Heart Diseases Prediction using Data Mining and its Techniques- A Survey

Muhammad Fahad [1] Sadiq Ur Rehman[2] Aqeel-ur-Rehman[3] Muhammad Kashif Alam[4]

Abstract

A process or way toward analyzing patterns of data as indicated by different points of view for classification into meaningful data, which is gathered and amassed in likely manner, e.g. data-warehouse for effective analysis, data mining algorithms, enabling business decisions to cut expenses and increase income. Areas including business, retail system, medical, sciences and engineering are indicating the worth of data mining. In this paper, ways to predict heart diseases using different algorithms/techniques are presented. To explore areas of data mining in health care is the key objective of this research. Medical industry is capable to produce data of different types i.e. non-real time or real-time and the amount of such data is increasing day by day. Due to the daily increment of medical data, medical industry is capable to provide huge contribution in the area of data mining which in result gives prediction of diseases and improves quality of services to the patients. This paper shows the combination and analysis of neural network and data mining, Fuzzy and genetic algorithm, data mining and machine learning.

Keywords: Data mining, Data analysis, Naïve Bayes, Heart disease, Data mining algorithms, Neural Networks, Decision Tree, Fuzzy- Logic, Machine Learning, Data mining.

1 Introduction

Data mining is one of the processes in which non-trivial data is being extracted. Data mining is a technique through which data can be gathered for further processing which is called knowledge/information. This technique has a key role in diseases prediction. There are multiple diseases which we can predict through data mining techniques, cancer, heart diseases, etc. are the major diseases which we can predict through data mining. In health industry data mining is very important. Since health industry has a lot of complex data for processing, this data can be in the form of hospital records, digital devices data, survey data of medical students, electronic gadget data of medical devices etc. The category of complex data used in health industry can be either real-time or non-real time. However, in both the cases error free data is of high importance as this data can be used in correct diagnosis and efficient treatment. Bad diagnosis is not acceptable in health care industry as it may result in death or big health hazard. Some of the main focused technologies of data mining are databank technology, machine learning and statistical analysis [1].

Heart is one of the mains of body [2]. Proper working or human body is mainly dependent on proper working of human heart. If due to any reason, working of heart gets disturbed then

[1] *Hamdard Institute of Engineering and Technology, Karachi | mfahad@hamdard.edu.pk*
[2] *Hamdard Institute of Engineering and Technology, Karachi | sadiqsr@gmail.com*
[3] *Hamdard Institute of Engineering and Technology, Karachi | aqeel.rehman@hamdard.edu.pk*
[4] *Hamdard Institute of Engineering and Technology, Karachi | kashif.alam@hamdard.edu.pk*

certain diseases may arise in human body.

Cardiovascular disease also known as heart disease is now a days most common disease in human body [3]. According to the statistics of 2012 world health report [4], there is a raised blood pressure complain in every one in three patients. Moreover, if we consider WHO (World Health Organization), around 17 million deaths were reported due to heart attack. There are various reasons which increase the probability of heart diseases, some of which are;

- Work load
- Mental stress
- Hypertension
- Physical inactivity
- Obesity
- Uncontrolled cholesterol
- Smoking
- Poor diet
- Blood pressure issue
- Genetic susceptibility to heart diseases

This paper is divided into five sections, Section I is related to the introduction of the topic. Details regarding heart diseases algorithms can be found in section II. Discussion about prediction can be found in Section III. Open source tools that are used in data mining are discussed in Section IV. Section V presents conclusion and future work.

2 Data Mining Techniques and Algorithms in Health Care

There are multiple techniques available to identify diseases of heart with the help of data mining. It includes classification, clustering, association, prediction etc. Classification is the machine learning technique and is responsible to categorize individual items into predefined groups. Statistics, decision trees, neural networks etc. Clustering is very beneficial to the group of substance having similar properties/features with the help of mechanical technique. Association is a technique which is considered to be the best data mining technique for the predication of heart diseases so far. In this technique, all the non-similar attributes that have been used for analyzing the heart disease are incorporated and patients with complete risk factors (important for prediction of disease) are sorted out. Prediction is a technique used in data mining to find the correlation between independent and dependent variables. Data mining algorithms and techniques are classified into the following sub categories,

A. Neural Network
B. Naïve Bayes
C. Decision Tree
D. Genetic Algorithm

Each category has different accuracy rates among which Neural Networks were found to be the most accurate classification technique having accuracy rate of 100%.

Intelligent Heart Disease Prediction System (IHDPS) which uses above mentioned data mining techniques can be seen in [5].

A Neural network

An artificial neural network (ANN) generally called as neural network (NN) is a model of mathematics based on biological NN. ANN works same like human brain. Neural network is composed of many parallel working nodes which are joint together with one-directional signal connections. Supervised learning and unsupervised learning are the two main categories of neural network

B Naïve Bayesian

Based on Bayes theorem, Naïve Bayes [6] is an algorithm that is used for classification. The key concept behind Bayes theorem is probability. In this theorem, probability of an on-going event is calculated given the probability of already occurred event.

C Decision Tree

Decision tree, as from its name, is a structure of large data set that gets divided into successive small data sets with the implementation of a sequence of decision rules. As increment takes place in successive division. The outcome of result will gets closer to other members of the set. There are various models of decision tree. Gain ratio decision tree is the most successful type of decision tree [7].

D Genetic Algorithm

In genetic algorithm, process of natural selection takes place by search method. This algorithm is responsible to provide optimization and solution to search problem by using advance techniques which includes mutation, crossover, inheritance and selection.

It would be a very powerful mechanism for efficient classification if genetic algorithm and fuzzy logic gets combined. Genetic algorithm helps in effectiveness while fuzzy logic helps to develop knowledge based system in health disease

Table 1: Data Mining Techniques with their accuracy rate. [18,20,30]

Reference	Classification Techniques	Accuracy	Recommended Data mining technique
Ahmed F et al. (2015)	Bayes Net	84.5%	SVM
	SVM(Support Vector Machines)	85.1%	
	FT(Functional Trees)	84.5%	
Salha M et al. (2014)	Neural Networks	91%	Decision Tree
	Decision Tree	99%	
	Naïve Bayes	96.5%	
Sivagowry et al. (2014)	Neural Networks	98%	Neural Networks
	Decision Tree	52%	
	Naïve Bayes	52.33%	
Rashe-Dur et al. (2013)	Decision Tree	75.5%	Fuzzy Logic
	Neural Network	79.19%	
	Fuzzy Logic	83.85%	
Indira S. FalDessai (2013)	BNN	80.4%	PNN
	PNN	94.6%	
	NB	84%	
	DT	84.2%	
Apte et al. (2012)	Naive Bayes	90.74%	Neural Networks
	Decision Trees	99.62%	
	Neural Networks	100%	
Jyoti K et al. (2012)	Neural Networks	100%	Neural Networks
	Decision Tree	99.62%	
	Naïve Bayes	90.74%	
Nidhi et al. (2012)	Naive Bayes	90.74%	Neural Networks
	Decision Trees	52.33%	
	Neural Networks	96.5%	
Chaitrali et al. (2012)	Neural networks	100%	Neural Networks
	Naive Bayes	90.74%	
	Decision Tree	99.62%	

Resul et al. (2009)	Neural networks	89.01%	Neural Networks
Resul et al (2009)	Neural networks	97.4%	Neural Networks
M. Anbarasi et al. (1999)	Classification through clustering Naive Bayes	88.3%	Decision Tree
	Decision Tree	99.2%	
	Naive Bayes	96.5%	
Matjaz et al. (1999)	Neural Networks	85%	Neural Networks
	ECG(Neural Networks)	74%	

Above mentioned table shows the comparison of different data mining techniques in which accuracy is measured. Different authors used different data sets for comparison with different techniques. After implementing techniques and algorithms the next phase is prediction that is discussed in the next section.

3 Prediction of Heart Diseases Using Data Mining

Many papers have been written related to heart diseases and data mining techniques for prediction of heart diseases. Different techniques like classification, dataset, algorithms are used to observe and show result which are efficient methods.

P.K Anooj [14], [17] has proposed CDSS for prediction and diagnosis of heart diseases based on fuzzy rules. The proposed system has two parts one was computerized and the other one was generalized. The process is simple it takes patient data automatically, and implements both phases. Proposed system is better than other systems. Result is better when applied fuzzy rules. Latha Parhiban [13], [14] also formulated the approach using co active nuero-fuzzy inference system (CANFIS).

Subbalakshmi, Ramesh and Chinarao [14], [15] proposed system which takes age, sex, blood pressure, cholesterol etc and other attributes as an input and then shows the result and predicts heart diseases. This model was good and predicts even complex queries with effective results. The system provides decision support in heart diseases using naïve Bayes data mining techniques.

Some of the solutions for heart diseases prediction are presented above which are based on Neural networks, Fuzzy logic, Decision Tree etc. Few open source tools that are available for data mining are discussed in detail in the next Section.

4 Open Source Tools for Data Mining Used In Health Care Applications

There are many available open source tools that are frequently being used in data mining especially in the case of heath care applications. Some of the most common open source tools are as follows,

A. Tanagra

It is also an open source software that is used for the purpose of academics and research. It covers that data mining concepts from several dimensions which include machine learning, exploratory data analysis, meta supervised learning, feature selection, database area etc. Tanagra is the graphical user interface based data mining tool [33]. Tanagra can be used to analyze both types of data (i.e. synthetic or real).

B. WEKA Tool

Weka [34] is a data mining software that has been developed in java language by University of Waikato (New Zealand). It is a set machine learning algorithm that is used for data mining. The beauty of algorithms used in Weka is that they are independent (can directly be practiced on data set or personal java code). Weka consists of specialized tools that are used for pre-processing of data, visualization, clustering, regression etc. Since Weka is an open source software, it facilitates developers to create new machine language techniques and also applications that are required to solve the data mining problems. The biggest plus point of Weka is the capability to be applied on big data. File format used in Weka is ARFF.

C. MATLAB

MATLAB [35] is highly recommended for the fast computation, visualization and for coding. Matlab is a GUI based software in which we can perform several complex tasks in efficient way. Matlab helps us in performing analysis of data, creating/ modifying complex algorithms, developing applications etc. The computation time of Matlab is faster than the computation time in C/C++ and other programming languages.

D. ORANGE

Orange [36], an open source machine learning and data mining suit used for the data analysis. It is a data mining tool in python. The use of Orange is very much simple, it can be used by professional programmers and also by beginners/ students who are working in the field of data mining. Library in this software is classified in hierarchical structure for the components of data mining.

E. Rapid Miner

Rapid Miner [37], is the top most open source software that is used for data mining. Rapid miner is most powerful and useful software for data mining. By using this software, data mining

as well as analysis can be done with integration of two products or a single product. It is used for data analyzing which includes environment for analytics, mining, deep learning, machine learning etc. Rapid miner shows result through visualization, optimization and validation. It uses client server model. It is very helpful and useful in data mining which optimizes, validates and visualizes results. There are various types of graphs which rapid miner shows after analyzing and mining which includes Pie Charts, Contour, 3-D, Density, Histogram, Survey Plots, and Quartiles etc.

Table 2: Open Source Tools Technical Overview [39,40,41,42,43]

S.No	Name	Version	OS Support	Reference
1	Tanagara	1.4.50	Windows	https://eric.univ-lyon2.fr/~ricco/tanagra/index.html
2	Weka	3.9	Cross Platform	https://www.cs.waikato.ac.nz/ml/weka/
3	MATLAB	9.0	Cross Platform	https://www.mathworks.com/
4	Orange	3.7	Cross Platform	www.orange.biolab.si
5	Rapid Miner	7.6	Cross Platform	www.rapidminer.com

Table 3: Open Source Tools Functions and Characteristics [38]

S.No	Name	Functions	Properties/Charactistics
1	Weka	Machine Learning	Visualization Classification Regression
2	MATLAB	Analysis, Computation and Visualization	Data Analysis Statics Integration
3	Orange	Machine Learning and Data mining/visualization	Interaction Visualization
4	Rapid Miner	Analysis and Data Mining	Shows real time macro values
5	Tanagara	Statistical Learning, Data Analysis and Machine Learning	Drag and drop Different Controls for functions.

5 Conclusion and Future Work

In this paper, each data mining technique had been shown with the result separately. Research also showed that when different techniques are used to predict heart diseases, there are some differences (in accuracy) in it. Same data and classifications show different result in different data mining techniques as discussed in table 1. The main purpose of the survey was to study and analyze different data mining techniques that are used to predict heart diseases. Result shows that Neural Network method provides more accuracy in different scenarios.

Text mining can also be performed on different data sets. Since, there is huge amount of data which is unstructured and we can utilize this data and apply above techniques an it.

References

[1] Singh, Kuldip, and Gurpreet Singh. "Alterations in Some Oxidative Stress Markers in Diabetic Nephropathy." Journal of Cardiovascular Disease Research 8, no. 1 (2017).

[2] Nagre, SurajWasudeo. "Mobile Left Atrial Mass–Clot or Left Atrial Myxoma." Journal of Cardiovascular Disease Research 8, no. 1 (2017).

[3] Bhatla, Nidhi, and KiranJyoti. "An analysis of heart disease prediction using different data mining techniques." International Journal of Engineering 1, no. 8 (2012): 1-4.

[4] Berenson, Gerald S., Sathanur R. Srinivasan, WeihangBao, William P. Newman, Richard E. Tracy, and Wendy A. Wattigney. "Association between multiple cardiovascular risk factors and atherosclerosis in children and young adults." New England journal of medicine 338, no. 23 (1998): 1650-1656.

[5] Palaniappan, Sellappan, and RafiahAwang. "Intelligent heart disease prediction system using data mining techniques." In Computer Systems and Applications, 2008. AICCSA 2008. IEEE/ACS International Conference on, pp. 108-115. IEEE, 2008.

[6] Ratnam, D., P. Himabindu, V. MallikSai, SP Rama Devi, and P. RaghavendraRao. "Computer-Based Clinical Decision Support System for Prediction of Heart Diseases Using Naïve Bayes Algorithm." Int. J. Comput. Sci. Inf. Technol. 5, no. 2 (2014): 2384-2388.

[7] Quilan, J. R. "Decision trees and multi-valued attributes." In Machine intelligence 11, pp. 305-318. Oxford University Press, Inc., 1988.

[8] JyotiKiran and BhatlaNidhi "An Analysis of Heart Disease Prediction using Different Data Mining Techniques" International Journal of Engineering Research & Technology (IJERT). ISSN 27278-0181, Vol. 1 Issue 8, October 2012.

[9] Dangare S. Chaitrali and S. ApteSulabha "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques" International Journal of Computer Applications (0975-888) Vloum 47-No. 10 June 2012.

[10] M. AlzahaniSalha, AlthopityAfnan, AlghamdiAshwag, AlshehriBoushra and AljuaidSuheer "An overview of DataMining Techniques Applied for Heart Diseases Diagnosis and Prediction" lecture Notes on Information Theory Vol. 2 No 4, December 2014.

[11] NikharSonam and Karandikar A.M "Prediction of heart diseases using Data Mining

Techniques – A Review" International Research Journal of Engineering and Technology (IRJET) Volume 3 Issue : 2 Feb 2016.

[12] Singh Williamjeet and KaurBeant "Review on Heart Disease Prediction system using data Mining Techniques" International Journal on Recent and Innovation Trends in Computing and Communication Volume : 02 issue : 10 October 2014 ISSN : 2321-8169.

[13] OyedotunKayode and Olaniyi Ebenezer "I.J. Intelligent Systems and Applications" Published online November 2015 MECS.

[14] Dubey Rishi and ChandarakarSantosh "Review on Hybrid Data Mining Techniques for the Diagnosis of Heart Diseases in Medical Ground" Indian Journal of Applied Research Volume: 5, Issue: 8 August 2015, ISSN 2249-555X.

[15] S. Sivagowry, M Durairaj and A Persia "An Empirical Study on applying Data mining Techniques for the Analysis and Prediction of Heart Disease"

[16] V. Manikantan and S. Latha, "Predicting the analysis of heart disease symptoms using medicinal data mining methods", International Journal of Advanced Computer Theory and Engineering, vol. 2, pp.46-51, 2013. W.-K. Chen, Linear Networks and Systems (Book style). Belmont, CA: Wadsworth, 1993, pp. 123–135.

[17] SellappanPalaniappan and RafiahAwang, "Intelligent heart disease prediction system using data mining techniques", International Journal of Computer Science and Network Security, vol.8, no.8, pp. 343-350,2008.

[18] K.Srinivas, Dr.G.Ragavendra and Dr. A. Govardhan, "A Survey on prediction of heart morbidity using data mining techniques", International Journal of Data Mining & Knowledge Management Process (IJDKP) vol.1, no.3, pp.14-34, May 2011.

[19] G.Subbalakshmi, K.Ramesh and N.ChinnaRao, "Decision support in heart disease prediction system using Naïve Bayes", ISSN: 0976-5166, vol. 2, no. 2.pp.170-176, 2011.

[20] K.Srinivas , B.Kavihta Rani, Dr. A.Govrdhan , Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks , IJCSE Vol. 02, No. 02, 2010, 250-255.

[21] D. P. Shukla, ShamsherBahadur Patel and Ashish Kumar Sen, "A literature review in health informatics using data mining techniques," International Journal of Software & Hardware Research in Engineering, vol. 2, no. 2, February 2014.

[22] D. P. Shukla, ShamsherBahadur Patel and Ashish Kumar Sen, "A literature review in health informatics using data mining techniques," International Journal of Software & Hardware Research in Engineering, vol. 2, no. 2, February 2014.

[23] Promad Kumar Yadav, K. L. Jaiswal, ShamsherBahadur Patel, D. P. Shukla, "Intelligent heart disease prediction model using classification algorithms," UCSMC, vol. 3, no. 08, pp. 102-107, August 2013.

[24] R.Chitra and V. Seenivasagam, "Review of Heart Disease Prediction System Using Data Mining and Hybrid Intelligent Techniques", ICTACT Journal on Soft Computing, vol.3, pp. 605–609, July 2013.

[25] JyotiSoni, "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", International Journal of Computer Applications, vol.17, pp. 43–48, Mar. 2011.

[26] S. Vijayarani and S. Sudha, "An Efficient Clustering Algorithm for Predicting Diseases from Hemogram Blood Test Samples", Indian Journal of Science and Technology, vol.8, pp. 1–8, Aug. 2015.

[27] B.Venkatalakshmi, M.V Shivsankar, "Heart Disease Diagnosis Using Predictive Data Mining", International Journal of Innovative Research in Science, Engineering and Technology, vol.3, pp. 1873–1877, Mar. 2014.

[28] Sultana Marjia, Haider Afrin and ShorifUdding Mohammad "Analysis of Data Mining Techniques for Heart Disease Preditction" iCEEiCT 2016.

[29] Govardhan A., RoaTaghavendra, Srinivas K. "Analysis of Conronary Heart Diseases and Prediction of Heart Attack in Coal Mining Regions Using Data Mining Techniques" The 5th International Conference on Computer Science &Education Hefei, China August 24-27, 2010

[30] TanejaAbhishek "Heart Disease Prediction System Using Data mining Techniques" Oriental journal of Computer Science and Technology ISSN: 0974-6471 Vol 6 No. 4. Dec 2013.

[31] ChandnaDeepali "Diagnosis of Heart Disease Using Data mining Algorithm" International Journal of Computer Science and Information Technologies Vol. 5(2), 2014. (IJCSIT)

[32] Rani Usha K. "Analysis of Heart Disease Dataset using neural Network approach" International journal of Data mining & knowledge manangement process Vol 1, No.5, September 2011.

[33] Krishnaiah, V., G. Narsimha, and N. Subhash Chandra. "Heart Disease Prediction System using Data Mining Techniques and Intelligent Fuzzy Approach: A Review." Heart Disease 136, no. 2 (2016).

[34] Russell, Ingrid, and Zdravko Markov. "An Introduction to the Weka Data Mining System." In Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education, pp. 742-742. ACM, 2017.

[35] Nakamura, Shoichiro. Numerical analysis and graphic visualization with MATLAB. Prentice-Hall, Inc., 1995.

[36] Demšar, Janez, TomažCurk, AlešErjavec, ČrtGorup, TomažHočevar, MitarMilutinovič, Martin Možina et al. "Orange: data mining toolbox in Python." The Journal of Machine Learning Research 14, no. 1 (2013): 2349-2353.

[37] Hofmann, Markus, and Ralf Klinkenberg, eds. RapidMiner: Data mining use cases and business analytics applications. CRC Press, 2013.

[38] Rangra, Kalpana, and K. L. Bansal. "Comparative study of data mining tools." International journal of advanced research in computer science and software engineering 4, no. 6 (2014).

[39] https://eric.univ-lyon2.fr/~ricco/tanagra/index.html

[40] https://www.cs.waikato.ac.nz/ml/weka/

[41] https://www.mathworks.com/

[42] www.orange.biolab.si/

[43] www.rapidminer.com/