Singaporean Journal of Scientific Research(SJSR)
An International Journal (AIJ)
Vol.16.No.1 2024,Pp.9-16
ISSN: 1205-2421
available at :www.sjsronline.com
Paper Received : 04-01-2024 Paper Accepted: 05-03-2024
Paper Reviewed by: 1.Prof. Cheng Yu 2. Dr.Yarab Baig
Editor : Dr. Chen Du Fensidal

A Survey on Heart Disease Prediction Using Data Mining Techniques

M. Ranjani
Research Scholar
Dept. of Computer Science
Periyar University
Salem

Dr.P.R.Tamilselvi
Assistant Professor
Dept. of Computer Science
Govt. Arts & Science College
Komarapalayam -638 183.

Abstract: Data Mining is the process of transforming the raw data into useful information for decision making. Medical Data mining has potential to explore the hidden pattern in data set of the medical domain. These hidden patterns can be used for clinical diagnosis and disease prediction. Heart disease is one of the common and major causes for death in India. According to the report of Global Burden of Disease, 1.7 million Indians died because of cardiovascular illnesses. The number of deaths due to cardiovascular illnesses has grown up by 53% since 2005. According to ASSOCHAM-Deloitte joint study, the cases of cardiovascular diseases are growing at 9.5% annually. At the same time, it is considered to be the most preventable and controllable disease. There are certain factors which may cause heart diseases. The factors include change in life style, food habits, mental stress, smoking, alcohol consumption, obesity, blood pressure and diabetes etc. This research paper intends to provide a survey of data mining techniques used in medical field particularly in heart disease prediction. Various research works on association rule mining, classification and Ontology approaches for disease prediction are analyzed and presented in this paper.

Keyword: - Association rule, Classification, Data mining, Disease Prediction, Heart disease.
examples for Predictive process. Descriptive process

1.INTRODUCTION

Knowledge Discovery in Databases (KDD) is the process of identifying useful and understandable pattern in data. The term ―Pattern‖ refers to a subset of data for representing the subset. KDD aims at discovering pattern which is potentially useful, valid and provides some benefits for the users to proceed further. In KDD Process, a sequence of steps is iteratively involved. The sequences of steps are (i) Selection (ii) Pre-processing (iii) Transformation (iv) Data Mining and (v) Interpretation. Data mining plays the main role in KDD Process.

Data mining is defined as the computational process that analyses large raw amount of data and extract useful information and patterns. In recent years, Data mining is recognized as a powerful tool in various fields like information technology, medical, geosciences and many more. Machine learning is an emerging field that has taken many of the methods and techniques from data mining.

Data mining comprises a lot of algorithms which can be applied independently or combined based on the need for the specific application. Data mining processes are generally classified into two types namely Predictive and Descriptive. Predictive process refers to building a model for predicting future behavior for a given input attributes. Classification, Prediction and Deviation detection are the refers to describing the data in an understandable and effective form. For example, data characterization, association rule discovery and clustering are the examples for descriptive type processes.

Data mining techniques are widely used in many important areas like marketing organization, Education, Health care, Manufacturing Engineering, Customer Relationship Management (CRM), Fraud Detection, Intrusion Detection, Lie Detection, Criminal Investigation and Bio-informatics. In health care systems, data mining techniques are used to improve care and reduce the cost. Data mining approaches like machine learning, soft computing, data visualization and statistics are used by researchers to predict the volume of patients in every category. Health care systems are being developed for the patients to receive appropriate care at the right time and at the right place. Recently, researchers are interested in developing Disease Prediction Systems based on the symptoms and physical readings of the patients. This kind of system is very helpful in reducing the risk of death.

The objective of this paper is to analyze the existing works on data mining which have been used for heart disease prediction. This paper has been structured as follows. The introduction is presented in Section I, Data mining techniques are briefly discussed in Section II. Section III covers the literature survey. Section IV provides the details about the databases, performance metrics and tools related with data mining techniques and finally paper is concluded in Sec.V.

2. TECHNIQUES USED IN DATA MINING:

This section focuses on Data mining techniques that are commonly used in Heart Disease Prediction. The techniques are briefly discussed below.

A. Association:

Association is one of the best known and widely used data mining techniques. In Association, a pattern is discovered based on the relationship of a particular item on other items in the same transaction. For example, the association technique is used in heart disease prediction as it tell us the relationship of different attributes used for analysis and sort out the patient with all the risk factors which are required to predict the disease. [17]

B. Classification:

Classification is a classic data mining technique used to classify each item in a set of data into one of the predefined set of classes or groups. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics etc. [17]

C. Prediction:

The prediction as its name implies, it is one of the data mining techniques that discovers relationship between the independent variables and dependent variables. For a specific item and a corresponding model, the capacity of prediction is the ability to predict the value of a specific attribute. For example, in a predictive model for treatment

schema, prediction is used to determine the next procedure in the sequence of treatment.

D. Ontology

Ontology defines the classes of entities and their interrelations. They are used to organize the data according to the theory of the domain and provide class definitions (i.e., the necessary and sufficient conditions for defining class membership). Disease ontology is the type of biomedical ontology which provides the coverage for the domain of diseases, disorders, illness, etc. (Ceusters and Smith 2010), [6].

3. LITERATURE REVIEW:

This section aims at analyzing the various data mining techniques introduced in recent years for heart disease prediction. Different data mining techniques have been used in the diagnosis of Cardio Vascular Disease (CVD) over different Heart disease datasets. We categorize the overall research works and present in three sub areas.

A. Association Rule Mining

Association rule mining in medical data will generate a large number of rules. Most of the rules are found as irrelevant and it will reduce the speed of the search process also. For rule-based Decision Support System (DSS), A. Dhanasekar and R. Mala [8] proposed a method to find the strength of association among attributes. Valid Association rules are generated using probability measure. The proposed method has been tested on synthetic and real data sets. This work concludes that Coronary Vascular Disease (CVD) risk is more in male gender and in the age group of 55 – 65 years.

C. Ordonez et al. [23] introduced an algorithm that uses search constraints to reduce the Association rules. The author focused on two prediction issues such as predicting the absence as well as predicting the existence of heart disease. To reduce the number of rules, four constraints such as item filtering, attribute grouping, maximum item set size and consequent rule filtering have been used. The elimination of unreliable rules and impact of constraints are studied by experiments using real data set. A study on Data mining techniques for heart disease prediction was carried out by Himigiri et al. [7]. This work used three search constraints for improved performance. They are (i) Producing only medically useful rules, (ii) Reducing the number of discovered rules, (iii) Improving runtime. The authors concluded that a hybrid set of attributes may be binned by the decision tree whereas other attributes may be manually binned by the user. It is also suggested that a family of small decision trees may be used as an alternative for number of association rules.

Jyoti Soni et al. [30] designed a GUI based Interface to enter the patient record and performed heart disease prediction using Weighted Association rule based Classifier (WAC). In WAC, different weights are assigned to different attributes based on their predicting capability. The system has been developed using Java and trained with benchmark data of Machine Learning Repository (UCI). It is proved that WAC is providing improved accuracy than other existing Associative classifier. The proposed work obtained maximum of 81.51% of accuracy. For association rule mining, many methods have been proposed with single constraint. Soni J et al. [31] analyzed few current techniques of data mining that are used in heart disease prediction. Number of experiment was conducted to compare the performance of predictive data mining technique on the same dataset. The outcome reveals that Decision Tree outperforms other predictive methods like KNN, Neural Networks. It is also concluded that the accuracy of the Decision Tree and Bayesian Classification further improves the accuracy by applying genetic algorithm to reduce the actual data size into sufficient number of attributes for heart disease prediction. Li Guang-yuan et al. [11] presented an efficient algorithm for mining association rules with multiple constraints. The proposed method consists of three phases. Experimental results proved that the proposed method outperforms the revised Frequent Pattern growth algorithm. Ibrahim Umar Said et al.[14] analyzed the heart disease data based on gender using Apriori algorithm. From the experimental results, it is found that females have more chance of being free from coronary heart disease than males. Carlos Ordonez et al. [25] used Greedy algorithm to discover significant association rules. The significance of association rules is evaluated using three metrics namely support, confidence and lift. This association rules mining approach was tested on real-time dataset.

R.Thanigaivel and K.Ramesh Kumar[36] analyzed various Data mining techniques for heart disease prediction. The authors found that neural network with offline model training are suitable for disease prediction in early stage. Preprocessing and Data normalization methods are helpful in improving the performance and better

classification accuracy can be achieved by feature reduction. S.P.Syed Ibrahim and K.R.Chandran [13] proposed a weighted class Association rule mining method which applies Weighted Association rule mining in the classification and Compact Weighted Associative Classifier (CWAC) is constructed. The weight of the data item is considered for generating weighted class Association rules. The proposed algorithm computes the weight using Hyperlink-Induced Topic Search HITS model and generates less number of rules which improves the classification accuracy.

Carlos Ordonez [26] compared the Association rules and predictive rules mined with decision trees. It is investigated that Decision trees are shown to be not as adequate for artery disease prediction as association rules. Decision tree produces simple rules, but they are not reliable and most of the rules refer to small sets of patient data. In contrast, Association rules produce simpler predictive rules and work well with user-binned attribute, rule reliability is higher and refer to larger sets of patient data.

Carlos Ordonez et al. [23] performed a study on prediction of heart disease with the help of Association rules. The authors used a simple mapping algorithm. This algorithm constantly treats attributes as numerical or categorical. This is used to the convert medical records to the transaction format. An improved algorithm is used to mine the constrained association rules. A mapping table is prepared and attribute values are mapped to items. The decision tree is used for mining data because they automatically split numerical values. The split point chosen by the Decision tree are of little use only. Clustering is used to get a global understanding of data.
Ordonez et al. [24] introduced a Greedy algorithm to discover only significant rules and to accelerate the search process. The significance of association rules is evaluated using three metrics namely support, confidence and lift.

B. Classification using heart disease prediction:

R. Bhuvaneswari et al. [5] used two well-known algorithms of data mining classification namely Back propagation Neural Network (BNN) and Naïve Bayesian (NB). Since Bayesian method works based on the probability theory, Naïve Bayes classification method is recommended by the authors for Decision support system. It is also found that statistics based algorithms need modifications to be used by Data mining approaches. Prediction of Heart Disease using classification algorithms was done by Hlaudi Daniel Masethe and Mosima Anna Masethe [20]. In this work, J48, Bayes Net, and Naïve Bayes, Simple Cart, and REPTREE algorithms are applied to classify and develop a model to diagnose heart attacks. Data set consists of 11 attributes which is generally referred by the health practitioner for predicting the heart disease is considered for experiment. Shadab Adam Pattekari and Asma Parveen [27] developed a web based Intelligent System using Naive Bayesian Classification technique. The system can answer complex queries of users for diagnosing heart disease. It also assists the health care practitioners to make clinical decisions.

Jagdeep Singh et al. [29] developed a framework using Associative classification technique for early diagnosis of heart diseases. This work was tested on UCI repository dataset. The various attributes like gender, age, chest pain type, blood pressure, blood sugar are used to predict the heart disease. Data mining algorithms such as Apriori , FP-Growth, Naïve Bayes, ZeroR, OneR,J48 and K- Nearest Neighbor are used for heart disease prediction. The prediction accuracy of 99.19% is achieved by this framework. Abhishek Taneja [35] designed few predictive models for heart disease detection using Data mining techniques. The author used three different supervised machine learning algorithms namely Decision Tree classifier, Bayesian Classifier and Neural Network. The performances of the models were evaluated using the metrics like accuracy, precision, recall and F-measure. The domain experts confirmed that the most of the generated rules are important in diagnosing heart diseases. Among various models, J48 classifier with selected attributes is found as an effective model. It provides 95.56% of classification accuracy. The resulting model has been recommended by the author for junior cardiologist for screening Cardiac patients. A new prediction model has been developed by N. Aditya Sundar, et al. [32] using data mining techniques such as Naïve Bayes and Weighted Associative Classifier (WAC). Data Mining Extension (DMX) query language has been used for model creation, training, prediction and content retrieval. This system uses the CRISP-DM (CRoss Industry Standard Process for Data Mining) methodology to build the mining models. This model extracts the hidden knowledge from a historical heart disease database and validated against a test dataset. Classification Matrix methods are used to evaluate the effectiveness of the models.

K.Thenmozhi and P.Deepika [37] explored the utility of various decision tree algorithms in classification and prediction of diseases. Different attribute selection measures like Information Gain, Gain Ratio, Gini Index and Distance measures are used for selecting the attributes. Theresa Princy and R. J. Thomas [38] analyzed the performance of different classification techniques for risk level prediction and survey report is presented. The authors

used various parameters like gender, blood pressure, cholesterol level, pulse rate and etc. The classification techniques such as Naïve Bayes, KNN, Decision Tree Algorithm, Neural Network are used for classification. The authors found that more number of attributes is required for better accuracy rate. Sudha et al. [34] used various classification algorithms like Naive Bayes, Decision tree and Neural Network for detecting the stroke diseases. In this work, Principle Component Analysis (PCA) was used for reducing the dimension. It is observed that neural network classifier provides better accuracy than other classifiers. A new model has been proposed by Mai Shouman, et al. [28] to diagnose the heart disease.

Moloud Abdar, et al. [1] compared the different data mining techniques used for heart disease prediction. After analyzing the features, prediction models using algorithms such as C5.0, Neural Network, Support vector Machine (SVM), K-Nearest Neighborhood (KNN) and Logistic Regression have been developed and validated. A model built using C5.0 Decision tree provides the highest accuracy of 93.02% where as KNN, SVM, Neural Network provided 88.37%, 86.05% and 80.23% respectively. Sick and healthy factors which contribute to heart disease are identified by Jasmine Nahar et al. [22]. UCI data repository is considered and rules are generated using Apriori, Predictive Apriori and Tertius. The relationship between heart disease risk factors in men and women is found. It is found that coronary heart disease risk in women is less than men. This paper concludes that Rest ECG should be considered as an important factor to predict heart disease in women. Among other algorithm, Predictive Apriori provides the better precision rate of 90%.

Gandhi Monika and Shailendra Narayan Singh[10] analyzed various data mining methods that are being used in today's research for prediction of heart disease. In this work, data mining methods namely Naive Bayes, Neural network, Decision tree algorithms are analyzed on medical data sets. Bahrami et al. [3] evaluated different classification techniques used in heart disease diagnosis. Classifiers like J48 Decision Tree, KNN, Naive Bayes, and Sequential Minimal Optimization (SMO) are used to classify the dataset. After classification, performance is evaluated using the measures like accuracy, precision, sensitivity, specificity, F-measure and area under ROC curve and comparison is made. The comparison results show that J48 Decision tree is the best classifier for heart disease diagnosis.

C. Heart disease prediction using ontology method Baydaa Al-Hamadani [2] proposed an Expert system called ―CardioOWL‖ to diagnose any kind of coronary artery diseases. CardioOWL recommends the drugs and other required surgery for the patients. CardioOWL was developed based on ontology knowledge about the patient's symptoms and it was developed using Semantic Web Rule Language (SWRL). The system has been tested by some general practitioners using several test cases. The system provides very good precision and recall rates.

S.T. Liawa et al. [19] analyzed so many works on ontology and recommended an ontological approach for Chronic Disease Management (CDM). The authors suggest that an ontology based design of information ystems enable more reliable use of routine data for measuring health mechanisms and impacts. Hamid Mcheick et al. [21] proposed a Stroke Prediction System (SPS) based on ontology and Bayesian Belief Networks (BBN). This helps to handle the stroke disease by determining the risk score level. This system is composed of four layers namely acquisition of data, aggregation, reasoning and application. Parminder Kaur and Aditya Khamparia[16] performed a comparative study on different medical ontologies, tools, models, features and languages. Sowkarthikaa and Sumathi V. P [33] addressed the problem of classifying the disease based on medical ontology. The techniques used in medical ontology such as Case profile ontology, decision support tool and particle swarm optimization model are analyzed. To evaluate the effectiveness of the intelligent system, three benchmark medical data sets, viz., Breast Cancer Wisconsin, Pima Indians Diabetes and Liver Disorders from the UCI Repository were used. The experimental results demonstrate that the hybrid intelligent system is found as an effective system in undertaking medical data classification tasks. Mike Uschold and Robert Jasper [15] proposed a framework for understanding and classifying ontology applications. The authors identified the three main categories of ontology applications. They are 1) neutral authoring, 2) common access to information, and 3) indexing for search. In each category, the authors identified specific ontology application scenarios. The intended purpose, the role of the ontology, the supporting technologies are identified for each category. The similarities and differences between scenarios are also identified.

Kamran Farooq et al. [12] proposed a hybrid clinical decision support framework which comprises of two key components. The two components are (1) ontology driven clinical risk assessment and recommendation system, (2) machine learning driven prognostic system. Few clinical case studies in the heart disease and breast cancer domains are considered for the development and clinical validation of the proposed framework. The proposed ontology driven

clinical risk assessment system is recommended for cardiovascular patients as a preventative solution. Fred Freitas, et al. [9] analyzed few ontology methods for Biology and Medicine. The authors introduced a framework and compare the systems in terms of their architectural elements, expressiveness and coverage. They identified International Classification of Diseases (ICD), Medical Subject Headings (MeSH), Gene Ontology (GO), Systematized Nomenclature of Medicine -Clinical Terms (SNOMED CT), Generalized Architecture for Languages, Encyclopedias and Nomenclatures (open GALEN), Foundational Model of Anatomy (FMA), Unified Medical Language System (UMLS) and Open Biomedical Ontologies (OBO) Foundry. Abinaya Sambath Kumar and A. Nirmala [18] proposed a cardiovascular decision support framework using an ontology driven approach. An ontology of cardiovascular diseases structured on Open Biomedical Ontologies (OBO) Foundry was presented by Barton A et al. [4].

4. DATABASES, METRICS and TOOLS

This Section provides the details about the databases, performance metrics and tools related with data mining algorithms.

A. Databases

For research purpose, any user can download the heart disease data from UCI respiratory [39]. The data set is available in http:// archive .ics. uci. edu/ ml/ datasets
/statlog +(heart).

B. Performance Metrics

To analyze the performance of Data mining algorithm, the following metrics are used.

Association Rules

* Minimum support and confidence
* Antecedent and Consequent

Predictive Model Performance Evaluation Metrics

* Accuracy *Sensitivity *Specificity * Precision

C. Tools

Many tools and software are used for Data Mining Techniques. Few of them are given in Table 1.

5. CONCLUSION

In this survey paper, we have analyzed various existing techniques used for predicting heart disease with the support of data mining and ontology methods. From this survey we gained the knowledge of how to apply data mining techniques to predict the heart disease. Previously existing system was designed with single algorithm which has not provided results with better accuracy. With the help of the research works done earlier we have observed that, by hybridization of two or more algorithms we can derive more accuracy than the existing methods. So, In future new algorithms and new techniques are to be developed which will certainly overcome the drawbacks of the existing system which will help to diagnose the heart disease accurately and to take preventive measures

REFERENCES

[1] Abdar, Moloud, et al. "Comparing Performance of Data Mining Algorithms in Prediction Heart Diseases." International Journal of Electrical and Computer Engineering (IJECE) 5.6 (2015): 1569-1576.

[2] Al-Hamadani B. CardioOWL: An ontology-driven expert system for diagnosing coronary artery diseases. InOpen Systems (ICOS), 2014 IEEE Conference on 2014 Oct 26 (pp. 128-132). IEEE.

[3] Bahrami, Boshra, and Mirsaeid Hosseini Shirvani. "Prediction and Diagnosis of Heart Disease by Data Mining Techniques." Journal of Multidisciplinary Engineering Science and Technology (JMEST) 2.2 (2015): 164-168.

[4] Barton A, Rosier A, Burgun A, Ethier JF. The Cardiovascular Disease Ontology. InFOIS 2014 Sep 5 (pp. 409-414

[5] Bhuvaneswari R, Kalaiselvi K. Naive Bayesian classification approach in healthcare applications. International Journal of Computer Science and Telecommunications. 2012 Jan;3(1):106-12.

[6] Ceusters W, Smith B. Foundations for realist ontology of mental disease. Journal of biomedical semantics. 2010 Dec;1(1):10.

[7] Danapana, Himigiri, and M. Sumender Roy. "Effective Data Mining Association Rules for Heart Disease Prediction System 1." (2011).

[8] A. Dhanasekar, Dr. R. Mala, ―Analysis of Association Rule for Heart Disease Prediction from Large Datasets‖ international Journal of Innovative Research in Science, Engineering and Technology (An ISO 3297: 2007 Certified Organization) Vol. 5, Issue 10, October 2016.

[9] Freitas F, Schulz S, Moraes E. Survey of current terminologies and ontologies in biology and medicine.RECIIS—Electronic Journal in Communication, Information and Innovation in Health. 2009 Mar;3(1):7- 18.

[10] Gandhi, Monika, and Shailendra Narayan Singh. "Predictions in heart disease using techniques of data mining." Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE), 2015 International Conference on. IEEE, 2015.

[11] Guang-Yuan L, Dan-yang C, Jian-Wei G. Association Rules Mining with Multiple Constraints. Procedia Engineering. 2011 Jan 1; 15:1678-83.

[12] Kamran Farooq, Amir Hussain Hefei, Anhui, China Bin Luo,‖ A Novel Ontology and Machine Learning Inspired Hybrid Cardiovascular Decision Support Framework‖, 2015 IEEE Symposium Series on Computational Intelligence 978-1-4799-7560-0/15© 2015 IEEE DOI 10.1109/SSCI.2015.122.

[13] Ibrahim SP, Chandran KR. Compact Weighted Class Association Rule Mining using Information Gain. arXiv preprint arXiv:1112.2137. 2011 Dec 9.

[14] Ibrahim Umar Said , Abdullahi Haruna Adam, and Dr. Ahmed Baita Garko ―Association Rule Mining on Medical Data to Predict Heart‖ International Journal of Science Technology and Management Vol. No.4,Issue08, August 2015
.

[15] Jasper R, Uschold M. A framework for understanding and classifying ontology applications. In Proceedings 12th Int. Workshop on Knowledge Acquisition, Modeling, and Management KAW 1999 Oct (Vol. 99, pp. 16-21).

[16] Kaur, Parminder, and Aditya Khamparia. "Review on Medical Care Ontologies." Liver (2012): 6.

[17] Kaur B, Singh W. Review on heart disease prediction system using data mining techniques. International journal on recent and innovation trends in computing & communication. 2014 Oct; 2(10):3003-8.

[18] Kumar, Abinaya Sambath, and A. Nirmala. "A Survey on Ontology Based Disease Diagonsis Algorithms." International Journal 5.10 (2015).

[19] Liaw ST, Rahimi A, Ray P, Taggart J, Dennis S, de Lusignan S, Jalaludin B, Yeo AE, Talaei-Khoei A. Towards an ontology for data quality in integrated chronic disease management: a realist review of the literature. International journal of medical informatics. 2013 Jan 1;82(1):10-24.

[20] Masethe HD, Masethe MA. Prediction of heart disease using classification algorithms. In Proceedings of the world congress on Engineering and Computer Science 2014 Oct 22 (Vol. 2, pp. 22-24).

[21] Mcheick H, Nasser H, Dbouk M, Nasser A. Stroke prediction context-aware health care system. InConnected Health: Applications, Systems and Engineering Technologies (CHASE), 2016 IEEE First International Conference on 2016 Jun 27 (pp. 30-35). IEEE.

[22] Nahar, Jesmin, et al. "Association rule mining to detect factors which contribute to heart disease in males and females." Expert Systems with Applications 40.4 (2013): 1086-1093.

[23] Ordonez C, Omiecinski E, De Braal L, Santana CA, Ezquerra N, Taboada JA, Cooke D, Krawczynska E, Garcia EV. Mining constrained association rules to predict heart disease. InData Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on 2001 (pp. 433-440). IEEE.

[24] Ordonez C, Ezquerra N, Santana CA. Constraining and summarizing association rules in medical data. Knowledge and information systems. 2006 Mar 1;9(3):1- 2.

[25] Ordonez C. Association rule discovery with the train and test approach for heart disease prediction. IEEE Transactions on Information Technology in Biomedicine. 2006 Apr;10(2):334-43.