Heart Disease Diagnosis using KNN Clustering Technique

Babitha M[1] Manikandan M[2]
[1,2]Department of Computer Science & Engineering
[1,2]Adhiyamaan College of Engineering, India

Abstract— Data mining is an iterative progress in which evolution is defined by detection, through usual or manual methods. Knowledge discovery and data mining have found various applications in scientific domain. Heart disease is a term for defining a huge amount of healthcare conditions that are related to the heart. This medicinal condition defines the unpredicted health conditions that directly control all the parts of the heart. Different data mining techniques such as association rule mining, classification, clustering are used to predict the heart disease in health care industry. The heart disease database is pre-processed to make the mining process more efficient. The pre-processed data is clustered using clustering algorithms like K-Nearest Neighbour (KNN) to cluster relevant data in database. International Classification of Diseases (ICD) Data is used for mining maximal frequent patterns in heart disease database. The frequent patterns can be classified using KNN algorithm as training algorithm using the concept of information entropy. The results showed that the designed prediction system is capable of predicting the heart attack with good accuracy.

Key words: KNN, Clustering Technique, ICD

I. INTRODUCTION

Among various life- threatening diseases, heart diseases have a great deal of attention in medical research. Also, it has more impact on human health. Various heart diseases was discussed and founded how they lead to heart attack (Jae-Hong Eom et.al). The number one cause of death in industrialized countries was due to cardiovascular disease. Cardiovascular diseases not only have a major impact on individuals and their quality of life in general, but also on public health costs and the countries' economies. Risk factors for these pathologies include diabetes, smoking, family history, obesity, high cholesterol etc (Jesmin Nahar et.al). Health information decision was enabled by particularly knowing about the anatomy and functioning of the heart. A new born infant also have the possibility of heart disease. Some of the symptoms of heart disease in people were chest pain and fatigue. It occurred while the heart does not meet the circulatory demands of the body (Chang-Sik Son et.al). The physician takes decision based on the patient's answers to questions and lab results (Hongmei Yan et.al). Blood flow to the heart muscles was decreased when block occurs in coronary arteries. The electrocardiogram recordings were analyzed to detect irregularity of heart beat problems occurred due to cardiovascular diseases (Swati Shilaskar). In advance of medical and surgical treatment the patient with heart disease reached adulthood (Petra A. Karsdorp et.al). There are many diseases that affect the heart and arteries but four are particularly prevalent. Myocardial infarction was linked to damage to the coronary arteries in 90% of cases. Strokes occurred as a result of impaired blood flow to the brain linked to a hemorrhage or a blockage of the arteries that supply blood to the brain. Heart failure was mainly linked to various changes in cardiovascular tissues, most often the result of ageing. High blood pressure was defined as the sustained elevation of arterial blood pressure in comparison to what is considered to be the "normal" value of 140/90 millimeters of mercury. There is a wide range of long-term consequences: heart failure, stroke, kidney failure etc. The different types of heart disease widely in the world are Coronary heart disease, Heart failure, Coronary artery disease, Ischemic heart disease, Cardiovascular disease, Hypoplastic left heart syndrome, Atherosclerosis, Chronic obstructive pulmonary disease, Congenital heart disease, Valvular heart disease. Mostly heart attacks are occurred when the plaque on the artery ruptures and a clot then forms, stopping blood flow. And the diagnosis of heart disease was based on medical knowledge occurred from patients. Correct diagnosis of the heart patient was delayed due to various problems. Diagnosis of heart disease was more costly and optimal decision path finder was used in terms of diagnostic accuracy while minimizing cost in diagnosis (Chih-Lin Chi et.al). Heart disease can strike suddenly and quick decisions have to be made. Prediction of heart diseases can provide some useful information about the health of patient. The prediction can be done with various computer aided diagnosis methods.

Generally, artificial intelligence techniques were used in medical diagnosis with an improvement in prediction of heart disease (Ismail Babaoglu et.al). Machine learning algorithm was used in medical diagnostic problem for heart disease (Evanthia E et.al). Case-based reasoning (CBR) was considered as a suitable technique for diagnosis, prognosis and prescription in the medical domain puts more stress on real cases than other domains (Yoon-Joo Park et.al). Coronary Artery Disease was diagnosed using two techniques called Binary Particle Swarm Optimization (BPSO) and Genetic Algorithm (GA). For the diagnosis of heart disease various classification and regression processes was used. It provides medical knowledge for diagnosis purpose.

II. RELATED WORK

Mu-Jung Huang et.al, have proposed a method by combining data mining and CBR to prognosis and diagnosis of chronic diseases. The implicit meaningful rules from health examination data was discovered by the process of adopting data mining techniques. The prognosis of chronic disease was identified by the extracted rules. Then diagnosis and treatment was supported by employing CBR. For the convenience of chronic diseases knowledge creating, organizing, refining and sharing the process was expanded to work within the system. After prognosis the suffering probability of new case chronic diseases was discovered by rules basically and then it trigger CBR. The mechanism of CBR was to retrieve most similar case from the case library. MJ health screening center collected health examination data and implemented through the system for prognosis and diagnosis of heart diseases and it was helpful reference for doctors and patients in chronic disease treatments.

Balakrishnan et.al proposed a concept for understanding the variation of voice in the coronary heart disease patients for detection of CHD. Computerized Speech Lab (CSL) model 4500 is used for processing the voice signal. CSL contains Multi-Dimensional Voice Program (MDVP) that breaks down and shows up to 22 voice parameters from a voice test. Voice samples of a group of 100 coronary heart disease patients (males and females) are compared with that of a group of 100 normal people (males and females). The study reveals variations in voice parameters like spectrogram, long term average spectrum (LTAS), jitter, shimmer, amplitude perturbation quotient (APQ) of the coronary heart disease patients in comparison with the normal people.

N.A. Setiawan, et al, have applied the three imputation methods namely Artificial Neural Network with Rough Set Theory (ANNRST), k-Nearest Neighbor (k-NN) and Concept Most Common Attribute Value Filling (CMCF) to University California Irvine (UCI) coronary heart disease data sets. The effect of missing attribute was investigated by comparing the three imputation methods with coronary heart disease data sets of University California Irvine. The rules were generated from the three data sets using the method called Rough Set Theory (RST). While filtering the generated rules the most complete data set of UCI coronary heart disease data is used as test data. Support filtering was applied on three sets of generated rules. In the case of UCI coronary heart disease data sets ANNRST could be considered as the best method.

Balakrishnan et.al presented a two way classification algorithm for the classification of breast cancer images into benign (tumour growing, but not dangerous) and malignant (cannot be controlled, it causes death) classes. Because of the sparse distribution of abnormal mammograms, the two-way classification data mining algorithms are used. First classification algorithm is k-means algorithm which is used to partition a given dataset into a user specified number of clusters. Second classification algorithm is Support Vector Machine (SVM) is used to find the best classification function to distinguish between members of the two classes in the training data.

III. PROPOSED MODEL

The proposed system architecture is given in the figure 1.

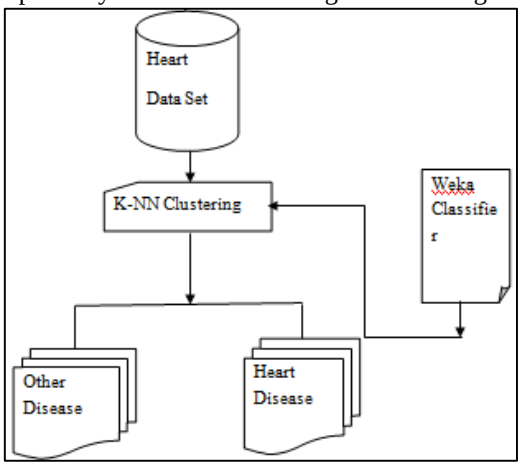


Fig. 1: Proposed System Architecture

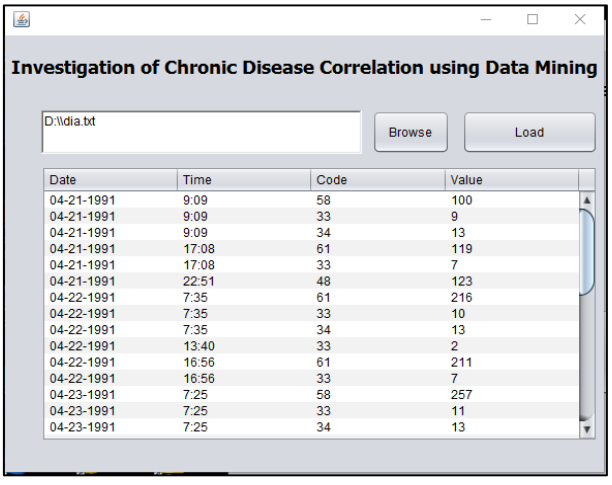
The information from the database (DB) is rebuilt i.e. the Beneficiary, Inpatient files and Outpatient files are connected using the primary key i.e. beneficiary id. The ICD9 diagnostic codes for the IP and OP files are also included in the data definition along with the chronic diseases. This definition helps to explore the ICD9 diagnostic codes and the chronic diseases effectively. Then the data is extracted by selecting the unique 2 digit and 3 digit ICD9 diagnostic codes, and the top N among them is extracted based on their occurrence. At that point division of information into tests in view of the interminable infections consolidated with each of the other endless ailments is finished. Here other chronic diseases along with diabetes and ischemic heart disease are extracted e.g. beneficiaries having diabetes and arthritis. Similarly every blend of 10 chronic disease with diabetes and ischemic heart disease is extracted. If a beneficiary has a particular chronic disease combination a value 1 is assigned this forms class C1 and value 2 for healthy patients which gives class C2. Data Mining Techniques Module: The next step is to apply various data mining techniques on the extracted data. It was observed that in the extracted data number of healthy individuals outnumber the beneficiaries with chronic diseases. This is a common observation in most of the healthcare datasets. This is leads to bias to a particular class during the learning process. To avoid this, the extracted data is resampled.

Resampling is one of the information mining procedure which guarantees uniform class dispersion, it appropriates the dataset consistently. At that point highlight choice systems are utilized to get the diminished arrangement of ICD9 codes. At that point these decreased arrangement of analytic codes are accepted utilizing an order calculation. Cross acceptance system is utilized for testing the decreased arrangement of codes.

IV. RESULTS AND DISCUSSION

This section explain about the results of our proposed system. The following screenshot shows the investigation of chronic heart disease using data mining techniques.

The following screenshot shows the accuracy of different diseases.



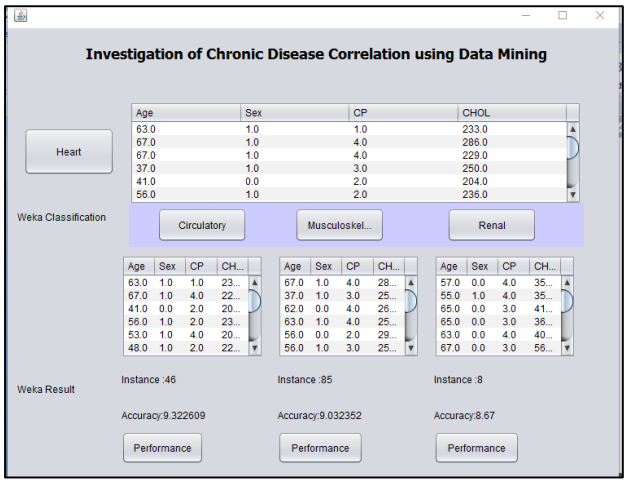


Fig. 2: Shows the accuracy of different diseases

V. CONCLUSION

The beneficiary file, IP record, OP document from the CMS data set was restructured and relevant information is extracted based on the combination of chronic diseases i.e. for diabetes and ischemic heart disease. To achieve this goal various types of data mining techniques are used. The conclusion is an ideal arrangement of ICD9 demonstrative codes connected with people having diabetes or ischemic coronary illness. We strongly believe that the performance of the proposed system can be further enhanced by framing new functionals that are more adaptable.

REFERENCES

[1] Jae-Hong Eom and Sung-Chun Kim, (2008) "AptaCDSS-E: A classifier ensemble-based clinical decision support system for cardiovascular disease level prediction", Journal of Expert Systems with Applications, Vol. 34 2465, PP.2479, 2008.

[2] Jesmin Nahar and Tasadduq Imam, (2013) " Association rule mining to detect factors which contribute to heart disease in males and females", Journal of Expert Systems with Applications Vol.40, PP.1086–1093, 2013.

[3] Chang-Sik Son and Yoon-Nyun Kim, (2012) "Decision-making model for early diagnosis of congestive heart failure using rough set and decision tree approaches", Journal of Biomedical Informatics, Vol.45, PP. 999–1008, 2012.

[4] Hongmei Yan and Jun Zheng, (2008) "Selecting critical clinical features for heart diseases diagnosis with a real-coded genetic algorithm", Journal of Applied Soft Computing, Vol.8, PP.1105-1111, 2008.

[5] Swati Shilaskar, (2013) "Feature selection for medical diagnosis: Evaluation for cardiovascular diseases", Journal of Expert System with Application, Vol.40, PP.4146-4153, 2013.

[6] Petra A. Karsdorp and Merel Kindt, (2009) "False Heart Rate Feedback and the Perception of Heart Symptoms in Patients with Congenital Heart Disease and Anxiety", International Journal of behavioral Medicine, Vol.16, PP.81-88, 2009.

[7] Chih-Lin Chi and W. Nick Street, (2010) "A decision support system for cost-effective diagnosis", Journal of Artificial Intelligence in Medicine, Vol.50, PP. 149-161, 2010.

[8] Ismail Babaoglu and Og˘uz Findik, (2010) "A comparison of feature selection models utilizing binary particle swarm optimization and genetic algorithm in determining coronary artery disease using support vector machine", Journal of Expert System With Applications, Vol.37, PP.3177-3183, 2010.

[9] Evanthia E. Tripoliti and Dimitrios I. Fotiadis, (2012) "Automated Diagnosis of Diseases Based on Classification: Dynamic Determination of the Number of Trees in Random Forests Algorithm", Journal of IEEE Transactions On Information Technology In Biomedicine, Vol. 16, No. 4, July 2012.

[10] Yoon-Joo Park and Se-Hak Chun, (2011), "Cost-sensitive case-based reasoning using a genetic algorithm: Application to medical diagnosis", Journal of Artificial Intelligence in Medicine, Vol.51, PP.133-145, 2011.

[11] Mu-Jung Huang and Mu-Yen Chen, (2007), "Integrating data mining with case-based reasoning for chronic diseases prognosis and diagnosis" Journal of Expert Systems with Applications, Vol. 32, PP.856–867, 2007

[12] Mishra S., Balakrishnan S., Babitha M. (2017) "Coronary Heart Disease Detection from Variation of Speech and Voice." In: Deiva Sundari P., Dash S., Das S., Panigrahi B. (eds) Proceedings of 2nd International Conference on Intelligent Computing and Applications. Advances in Intelligent Systems and Computing, vol 467. Springer, Singapore.

[13] N.A. Setiawan, (2008) " A Comparative Study of Imputation Methods to Predict Missing Attribute Values in Coronary Heart Disease Data Set", Journal in Department of Electrical and Electronic Engineering,Vol.21, PP. 266–269, 2008.

[14] P. Palanikumar, S. Geofrin Shirly and S.Balakrishnan, (2015) "An Effective Two Way Classification of Breast Cancer Images", International Journal of Applied Engineering Research, ISSN 0973-4562, Volume 10, Number 21, pp 42472-42475.

