Some of the parameters are age, sex, cholesterol, air pollution, radiation on chest, smoking, balance diet, alcohol use, tobacco, mental trauma, hypertension ,up take of red meat, obesity, genetic reason, history of preeclampsia, back pain and previous health record etc.

Here age, sex, hypertension, pollution,chest pain, cholesterol, genetic problems are the parameters which can`t be ignored[8].

Australian authors believe that the chances of CVD is more in women than in men but American and UK based authors believes the vice-versa. It gives clearer understanding that environmentis a greatest factor for the variation. Similarly in India the impact of CVD is more in women.The presence of vasa vasorum(VV) neovascularization on the plague has been identified as a common feature inflammation and has been identified as plague vulnerability index[2].

Some authors believe that using single data mining technique is not firm in case of medical disorders. But hybrid data mining is essential [5]. Some of the aspects are taken by applying both the techniques on the same dataset and found that best accuracy achieved using single data mining technique is 84.14% by naïve bayes. However the best accuracy achieved by using hybrid data mining technique is 89.01% by neural network ensemble[10].

III. DATA MINING ALGORITHMS AND METHODS

Each author has applied their preferred algorithm on their relevant data sets. Its mean that there is no such standard datasets upon which these respective techniques are applied.

A. Classification

Classification is the most commonly used data mining technique but it scores the least in detection of the CVD. It employees neural network and decision tree classification algorithm[6]. Yan,et al(2003) used decision tree classification and found that it is 75.738% efficient in detecting the CVD[5]. On the other hand De Beule,et al(2007) employees artificial neural network can score 82% in detecting the disease. Similarly other authors employees classification on their respective datasets and found:

Table I.

Author	Year	Technique	Accuracy
Palaniappan et al	2007	Decision tree	94.93%
Sitar-Taut, et al	2009	Decision tree	60.40%
Kangwanariyakul, et al	2010	Neural network	70.59%

B. Clustering

Clustering can be said as identification of similar classes of objects. By using clustering techniques we can further identify dense and sparse regions in object space and can discover overall distribution pattern and correlations among data attributes. It is famous for its Naïve Bayes and K-means clustering techniques[4]. The most relevant and efficient till date is the Naïve Bayes but it is complex. The maximum score for the Naïve Bayes is 96.5% and minimum score is 78.6%[7]. Similarly other authors employees clustering on their respective datasets and found:

TABLE II.

Author	Year	Technique	Accuracy
Yan, et al	2006	Naïve Bayes	78.6%
Palaniappan et al	2007	Naïve Bayes	95%
Anbarasi, et al	2010	Classification via clustering	88.3%

C. Data mining with Genetic algorithm

Moving from classification concepts GA or Genetic algorithm in applied when the amount of data is huge. Genetic algorithm is the first preference of modern DM researcher scholars.Because GA is once started it will keep finding the optimum resultant of the previous. Suppose β is the first randomly generated generation. Now new generation is produced by applying reproduction operator. This process is repeated until a better solution or a better generation is reached[11]. All the genes are mutated before applying the next reproduction operator.

Genetic algorithm is used to reduce the actual data size & to get the optimal subset of the attributed data from CVD prediction. Decision tree, Naïve Bayes & classification via clustering is used to predict. A pair of generation is taken & crossover is performed is repeated until sufficient generation is reached. GA applied on different data sets as:

TABLE III.

Author	Year	Technique	Accuracy
Anbarasi, et al	2010	Genetic with decision tree	99.2%
Anbarasi, et al	2010	Genetic with Naïve Bayes	96.5%
Anbarasi, et al	2010	Genetic with Naïve Bayes Classification via clustering	88.3%

D. Association rule discovery

Associativity rules plays a vital role in the discovery of CVD. But the main issue is that when rules are applied on medical data, a huge set of rules are formed. Which can be irrelevant& time impractical. So to decrease this list of rules four attributes to be concerned:

1. Item filtering
2. Attribute grouping
3. Maximum item set size
4. Antecedent/Consequent rule

But such factors were adequate until the discovery of MRI and CT scans in 2001[11]. Because these two techniques generate a huge amount of data. So such factors are not enough in today`s theory and to figure out which rule is more sensitive than others.

Then Aqueel ahmend & haikh abdul hannan(2012) added a new parameter in Associativity rule discovery i.e.,lift. This factor check the sensitivity & specificity of the disease. Thus both of them have given the following advanced algorithm:

1. Transform the categorical & numerical data into transactional data.
2. Apply four attribute filtering on concerned rules.
3. Train and test the data.

E. Rough set theory

After associativity rules there is another aspect which is mostly ignored it is the rough set theory. The resultant knowledge discovery left in decision tree , association rules etc. Which contribute a large set of databases. In order to interpret data it takes enormous time to get work[14]. To optimize these impractical rules rough set theory was introduced by Zdzislaw pawlak(1989). Rule pruning is a mathematical method to manage uncertainties, ambiguity and vagueness from incomplete and noisy information.

IV. CONCLUSION

In this paper we briefly reviewed the various data mining algorithms on heart databases from its inception to the future. This review would be helpful to researchers to focus on the various issues of data mining. Our main goal is to find a better risk assessment index for an individual`s risk of coronary heart events. If this score is used as a predictor of the future adverse events, it may provide best results in subclinical information. We can compare the images of heart events to find out the disease,in future.

REFERENCES

[1] National Institute of Health,"Disease statistics", National heart Lung and blood Institute and Technology.Rep.,2007
[2] I.A. Kakadiaris and U. Kurkure, "Towards Cardiovascular Risk Stratification Using Imaging Data",Minneapolis Minnesota, USA, 2009
[3]Chen, J., Greiner, R.: Comparing Bayesian Network Classifiers. In Proc. of UAI-99, pp.101–108,1999.
[4] Aqueel Ahmed, Shaikh Abdul Hannan, "Data Mining Techniques to Find Out Heart Diseases: An Overview",Mahavidyalaya, Aurangabad, 2012
[5]Mai Shouman, "Using data mining techniques in heart disease and treatment", Canberra,2012
[6]Polat , K., S. Sahan, and S. Gunes, "Automatic detection of heart disease using an artificial immune recognition system (AIRS) with fuzzy resource allocation mechanism" 2007
[7] Srinivas, K., B.K. Rani, and A. Govrdhan, "Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks". International Journal on Computer Science and Engineering (IJCSE), 2010.
[8] Simons, L.A., et al., "Risk functions for prediction of cardiovascular disease in elderly Australians: the Dubbo Study". Medical Journal of Australia, 2003.
[9] Wilson, P.W.F., et al., "Prediction of Coronary Heart Disease Using Risk Factor Categories". American Heart Association Journal, 1998.
[10] Das, R., I. Turkoglu, and A. Sengur, "Effective diagnosis of heart disease through neural networks ensembles"., Elsevier, 2009.
[11] Porter, T. and B. Green, "Identifying Diabetic Patients: A Data Mining Approach. Americas" Conference on Information Systems, 2009.
[12] Tu, M.C., D. Shin, and D. Shin, "Effective Diagnosis of Heart Disease through Bagging Approach"., IEEE, 2009
[13] Podgorelec, V., et al., "Decision Trees: An Overview and Their Use in Medicine". Journal of Medical Systems, 2002.
[14] Andreeva, P., "Data Modelling and Specific Rule Generation via Data Mining Techniques". International Conference on Computer Systems and Technologies - CompSysTech, 2006.
[15] N.Satyanandam, Dr. Ch. Satyanarayana, Md.Riyazuddin, A.Shaik."Data Mining Machine Learning Approaches and Medical Diagnose Systems" A Survey. International journal of computer applications, 2009