I.J. Education and Management Engineering, **2016, 6, 45-52**
Published Online November 2016 in MECS (http://www.mecs-press.net)
DOI: 10.5815/ijeme.2016.06.05

Available online at http://www.mecs-press.net/ijeme

Utilization of Data Mining Classification Approach for Disease Prediction: A Survey

Divya Jain [a], Vijendra Singh [b]

[a] *Computer Science and Engineering, The NorthCap University, Gurgaon, 122017, India*
[b] *Computer Science and Engineering, The NorthCap University, Gurgaon, 122017, India*

Abstract

Early diagnosis of a disease is a vital task in medical informatics. Data mining is one of the principal contributors in this discipline. Utilization of Data Mining Technology in Disease Forecasting System is a recognized trend and is successfully emerging in this domain. In today`s world, Heart Disease is the one of the most prevalent disease among people with a high mortality rate. It is essential to classify the reports of heart patients into correct subclasses to lower fatality rate. Over the years, Data mining classification and prediction approaches has been used extensively for disease prediction. This paper comes out with the compilation, analysis as well as comparative study of numerous classification approaches used for predictive analysis of several diseases. The goal of the survey is to provide a comprehensive review of the work done on disease prediction using different classification approaches in data mining.

Index Terms: Data Mining, Classification algorithms, Disease prediction, Healthcare Sector.

1. Introduction

With the rapid accumulation of advanced data mining algorithms and high-throughput technologies, doctors are benefitted extensively in healthcare sector as patient`s records are accessible rapidly in an effective manner. Hospitals maintain a database of patient`s data electronically. A large amount of unstructured, heterogenous data is generated and maintained in a database on a daily basis. We can make data structured using many techniques. This data can made useful using various mining techniques and analyzed to make effective decisions in different situations.

For proper diagnosis of a particular disease, a patient has to undergo several tests in hospitals. In developing countries, this process is more of a time-consuming manual process. As there is a lack of proper medical care and limited access to medical facilities in numerous areas, Disease Control should be prioritized among people. So, there is an essential need to solve this problem and to design a novel data mining technique which is self-

* Corresponding author:
E-mail address:

automated and self-configured having least complexity and better accuracy.

Data mining technology [1] is emerging as a promising field and is used in widespread application areas like ecommerce, bank transactions, microarray gene expression data, scientific experiments etc. This technology blends various analytic methods with advanced and sophisticated algorithms which helps in exploring large volumes of data [2]. It also plays a crucial role in the early detection of diseases. There exist numerous application areas of data mining in medical industry. It is essential that data mining techniques like classification, clustering etc. should be applied to hospital databases so that the right treatments can be provided to patients at the right time which in turn will lower mortality rate.

Classification approach [3,4,5] works by first building a model from training data and then it is applied on testing data for the prediction of unknown data. In healthcare sector, classification and prediction is used predominantly in disease forecasting. There exist numerous techniques for classification of data like KNN, Naïve bayes, support vector machines,decision trees which plays a promising and significant role for the early disease detection.

The remaining portion of the survey is organized as follows. Section 2 presents the Related Work. A summarized conclusion of literature survey and a detailed comparative study is presented in Section 3. Section 4 gives conclusion.

2. Related Work

An intelligent prediction system was proposed in [6] for the diagnosis of heart disease using three commonly used classification approaches - Naïve Bayes, Decision Trees and Neural Network. This intelligent system was scalable, user-friendly, expandablable and was able to answer complex queries effectively. The proposed system discovered hidden information from a historical database of heart disease using various medical factors which can be very helpful in taking clinical decisions and in reducing costs of various treatments. This system can be helpful for the training of medical students and nurses in hospitals for heart disease prediction which can be beneficial in assisting doctors. With the results obtained, it was found that the performance of Naïve Bayes was the best for identification of heart disease compared to Neural Networks and Decision Trees.

Prediction of Kidney Disease was done in Dr. S. Vijayarani et. al. [7] using SVM algorithm and Naïve Bayes approach. Authors tried to classify various stages of Kidney disease through the proposed algorithm called ANFIS. The experiments were conducted in MATLAB. The goal of the research was to find the efficient classification technique through various evaluation measures like accuracy and execution time. While SVM Algorithm gave greater classification accuracy, Naïve Bayes performed better as it executed results in minimum time. The results indicate that SVM overall performs better compared to Naïve Bayes Approach to predict Kidney Disease.

Fuzzy approach using a membership function was applied for the prediction of Heart Disease in V. Krishnaiah et. al. [8]. Authors tried to remove ambiguity and uncertainity of data using Fuzzy KNN Classifier. Dataset containing 550 records was divided into 25 classes, each class consisting of 22 records. Dataset was equally divided into training and testing sets. After applying preprocessing techniques in WEKA, fuzzy KNN approach was implemented. This approach was evaluated through various evaluation measures like accuracy, precision and recall etc. With the results obtained, it was found that that the performance of fuzzy KNN classifier pwas better in comparison to KNN classifier in terms of accuracy.

A novel approach was developed in [9] using ANN algorithm for the prediction of heart disease. The researchers developed an interactive prediction system using the classification through artificial neural network algorithm with the consideration of 13 most important clinical factors. The proposed approach was very effective and user friendly for heart disease prediction with 80% accuracy and can be of great used for healthcare preofessionals.

An efficient prediction system was designed in [10] to predict the risk level of heart patients. The system could discover rules efficiently from the dataset using decision tree approach according to the given parameter related to patient`s health. Authors concluded that the system can predict the risk level of heart disease risk

level to a great extent.

A useful system was presented for the prediction of heart attacks [11]. The prediction system was developed with the inclusion of classification and clustering techniques for predicting risk level of heart attacks.

Three classification based approaches were applied on healthcare data for the diagnosis of heart disease [12]. The approaches used were KNN, Naïve Bayes and C4.5 Algorithm. The experiments were conducted on the heart disease data set using WEKA tool to find the best technique for the prediction of heart disease using various evaluation techniques like sensitivity, specificity, accuracy, error rates etc. With the results obtained, it was found that KNN performed best in terms of accuracy and C4.5 Algorithm works best for the purpose of prediction.

A prediction model was proposed for the prediction of Alzheimer Disease using decision tree approach [13]. Authors considered five major risk factors related to Alzheimer's disease. In this research, the decision tree induction used Entropy or Information Gain as a measure for predicting Alzheimer's disease in patients. The model can be of great help to healthcare professionals for determining the status of this disease.

The researchers focused on different classification techniques with their merits and demerits used for the heart disease prediction [14].

An automated system to answer complex queries for heart disease predicition was proposed in [15]. This intelligent System was implemented using Naive Bayes approach in Java platform. The system was designed to give fast,better and accurate results. It could help medical practictioners in taking clinical decisions related to heart attacks. This system can be expanded by incorporating SMS facility, designing Andriod and IOS mobile applications, addition of pacemaker in the system.

An effective diabetes mellitus prediction system using decision tree approach was designed in [16] for predicting the risk level of diabetic patients. The results were evaluated with 2 classifiers namely C4.5 algorithm and patial trees. With the results obtained, it was found that C4.5 algorithm performed better with 81.27 % accuracy.

The researchers experimented the application of different classification techniques and developed models to diagnose heart attacks [17]. Researchers also did comparison of these models to find out which model is better for the prediction of heart attacks and can be very helpful to handle complex queries related to heart attacks.

An intelligent system using Naïve Bayes Approach and K-means clustering was proposed to predict heart disease [18]. While clustering was used for grouping of attributes and for increasing efficiency of results, Naïve Bayes approach was used for heart disease prediction.

3. Comparative Study

This section presents summarization of the literature survey in two different tables. Table 1 is more specifically concerened with the utility driven from the paper and gives a scope for further research work. Table 2 gives an analysis on the application of classification technique applied on various datasets.

Table 1. Summarized Conclusion of Literature Survey

Disease/Dataset Used/Data Source	Work Done	Utility/ Conclusion	Future Scope
Cleveland Heart Disease database	The researchers proposed an intelligent, scalable, user-friendly prediction system to answer complex queries effectively for the identification of heart disease using Neural Network, Naïve Bayes and Decision Trees and [6].	a) The results indicated that Naïve Bayes Model was the best to answer complex queries related to heart disease. b) This system can be helpful for the training of medical students and nurses in hospitals for heart disease prediction which can be beneficial in assisting doctors.	a)The system worked only with categorical data. In future, we can also work with continuous data b) Size of dataset can be increased as well as system can be made to work with more attributes. c) Additionly, system should be tested from cardiologists before being deployed in hospitals.

Synthetic Kidney Function Test (KFT) Dataset	Authors classified four types of kidney diseases using SVM and Na ïve Bayes Approach [7].	The results indicated that SVM overall performed better to predict Kidney Disease as it was better in giving accurate prediction while Na ïve Bayes performed better as it executed results in minimum time.	Using the proposed approach, we can work on the prediction of other treacherous diseases. Also we can add the application of other mining techniques to get better results.
Statlog Heart disease dataset and Cleveland Heart Disease dataset	Prediction of Heart disease was done with the inclusion of Fuzzy approach using a membership function in [8]. Fuzzy KNN Classifier was applied to remove ambiguity and uncertainity of data.	The proposed fuzzy KNN approach removed the redundancy and ambiguity of the data and provided better accuracy in comparison to KNN classifier.	Can further expand the system by incorporating more attributes in the current system.
Heart disease data from UCI repository	The researchers proposed a novel approach using Artificial Neural Network algorithm for heart disease prediction system [9].	The resulted prediction system was 80% accurate which can be very effective in helping healthcare professionals to predict the status of heart disease.	The prediction system can be improved with the consideration of more risky factors for heart disease.
a) V.A. Medical Center and Cleveland Clinic Foundation b) Donor	The researchers developed an effective prediction system using decision tree approach to predict the risk level of heart patients [10].	This system can be used for predicting risk level of heart patients to a great extent which can be very helpful for medical practitioners in taking effective clinical decisions.	The resulted system can incorporate more data mining techniques to make the system more efficient.
Heart Data Set from the UCI Learning Repository	Three classification based approaches - KNN, Na ïve Bayes and C4.5 algortihm were applied on the heart data set for the identification of heart disease [12].	The results concluded that decision tree approach(C4.5 Algorithm) worked best for predition while KNN performed best in terms of accuracy for the diagnosis of heart disease.	In future, other classification approaches like SVM Algorithm can be applied on the dataset to get better results.
Clinical Data	Authors proposed a prediction model using decision tree induction towards Alzheimer's disease prediction [13].	The model can be of great help for healthcare professionals to determine the status of Alzheimer's disease.	Can improve this prediction model by considering more risk factors related to Alzheimer's disease.
Clinical Data	The researchers proposed an intelligent and automated prediction system to answer complex queries for the diagnosis of heart disease [15].	The proposed system can assist medical professionals to make clinical decisions by answering complex queries related to heart.	Can work on developing mobile apps on IOS and Andriod for better availability of system. Furthurmore, SMS facility can also be added. Also we can add pacemaker to the system for better functioning.
Pima Indians Diabetes Database	Authors designed a prediction system for diabetes mellitus using decision tree approach for predicting risk level of diabetic patients [16].	The proposed system was 81.27 % accurate with the usage of C4.5 algorithm which can be beneficial for medicial professionals.	In future, we can work with other classifiers to get better prediction with more accurate results.
Data given by medical practitioners of South Africa	The researchers developed models using various classification approaches to diagnose heart attacks. [17].	The model can be very helpful for the physians in risky cases to handle complex queries related to heart attacks.	Can work on building the resulted predicitive model more accurate and efficient using the application of other techniques.

Table 2. Summarized Objectives of Related Work Done by Different Authors

Author	Title	Year	Data Source/ Dataset Used	Classification Techniques Used	Disease Examined	Objectives
Ritika Chadha et. al. [22]	Application of Data Mining Techniques on Heart Disease Prediction: A Survey	2016	N/A	Decision Trees, Genetic Algorithm ,Naïve Bayes and Neural Network	Heart Disease	To analyze different mining techniques that have been implemented in the recent years for identification of heart disease.
Dr. S. Vijayarani et. al. [7]	Data Mining Classification Algorithms For Kidney Disease Prediction	2015	Synthetic Kidney Function Test (KFT) Dataset	Naïve Bayes and SVM Algorithm	Kidney Disease	To predict kidney disease using various classification approaches and finding the efficient classification algorithm
V. Krishnaiah et. al. [8]	Heart Disease Prediction System Using Data Mining Technique by Fuzzy K-NN Approach'	2015	Statlog Heart disease database and Cleveland Heart disease database	Fuzzy K-NN classifier	Heart Disease	To remove the uncertainty and ambiguity of data using Fuzzy KNN classifier for the identification of heart disease
Purushottam et. al. [10]	Efficient Heart Disease Prediction System using Decision Tree	2015	V.A. Medical Center, Long Beach and Cleveland Clinic b) Donor	Decision Trees (C4.5 Algorithm)	Heart Disease	To design an efficient heart disease prediction system for predicting the risk level of heart patients.
Sujata Joshi et. al. [12]	Prediction of Heart Disease Using Classification Based Data Mining Techniques	2015	Heart data from the UCI Repository	Decision Trees, KNN and Naïve Bayes	Heart Disease	To diagnose the occurrence of heart disease using classification approaches.
Monika Gandhi et. al. [14]	Predictions in Heart Disease Using Techniques of Data Mining	2015	N/A	Naïve Bayes, Neural Network and Decision Trees,	Heart Disease	To study different classification techniques for heart disease prediction.
Sana Shaikh et. al. [15]	Electronic Recording System - Heart Disease Prediction System	2015	N/A	Naïve Bayes and Decision Trees	Heart Disease	To propose an automated system to answer complex queries for the diagnosis of heart disease.
Purushottam et. al. [16]	Diabetes Mellitus Prediction System Evaluation Using C4.5 Rules and Partial Tree	2015	Pima Indians Diabetes Database	Partial Tree and C4.5 Algorithm	Diabetes Mellitus	To predict risk level of diabetes mellitus using Partial Tree and C4.5 Algorithm.
Rucha Shinde et. al. [19]	An Intelligent Heart Disease Prediction System Using K-Means Clustering and Naïve Bayes Algorithm	2015	N/A	K-Means Clustering and Naïve Bayes Algorithm	Heart Disease	To Propose and implement heart disease prediction system with the inclusion of Naïve Bayes classifier and K-Means Clustering
Dr. S. Vijayarani et. al. [20]	Kidney Disease Prediction Using SVM And ANN Algorithms	2015	Synthetic kidney function test (KFT) dataset	SVM and ANN Algorithm	Kidney Disease	To classify the kidney Disease into four types using SVM and ANN Algorithm.
M.A.Nishara Banu. et. al.[11]	Disease Forecasting System Using Data Mining Methods	2014	Cleveland Heart disease dataset	C4.5 Algorithm	Heart Disease	To design a useful system for the prediction of heart attacks.

Dana AL-Dlaeen et. al. [13]	Using Decision Tree Classification to Assist in the Prediction of Alzheimer's Disease	2014	Clinical Data	Decision Trees	Alzheimer's Disease	To develop an effective model for the prediction of Alzheimer`s disease.
Hlaudi Daniel Masethe et. al. [17]	Prediction of Heart Disease using Classification Algorithms	2014	Data given by medical practitioners of South Africa	C4.5, Naïve Bayes, REPTREE, CART, and Bayes Net	Heart Disease	To develop different models using various classification techniques to diagnose heart attacks.
Rashedur M. Rahman et. al. [21]	Comparison Of Various Classification Techniques Using Different Data Mining Tools For Dia betes Diagnosis	2013	Pima Indian Diabetes Data (PIDD) set	Fuzzy Logic, Decision Tree and Neural Network.	Diabetes	To analyze the performance of different classification techniques on a large dataset
AH Chen et. al.[9]	HDPS: Heart Disease Prediction System	2011	Heart data from UCI repository	Artificial Neural Network Algorithm	Heart Disease	To develop a novel user friendly prediction system for heart disease using ANN algorithm
Jyoti Soni et. al. [18]	Intelligent and Effective Heart Disease Prediction System using Weighted Associative Classifiers	2011	Heart Data Set from the UCI Repository	Weighted Associative Classifier (WAC)	Heart disease	To develop an intelligent prediction system for the prediction of heart attacks.
Sellappan Palaniappan et. al. [6]	Intelligent Heart Disease Prediction System Using Data Mining Techniques	2008	Cleveland Heart disease database	Naïve Bayes, Neural Network and Decision Trees	Heart Disease	To design a user friendly, intelligent System for the prediction of heart disease using different classification approaches.

4. Conclusion

Early Disease Prediction is a major challenge in the healthcare sector. Over the last few years, a lot of work has been done in the predictive analysis of diseases using numerous classification approaches. Data mining classification approaches have been utilized extensively for disease prediction. Each approach has its own merits and demerits but Naïve Bayes Approach and the C4.5 Algorithm are found to be the most promising techniques for the diagnosis and prediction of numerous medical diseases in less time with high accuracy and least complexity.

References

[1] Ian H. Witten and Eibe Frank, "Data Mining: Practical machine learning tools and techniques". Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2nd edition.

[2] D. T. Larose, *Discovering Knowledge in Data: An Introduction to Data Mining.* ISBN 0-471-66657-2, John Wiley & Sons, Inc., 2005.

[3] P.N. Tan, M Steinbach, V. Kumar, *Introduction to Data Mining*. 4th edn. Pearson Publications, Boston.

[4] J. Han, M. Kamber, *Data Mining: Concepts And Techniques*. Morgan Kaufmann, San Francisco (2001).

[5] M. H. Dunham, S. Sridhar, *Data Mining: Introductory and Advanced Topics*, Pearson Education, New Delhi, ISBN: 81-7758-785-4, 1st Edition, 2006.

[6] S. Palaniappan., R. Awang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques", IJCSNS International Journal of Computer Science and Network Security 8(8) (August 2008).

[7] S. Vijayarani, S. Dhayanand ,"Data Mining Classification Algorithms for Kidney Disease Prediction", International Journal on Cybernetics & Informatics (IJCI) Vol. 4, No. 4, August 2015 DOI: 10.5121/ijci.2015.4402 13.

[8] V. Krishnaiah, G. Narsimhaand N. Subhash Chandra, "Heart Disease Prediction System Using Data Mining Technique by Fuzzy K-NN Approach", Emerging ICT for Bridging the Future – Volume 1,Advances in Intelligent Systems and Computing 337, DOI: 10.1007/978-3-319-13728-5_42.

[9] AH Chen, SY Huang, PS Hong, CH Cheng, EJ Lin,"HDPS: Heart Disease Prediction System", Computing in Cardiology 2011;38:557-560.

[10] Purushottam**,** Kanak Saxena and Richa Sharma, "Efficient Heart Disease Prediction System using Decision Tree", International Conference on Computing, Communication and Automation (ICCCA2015).

[11] M. A. Nishara Banu**,** B. Gomathy ,"Disease Forecasting System Using Data Mining Methods", 2014 International Conference on Intelligent Computing Applications.

[12] Sujata Joshi and Mydhili K. Nair, "Prediction of Heart Disease Using Classification Based Data Mining Techniques", Computational Intelligence in Data Mining - Volume 2, Smart Innovation, Systems and Technologies 32, DOI 10.1007/978-81-322-2208-8_46.

[13] Dana AL-Dlaeen and Abdallah Alashqur, "Using Decision Tree Classification to Assist in the Prediction of Alzheimer's Disease", 2014 6th International Conference on CSIT ISBN:987-1-4799-3999-2.

[14] Monika Gandhi and Shailendra Narayan Singh, "Predictions in Heart Disease Using Techniques of Data Mining", 2015 1st International Conference on Futuristic trend in Computational Analysis and Knowledge Management (ABLAZE-2015).

[15] Sana Shaikh, Amit Sawant, Shreerang Paradkar and Kedar Patil, "Electronic Recording System - Heart Disease Prediction System", 2015 International Conference on Technologies for Sustainable Development (ICTSD-2015), Feb. 04 – 06, 2015, Mumbai, India.

[16] Purushottam**,** Kanak Saxena and Richa Sharma, "Diabetes Mellitus Prediction System Evaluation Using C4.5 Rules and Partial Tree", 2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions).

[17] Hlaudi Daniel Masethe, Mosima Anna Masethe, "Prediction of Heart Disease using Classification Algorithms", Proceedings of the World Congress on Engineering and Computer Science 2014 Vol II WCECS 2014, 22-24 October, 2014, San Francisco, USA.

[18] Jyoti Soni, Uzma Ansari and Dipesh Sharma, "Intelligent and Effective Heart Disease Prediction System using Weighted Associative Classifiers", International Journal on Computer Science and Engineering (IJCSE), Vol. 3 No. 6 June 2011.

[19] Rucha Shinde, Sandhya Arjun, Priyanka Patil and Jaishree Waghmare "An Intelligent Heart Disease Prediction System Using K-Means Clustering and Naïve Bayes Algorithm",(IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (1) , 2015, 637-639.

[20] S. Vijayarani and S.Dhayanand ,"Kidney Disease Prediction Using Svm And Ann Algorithms", International Journal of Computing and Business Research (IJCBR), ISSN (Online) : 2229-6166, Volume 6 Issue 2 March 2015.

[21] Rashedur M. Rahman and Farhana Afroz , "Comparison Of Various Classification Techniques Using Different Data Mining Tools For Diabetes Diagnosis", Journal of Software Engineering and Applications, 2013, 6, 85-97.

[22] Ritika Chadha, Shubhankar Mayank, Anurag Vardhan and Tribikram Pradhan, "Application of Data Mining Techniques on Heart Disease Prediction: A Survey", Emerging Research in Computing, Information, Communication and Applications, DOI 10.1007/978-81-322-2553-9_38.

Authors' Profiles

Divya Jain is pursuing PhD from NorthCap University and is M.Tech. holder in Computer Science with First Division from the NorthCap University, Gurgaon, Haryana, India. Her current research interests include Data Mining classification algorithms and clustering algorithms. Divya Jain received the B.Tech in CSE with Honors from Maharishi Dayanand University, Rohtak in 2012. She is author of one book and four research papers in International Journals/Publishing House.

Singh Vijendra received his PhD degree in Engineering and M Tech degree in Computer Science and Engineering from Birla Institute of Technology, Mesra, Ranchi, India. He is currently working as an Associate Professor in the department of computer science and engineering at The NorthCap University, Gurgaon, India. Dr. Singh major research concentration has been in the areas of Data Mining, Pattern Recognition, Image Processing, Big Data and Soft Computation. He has more than 30 scientific papers in this domain. Singh Vijendra served as Editor of the International Journal of Multivariate Data Analysis, Inderscience, UK; International Journal of Internet of Things and Cyber-Assurance, Inderscience, UK; BMC Bioinformatics, Springer; Journal of Next Generation Information Technology, Korea; International Journal of Intelligent Information Processing, Korea; Research Journal of Information Technology, USA and Lead Guest Editor, Computational Intelligence in Data Science and Big Data, USA. He is a reviewer of Springer and Elsevier journals. He is a member of programme committee and technical committee of over 30 international conferences including: (SCDS2015), Malaysia; 2015 International Conference on Data Mining (DMIN15), Las Vagas, USA; (CISIA2015), Bangkok, Thailand; (ETCA2015), Beijing, China; (CIS 2015), Beijing, China; ENCINS' 2015, Casablanca, Morocco; ICCVIA, 2015, Sousse, Tunisia and eQeSS 2015, Dubai; DMIN14, USA; DMIN13, USA; DMIN12, USA.

How to cite this paper: Divya Jain, Vijendra Singh,"Utilization of Data Mining Classification Approach for Disease Prediction: A Survey", International Journal of Education and Management Engineering(IJEME), Vol.6, No.6, pp.45-52, 2016.DOI: 10.5815/ijeme.2016.06.05