International Journal of Engineering Sciences Paradigms and Researches (IJESPR)
(Vol. 20, Issue 01) and (Publishing Month: May 2015)
(An Indexed, Referred and Impact Factor Journal)
ISSN (Online): 2319-6564
www.ijesonline.com

A Literature Survey on Applications of Data Mining Techniques to Predict Heart Diseases

Shimpy Goyal[1] and Dr. Rajender Singh Chhillar[2]

[1]Student, Department of Computer Science & Applications
Maharshi Dayanand University, Rohtak (Haryana) India
shimpygoyaluiet@gmail.com

[2]Professor, Department of Computer Science & Applications
Maharshi Dayanand University, Rohtak (Haryana) India
chhillar02@gmail.com

Abstract

Data mining is a phenomenon which is used to analyze large volumes of data and extracts patterns that can be converted to useful knowledge. The data mining techniques can be applied on medical data. By using software to look for patterns in large batches of data, businesses can learn more about their customers and develop more effective knowledge. As in medical science record we found 25 per cent of deaths in the age group of 26- 69 years occur because of heart diseases. If all age people are included, heart diseases account for about 19 per cent of all death rates. It is the leading cause of death among males as well as females. However researchers have proposed various tools to detect and diagnosing disease using various algorithms and prototypes like naïve bayes and weighted associative classifier (WAC) but still many cases come to us by random sampling where we fail to predict the original case. So our aim here to analyzes how data mining techniques are used for predicting different types of diseases so this paper reviewed the research papers which mainly concentrated on predicting heart disease, Diabetes and Breast cancer. The systematic literature survey is based on collection of different international journals and conferences and world health organization (WHO) reports. These journals have been published within the time span of 2008 to 2015.

Keywords: *Naïve bayes, weighted associative classifier (WAC), data mining, kidney failure, heart disease**,** A-prior and k-mean algorithm.*

1. Introduction

Knowledge discovery in databases is well-defined process consisting of several distinct steps. Data mining is the core step, which results in the discovery of hidden but useful knowledge from massive databases. Quality service implies diagnosing patients correctly and administering treatments that are effective. Poor clinical decisions can lead to disastrous consequences which are therefore unacceptable.

Treatment records of millions of patients can be stored and computerized and data mining techniques may help in answering several important and critical questions related to health care.

The availability of integrated information via the huge patient repositories, there is a shift in the perception of clinicians, patients and payers from qualitative visualization of clinical data by demanding a more quantitative assessment of information with the supporting of all clinical and imaging data. For instance it might now be possible for the physicians to compare diagnostic information of various patients with identical conditions. The term Kidney failure and heart disease applies to a number of illnesses that affect the circulatory system, which consists of heart and blood vessels. It is intended to deal only with the condition and the factors, which lead to such condition. Acute kidney injury (also called acute renal failure) means that your kidneys have suddenly stopped working. This can cause problems that can be deadly. So Detecting and diagnosing disease by using the k-means and Apriori.

1.1 The *k*-means algorithm:

The k-means algorithm is a simple iterative method to partition a given dataset into a specified number of clusters, *k*. This algorithm has been discovered by several researchers across different disciplines. The algorithm operates on a set of *d*-dimensional vectors, $D = \{\mathbf{xi} \mid i = 1, \ldots, N\}$, where $\mathbf{xi} \in \mathrm{R}d$ denotes the *i*th data point. The algorithm is initialized by picking *k* points in R*d* as the initial *k* cluster representatives or “centroids”. Techniques for selecting these initial seeds include sampling at random from the dataset, setting them as the solution of clustering a small subset of the data or perturbing the global mean of the data *k* times. The simple way to understand K-means is:

- Requires real-valued data.
- We must select the number of clusters present in the data.
- Works best when the clusters in the data are of approximately equal size.
- Attribute significance cannot be determined.
- Lacks explanation capabilities.

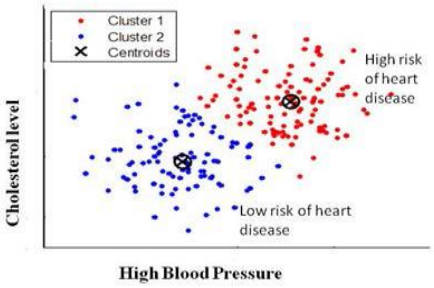

Figure1: K-means clustering for Heart Disease Patients

1.2 The Apriori algorithm

Apriori algorithm for association is proposed by R.Agarwal., in 1994. It finds out the relationships among item sets using two inputs support and confidence. One of the most popular data mining approaches is to find frequent item sets from a transaction dataset and derive association rules. Finding frequent item sets (item sets with frequency larger than or equal to a user specified minimum support) is not trivial because of its combinatorial explosion. Once frequent itemsets are obtained, it is straightforward to generate association rules with confidence larger than or equal to a user specified minimum confidence. Apriori is a seminal algorithm for finding frequent itemsets using candidate generation. It is characterized as a level-wise complete search algorithm using anti-monotonicity of itemsets, “if an itemset is not frequent, any of its superset is never frequent”. By convention, Apriori assumes that items within a transaction or itemset are sorted in lexicographic order.

2. Review of Literature

In the diagnosis of heart disease large number of work is carried out, researchers have been investigating the use of data mining techniques to help professionals. Many risk factors associated with heart disease like age, sex, chest pain, blood pressure, cholesterol, blood sugar, family history of heart disease, obesity, and physical inactivity. Knowledge of these risk factors medical professionals can diagnosis the heart disease in patients easily.

World health organization [1] presented a the ten leading causes of death by broad income group 2008 .According to WHO (World health organization) report cardiovascular diseases kill more people every year than any others. In 2008, 7.3 million people died from ischaemic heart disease, 6.2 million from stroke or another form of cerebrovascular disease. Tobacco use is a major cause of many of the world’s top killer diseases – including cardiovascular disease. In high-income countries more than two thirds of all people live beyond the age of 70 and predominantly die of chronic diseases: cardiovascular disease, chronic obstructive lung disease, cancers, diabetes or dementia. In middle-income countries, nearly half of all people live to the age of 70 and chronic diseases are the major killers, just as they are in high-income countries. In low-income countries less than one in five of all people reach the age of 70, and more than a third of all deaths are among children under 15.More than 8 million deaths in 2008 were among children under five years of age, and 99% of them were in low- and middle-income countries.Report conclude that counting the dead matters because Measuring how many people die each year and why they have died is one of the most important means – along with gauging how various diseases and injuries are affecting the living– for assessing the effectiveness of a country’s health system. Having those numbers helps health authorities determine whether they are focusing on the right kinds of public health actions.

Vikas Chaurasia *et al*.[2]The objective of this research work is to predict more accurately the presence of heart disease with reduced number of attributes. Originally, thirteen attributes were involved in predicting the heart disease. Thirteen attributes are reduced to 11 attributes. Three classifiers like Naive Bayes, J48 Decision Tree and Bagging algorithm are used to predict the diagnosis of patients with the same accuracy as obtained before the reduction of number of attributes. The empirical results show that they can produce short but accurate prediction list for the heart patients by applying the predictive models to the records of incoming new patients. This study will also work to identify those patients which needed special attention for treatment.

My Chau Tu *et al.* [3], proposed the use of a bagging algorithm to indentify the warning signs of heart disease in patients and also to compare the effectiveness of the bagging algorithm with the decision tree algorithm that has been used by many researchers. As the diagnosis of heart disease is important issue, prompting many researchers to work on development of intelligent medical decision support systems to improve the ability of physicians.

M. Akhil jabbar *et al.* [4] proposed a efficient associative classification algorithm using genetic approach for heart disease prediction. The main motivation for using genetic algorithm in the discovery of high level prediction rule is that discovered rules are highly comprehensible, having predictive accuracy and of high interestingness values. Experimental results show that most of the classifier rules help in the best prediction of heart disease which even help doctor in their diagonose decision. It has certain limitation that it uses large number to predict the heart disease using data mining approach.we can reduce the number of attributes to make it less complex and better.

N. Aditya Sundar *et al* .[5] presented a training tool to train nurses and medical students to diagnose patients with heart disease using Naïve Bayes and WAC(weighted associative classifier). It is a web based user friendly system and can be used in hospitals if they have a data ware house for their hospital. Presently they are analyzing the performances of the two classification data mining techniques by using various performance measures such as matrix, lift chart & bar chart. But it has some drawbacks like it uses only for categorical data and other limitations only 2 data mining techniques are used .we can use more techniques for providing more detailed explanation.

C Y Hsu *et al.* [6] presented Few studies have defined how the risk of hospital-acquired acute renal failure varies with the level of estimated glomerular filtration rate They conclude a heightened risk of ARF is another adverse sequela of chronic kidney disease that becomes apparent at an estimated GFR of below 60ml per min per 1.73m^2. Even subjects with estimated GFR 45–59ml per min per 1.73m^2 had a twofold increase in adjusted odds ratio of ARF compared with subjects with estimated GFR 60ml per min per 1.73m^2 or above. Their study is distinguished from prior investigations of ARF that have mostly focused on hospital-based or intensive care unit-based risk factors. Their results suggest that other markers of chronic kidney disease–such as proteinuria–and other risk factors for chronic kidney disease–such as hypertension and diabetes–are also risk factors for ARF. Limitations of this study include the fact that we did not study hospitalized patients without serum creatinine determinations before admission. We may have missed cases of dialysis-requiring ARF, as it was not feasible to review medical records of all cases. However, recent data [7] show that administrative codes for dialysis-requiring ARF have a sensitivity of over 90%.

Mohammed Abdul Khaleel *et al.* [8] presented methodology to discover locally frequent diseases with the help of Apriori data mining technique. They also used visualization techniques to present the trends graphically. They built a prototype application that demonstrates the efficiency of the method. The empirical results revealed that the prototype is useful and can be used in real world Healthcare tools. One future direction we have in mind is working on discovering temporally frequent diseases in near future.

Chris Ding *et al* [9] presented a Mapping data points into a higher dimensional space via kernels, they prove that principal components are the continous solutions to the discrete cluster membership indicator for K-mean clustering. Experiments indicate that newly derived lower bounds for K-means objective are within 0.5-1.5%of the optimal values. On learning, our results suggest effective techniques for K-means clustering. DNA gene expression and Internet newsgroups are analyzed to illustrate the results.

K.R. Lakshmi *et al.*[10] analyzing the performances of the ten classification data mining techniques by using various performance measures. For implementation of the work a real time patient database is taken and the patient records are experimented and the final best classifier is identified with quick response time and least error rate. A typical confusion matrix is furthermore displayed for quick check. The study describes algorithmic discussion of the heart disease dataset from Cleveland Heart Disease database, on line repository of large datasets. The Best results are achieved PLS-DA algorithm. PLS-DA shows better results. It also results sequence based classification with very least error rate and which increases the accuracy rate. The performance of PLS-DA shows the high level compare with other classifiers. Hence PLS-DA shows the concrete results with different Heart disease of patient records. Therefore PLS-DA classifier is suggested for Heart disease based classification to get better results with accuracy and performance.

Boshra Bahrami *et al*.[11] evaluate different classification techniques in heart disease diagnosis. Classifiers like J48 Decision Tree, K Nearest Neighbors(KNN), Naive Bayes(NB), and SMO are used to classify dataset. After classification, some performance evaluation measures like accuracy, precision, sensitivity, specificity, F-measure and area under ROC curve are evaluated and compared. The comparison results show that J48 Decision tree is the best classifier for heart disease diagnosis on the existing dataset.

3. Conclusion

The systematic Literature survey investigated existing tools based on K-means and apriori algorithm. The researchers also presented challenges in detecting & diagnose the diseases and analyze results of research .As seen from the year of publication of the articles, it can be seen that the researchers presented different problems based approaches from the year 2008 to 2015. They studied of exiting techniques and software for Detecting and diagnose disease using K-means and Apriori algorithm over disease prediction system .There was good coverage in terms of research in understanding the behaviour of components, interactions and compatibility of components The author contributed in the field of predicting of heart disease or kidney failure etc which even helps doctors in their diagnosis decisions by integrating A-prior and k-mean algorithm.

4. Future Scope

The paper provides a comparison study of many data mining techniques; like K-means and Apriori algorithm over disease prediction system Experimental Results shows that many of the rules help in the best prediction The Future scope for this paper is that it can be implemented for more number of diseases. More effective if implemented using Artificial Intelligence in future. Experimental Results will show that many of the rules help in the best prediction of heart disease and kidney failure which even helps doctors in their diagnosis decisions by using combination of A-prior and k-mean algorithm. By using combination of A-prior and k-mean algorithm we can proposed easy and efficient way in which we can find the stage of the kidney failure and heart disease.

References

[1] World Health Organization. 2008 May 2011]; Available from http://www.who.int/mediacentre/factsheets/fs310_2008.pdf

[2] Vikas Chaurasia, Saurabh Pal Data Mining Approach to Detect Heart Dieses International Journal of Advanced Computer Science and Information Technology (IJACSIT) Vol. 2, No. 4, 2013, Page: 56-66, ISSN: 2296-1739 © Helvetic Editions LTD, Switzerland www.elvedit.com

[3] My Chau Tu, Dongil Shin, Dongkyoo Shin ,"Effective Diagnosis of Heart Disease through Bagging Approach", 2nd International Conference on Biomedical Engineering and Informatics,2009.

[4] M.Akhil jabbar, Dr.Priti Chandra, Dr.B.L Deekshatulu " Heart Disease Prediction System using Associative Classification and Genetic Algorithm". ICECIT, 2012

[5] N. Aditya Sundar, P. Pushpa Latha, M. Rama Chandra," Performance analysis of classification data mining techniques over heart disease data base," International journal of engineering science & advanced technology Volume-2, Issue-3, 470 – 478.

[6] C Y Hsu, J D Ordoñez "The risk of acute renal failure in patients with chronic kidney disease". 2 April 2008

[7] Waikar SS, Wald R, Chertow GM *et al*. Validity of International Classification of

Diseases, Ninth Revision, Clinical Modification Codes for acute renal failure. *J Am Soc Nephrol* 2006; 17: 1688–1694

[8] Mohammed Abdul Khaleel, Sateesh Kumar Pradhan,” Finding Locally Frequent Diseases Using Modified Apriori Algorithm,” International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 10, October 2013.

[9] Chris Ding, Xiaofeng He, ”K-means Clustering via Principal Component Analysis, Chris”, Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720

[10] K.R. Lakshmi, M.Veera Krishna, S.Prem Kumar “Performance Comparison of Data Mining Techniques for Predicting of Heart Disease Survivability,”International Journal of Scientific and Research Publications.

[11] Boshra Bahrami, Mirsaeid Hosseini Shirvani “Prediction and Diagnosis of Heart Disease by Data Mining Techniques ”Journal of Multidisciplinary Engineering Science and Technology (JMEST) ISSN: 3159-0040 Vol. 2 Issue 2, February–2015.