International Journal of Scientific Research & Engineering Trends
Volume 7, Issue 2, Mar-Apr-2021, ISSN (Online): 2395-566X

Study and Analysis in Heart Disease Analysis Using K-Nearest Neighbor Classifier

Research Scholar Wasim Akaram, Professor Dr. Rohit Kumar Singhal
Institute of Engineering and Technology, Alwar

Abstract- Data mining refers to analysis of complex data. The prediction is the process of determining what will happen next. Recently, various techniques have been applied for the prediction analysis. A SVM technique is applied to the prediction analysis. The technique divides data into training and testing stages. The first class of test data is for the most part related to the individuals who have little to no risk of having a heart disease . The second class of test data all have risk-of-heart-disease levels above 50%. This research work proposes to improve this existing method using decision tree classifier. The proposal would improve accuracy and reduce the execution time.

Keywords- Data mining, testing stages, SVM technique

I.INTRODUCTION TO DATA MINING

Data mining is the method of going over the data and extracting the details and patterns that could be used to determine something significant about the user. There are various types of data mining software that can be used to analyse different types of data. This method of using the knowledge embedded in the different fields has evolved into a major one, one that can be used by several fields. One of the main functions of the EDI is to analyse the store information such that only effective information can be used. [1] We have some experience of looking at data for relational databases such as personal records (medical records), multimedia assets and object relational databases.

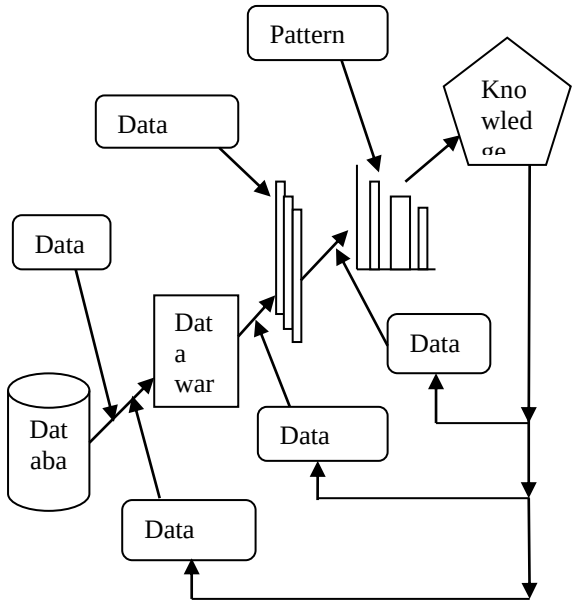

Fig. 1 Data Mining Process

In order to get a good rate, the treatment should first clean the data such that it will remove noise and irrelevant data. The next step is data integration, which is used to place several data sources into one cohesive unit. When selecting the data to be gleaned from the database during the second step, information is detected by the computer [2]. It is a data collecting tool that gathers a large amount of data so that by acquiring knowledge, that isn't really true in how the data was gathered. In today's world, a large amount of data is collected and analysed about the general public to gauge how the public is adapting to social transition.

In order to go through the data, one might need to slightly manipulate the data in order to extract accurate result from it. However, gold is basically obtained from rocks or sand and is thus referred to as gold mining, rather than rock or sand mining. Mining is the process by which raw materials are extracted, and is defined as the act of extracting something from the earth. Many terms are used to describe what knowledge extraction is and how it is relevant: data archaeology, knowledge mining from databases and data dredging. It is a world where there is a lot of information floating around, but only the most powerful can make use of this information because of sophisticated technological equipment such as satellites and computers. An easy way to access so many different types of information is through the use of the main computer systems which stores a wide range of data, including the medical information we need to keep a check on all patients.

1. Techniques used in data mining

Association: In the data mining technique, association is the best technique among others. On the basis of the relationship a pattern is discovered with the help of association for the particular item on other items for the transaction of the similar data. For the prediction of the

heart disease this association technique is utilised which provide information and used for analysis about the relationship of different attributes. All the risk factor in the patients are sorted out that utilised for prediction of disease.

2. Classification- In the data mining technique, classification is a classic method that is based on machine learning. In the classification of the data, each item in the data set is classified into predefined set of classes or groups. Decision trees, linear programming, neural network and statistics are the mathematical techniques that have been utilised by the classification method [6].

3. Clustering- Clustering is a data mining technique in which clustering of objects are identified. This can be done using automatic technique as it has similar characteristics. The classes have been defined by the clustering technique and objects are placed unlike classification as in this process objects are assigned into predefined classes. With the help of clustering technique for the prediction of heart. disease, clusters sets are obtained in which contains list of patients which have same risk factor. Therefore, using this technique a separate list of patients is made.

4. Prediction- In the data mining technique, prediction is one of the techniques that discover the relationship between independent variables and dependent variables. This technique can be utilised in various fields such as in sale in order to predict profit for future hence, profit is referred as a dependent variable and sale as an independent variable. A fitted regression curve can be drawn using historical sale and profit data that is utilised for the prediction of profit.

5. Issues in Data Mining

DM algorithms represent the methods which are sometimes existed for a number of years. However, their utilization as reliable and scalable tools can be done later on and it performed better than the earlier traditional statistical techniques. Though, the DM is in its premature age, it has acquired a great attention and becomes ever-present. There are some issues whose tackling is still required prior to advancement of DM into a traditional, mature and trusted discipline. These issues are described as:

- Security and social issues: It is an essential issue in any data collection shared and suggested to carry out strategic decision-making. Furthermore, the collection and storage of a huge amount of sensitive and private information regarding the individuals or companies have performed after gathering the data for various purposes. The confidential nature of some of this data and the illegal accessing of the information make it notorious. Additionally, new implicit information regarding individuals or groups will be revealed through DM which can be against the privacy policies, particularly, in case of potential distribution of discovered information.
- User interface issues: The knowledge investigated from the DM tools is valuable till it seems interesting and understandable for the user [25]. The DM outcomes can be easily interpreted through the fine visualization of data that also assists the users in understanding their requirements. The screen real-estate and interaction are main concerns regarding the user interfaces and visualization.
- Mining methodology issues: these issues are related to implementation of DM technique and their drawbacks. The concepts such as flexibility of the mining techniques, to estimate the investigated knowledge, to utilize the background knowledge and metadata, to control and handle the noise in data, etc. are all the instances using which mining methodology choices is dictated. To illustrate, the availability of various DM techniques is required as dissimilar approaches can be run differently on the basis of data. Furthermore, different approaches are suitable and are capable of dealing with the requirements user in different manner [26].
- Performance issues: A number of AI and statistical techniques are available in order to analyze and interpret the data. But, the formation of these techniques is not done for enormous data sets which the DM is dealing recently. TB sizes are very common. The issues regarding the scalability and efficacy of DM techniques are developed while processing the significant huge amount of data. The incremental updating and parallel programming are the other issues. The parallelism assists in tackling the size problem only when the dataset can be subdivided and the integration of results can be done later on. Incremental updating is essential in order to combine the outcomes from parallel mining or update results of DM at time of availability of novel data without re-analyzing the entire dataset [27].

6. Machine Learning

Machine learning is a branch of artificial intelligence that allows computer systems to learn directly from examples, data, and experience. Through enabling computers to perform specific tasks intelligently, machine learning systems can carry out complex processes by learning from data, rather than following pre-programmed rules. Recent years have seen exciting advances in machine learning, which have raised its capabilities across a suite of applications. Increasing data availability has allowed machine learning systems to be trained on a large pool of examples, while increasing computer processing power has supported the analytical capabilities of these systems [30]. Within the field itself there have also been algorithmic advances, which have given machine learning greater power. As a result of these advances, systems which only a few years ago performed at noticeably below-human levels can now outperform humans at some specific tasks

.

II.LITERATURE REVIEW

Tülay Karayilan, et.al, “Prediction of Heart Disease Using Neural Network”, 2017 Heart disease is the fatal disease from which large number of population is currently suffering as its detection and prevention is major and required to diagnose at the early stage. The process of diagnosis for this diseases is complicated as it requires

proper monitoring therefore, early detection of this disease is necessary and accurately. This disease cause maximum numbers of causalities [28]. In the traditional methods there are various limitations as analyzed doing experiments, therefore enhanced methods have been proposed in this paper.

Ms. Tejaswini U. Mane, et.al "Smart heart disease prediction system using Improved K-Means and ID3 on Big Data", 2017 The survey performed by the world health organization in the worldwide for the heart disease is presented here in which every year more than 12 million deaths occur due to this fatal disease. Therefore maximum casualties are caused due to which detection of this disease is necessary. Heart disease sometimes referred as the big data approach and for the reduction of such big data Hadoop Map platform has been utilized.

Kanika Pahwa, et.al, "Prediction of Heart Disease Using Hybrid Technique For Selecting Features", 2017 In the health care industry there is large amount of data is available which needs to be discovered with the help of hidden patterns as per the requirement. In order to provide effective decision making, the advancement in the data mining techniques is required within this field. The features that are embedded in the proposed technique can be utilized with the help of Random forest and Naïve bayes techniques. On the basis of the obtained results the performance level of the method can be increased [30]. For the selection of the features from the dataset the SVM-RFE and gain ratio algorithms has been utilized. A particular weight is assigned to each feature after applying this technique. As per performed experiments, it is concluded that the proposed method provide the optimal accuracy and minimization of computational time.

Sarath Babu, et.al, "Heart Disease Diagnosis Using Data Mining Technique", 2017 A data mining technique is used to extract the patterns and analyze huge volume of data in order to convert it in useful information. It is the process of extracting useful information that can be utilized for implementation. In the medical dataset or medical domain a key role is played by the medical data mining which has the great potential to discover the hidden patterns. These patterns have been utilized for the clinical diagnosis for which data need to be collected in a standardized form. Age, sex, blood pressure and blood sugar and many more are the attributes that are extracted from the medical profiles in order to determine whether patient is suffering from heart disease or not [31].

Monira Islam et.al "A Non-Invasive Technique of Early Heart Diseases Prediction from Photoplethysmography Signal," 2017 Non-invasive techniques are proposed to detect the heart rate from Photoplethysmography (PPG) signal. [32] This proposed technique plays a very significant role in the diagnosis of heart related disease. This creates very less discomfort to any patient because traditional sensors can damage the tissues during extraction of cardiac signals. A convenient heart diseases detection mechanism is used with the help of PPG and extracted from human facial videos. This can overcome the expensive ECG machine for heart related disease detection. The FFT can detect the heart rate and re-check with the conventional heart rate using ECG machines.

Tahira Mahboob et.al, "Evaluating Ensemble Prediction of Coronary Heart Disease using Receiver Operating Characteristics," 2017 Authors studied about various learning practices which assist the detection of innumerable the heart disease [33]. There are certain techniques were used like Hidden Markov Models, Support Vector Machine, Computational intelligent classifier, and data mining and so on. As the cardiac diseases treatment is very expensive and unaffordable to any normal individual so, these types of advanced technology are developed to overcome this problem. These techniques are also useful in early stage predictions. It avoids any other future sufferings by making slight changes in daily routine. Hence, the author concludes that the predicted approach has several advantages and is very useful.

Procheta Nag et.al, "A Simple Acute Myocardial Infarction (Heart Attack) Prediction System Using Clinical Data and Data Mining Techniques," 2017 Authors proposed a very proficient approach which can predict the early chances of cardiac rest when a person is having any kind of chest pain or any other early symptoms [34]. A prototype was developed by dividing the clinical data collected from hospitalized patients who are suffering from Acute Myocardial Infarction (AMI). Heat attack contain several symptoms like pain in chest, problem in breathing, palpitation, nausea, vomiting, continuous sweating and so on. Decision tree which comes under data mining and another technique random forest are use to study the heart attack rate which further classify the more symptoms of heart attack. Therefore, the result of the technique concludes that the advancement of computer technology in medical and health region provides useful aids and people are becoming more dependent on these technologies. The results of data mining are very beneficial and are used for the better assistance to.

III.PROBLEM FORMULATION

The prediction analysis is the technique which can predict the future possibilities from the existing data. The prediction analysis techniques are based on the clustering and classification. In the basepaper, medical data is analyzed to predict the regional diseases. The data is collected from the central china in 2013-2015 to prepare modal for predicted analysis. The base paper modal for the prediction analysis is based on the neural networks.. The clustered data is given as input to the classification algorithm which can divide the dataset into two parts testing and training. The SVM

classifier is used to classify the data into certain number of classes. In the k-mean clustering algorithm, the centered points are calculated by taking arithmetic mean of the whole dataset which can reduce accuracy of prediction analysis. When the dataset is complex, it is difficult to establish relationship between the attributes of the dataset. In this research work, the decision tree classifier is applied which can classify the wheat production in certain number of classes. The decision tree classifier can be replaced with some other classifiers to improve accuracy of classification.

V.RESEARCH METHODOLOGY

This research work is based on the prediction analysis of heart diseases. The prediction analysis is the technique in which future possibilities can predicted based on the current dataset. In this research work, technique of decision tree is applied previously for the prediction analysis. One of the simplest algorithms amongst all the learning machine algorithms is the decision tree algorithm. Since there are no assumptions made on the underlying data distribution, decision tree is known to be a non-parametric supervised learning algorithm. Here, on the basis of nearest training samples present within the feature space, the samples are classified. The feature vectors are stored along with the labels of training pictures within the training process. Towards the label of its k-nearest neighbors,

Following are the techniques used in Research Methodology :On the basis of the given output prediction for the certain outcomes can be done by classification process. The outcome is predicted by the processed algorithm as different set of attributes are present in the training set and in the respective outcome which is called as goal or prediction attribute. The relationship between the attributes is discovered by the algorithm that will be helpful in the prediction of the outcome. A data set is provided by the algorithm known as the prediction set in which the same set of attributes are present but the prediction attributeis absent that is not well known. The input is analyzed by the algorithm that is helpful in the prediction process. The accuracy of the algorithm is defined by the prediction accuracy [19].

1.Genetic programming

Genetic programming (GP) has been widely used in the area of research as it solves data mining classification problems. The major fact that genetic programming is so widely used is because it predict the rules effectively that are naturally represented in GP. Optimal results are produced by the GP with global search problems like classification. Several ‘peaks’ are present in the search space for classification this cause local search algorithms also known as simulated annealing that performs badly.

2. Neural networks

Neural networks are the interconnectivity between the processing elements also called units, nodes, or neurons. These networks are designed after the cognitive processes of the brain. These networks are used to predict new outcomes from the previous observations. In order to produce an output function all the present neurons within the network work together. The collective neurons performed the computational functions within a neural network still it is capable of producing the output function even some of the individual neurons are malfunctioning. Instead network remains robust and fault tolerant. Within a neural network, each neuron has an associated activation number and also a weight associated in each connection between the neurons [20].

3.Ant colony

Ant Colony algorithms is the natural inspired technique and by the behavior of ants as they help in finding the optimal path from the colony to food. They use the good paths within a graph in order to find optimal ways [21]. Chemical called pheromones are deposited by the ant on their trails when they are travelling from one place to another in the search of food. Ants used these trails to find their way back to the colony without distracting and it is followed by other ants as well if they find path safe. Due to this effect more pheromone are deposited on the trail which cause the effect of reinforcing. If for the long way, these trails have not been utilized than the pheromone starts to evaporate. The density of pheromone remains high as the short and optimal paths are utilized again and again that provides faster rate to find food. Therefore, large amount of ants travelling on the shortest path due to which density of the pheromone is increased that is followed by all ants as well. This behavior of the ants is copied by the Ant Colony algorithms by which an optimal within a graph is determined. In the initial stage small amount of pheromone is deposited on the trails of the ant randomly.

4. Statistical algorithms: ID3 AND C4.5

The development of the ID3 algorithm is done by J. Ross Quinlan at the University of Sydney that was published in the book of “machine learning” in the year 1975. Using data, classification models are induced by the ID3 algorithm. This method is also known as the supervised learning algorithm for different classes and also they are trained due to which the prediction of the class for a new item becomes easy. Attributes are identified by the ID3 that differentiate one class from another and all these attributes must be well known selected from the known set of values. Temperature and country of citizenship are the valid attributes. The statistical property of entropy has been utilized by the ID3 in order to determine the importance of the attributes. The amount of information in an attribute is measured by the entropy measure. Therefore, in this way the decision tree is build that is utilized to test future cases.

VI.CONCLUSION

The relevant information is fetched from rough dataset using data mining technique. The similar and dissimilar data

is clustered after calculating a similarity between input dataset. The SVM used to classify both similar and dissimilar data type in which central point is calculated by calculating an arithmetic mean of the dataset. The central point calculated Euclidian distance is used to calculate a similarity between different data points. According to the type of input dataset a clustered data is classified using decision tree classifier scheme in the last step. In this research work, the SVM classifier will be replaced with the decision tree classifier. It is analyzed that decision tree has less execution time and high accuracy of classification

Future Work

Following are the various future possibilities:-

1.The proposed algorithm can be further improved using the hybrid type of classifier for the prediction analysis

2.The proposed algorithm can be compared with the other classifiers to analyze its reliability.

REFERENCES

[1] Monali Dey, Siddharth Swarup Rautaray, Study and Analysis of Data mining Algorithms for Healthcare Decision Support System, International Journal of Computer Science and Information Technologies, vol. 6, issue 3, pp. 234-239, 2014.

[2] Oyelade, O. J, Oladipupo, O. O and Obagbuwa, I. C (2010), "Application of k-Means Clustering algorithm for prediction of Students' Academic Performance", International Journal of Computer Science and Information Security, vol. 7, issue 4, pp. 123-128, 2010.

[3] Azhar Rauf, Mahfooz, Shah Khusro and Huma Javed (2012), "Enhanced K-Mean Clustering Algorithm to Reduce Number of Iterations and Time Complexity", Middle-East Journal of Scientific Research, vol. 12, issue 6, pp. 959-963, 2012.

[4] Indira S. Fal Dessai, Intelligent Heart Disease Prediction System Using Probabilistic Neural Network, International Journal on Advanced Computer Theory and Engineering, vol. 7, issue 4, pp-56-62, 2013.

[5] Abhishek taneja, Heart Disease Prediction System Using Data Mining Techniques, Oriental Scientific Publishing Co., India, vol. 5, issue 4, pp. 959-963, 2013.

[6] Kajal C. Agrawal and Meghana Nagori (2013), "Clusters of Ayurvedic Medicines Using Improved K-means Algorithm", International Conf. on Advances in Computer Science and Electronics Engineering, vol. 23, issue 3, pp. 546-552, 2013.

[7] Swain Sunita, Badajena J Chandrakanta and Rout Chinmayee, A Hybrid Approach of Intrusion Detection using ANN and FCM, European Journal of Advances in Engineering and Technology, vol. 3, issue 2, pp. 6-14, 2016.

[8] Tetiana Gladkykh, Taras Hnot and Volodymyr Solskyy, Fuzzy Logic Inference for Unsupervised Anomaly Detection, IEEE First International Conference on Data Stream Mining & Processing vol. 4, issue 1, pp. 42-47, 2016.

[9] Jesmin Nahar, Tasadduq Imama, Kevin S. Tickle, Yi-Ping Phoebe Chen, Association rule mining to detect factors which contribute to heart disease in males and females, Elsevier, vol. 8, issue 1, pp. 23-48, 2013.

[10] Resul Das, Ibrahim Turkoglu, Abdulkadir Sengur Diagnosis of valvular heart disease through neural networks ensembles, Elsevier,vol. 4, issue 1, pp. 23-48, 2009.

[11] Mohammed Mahmood Ali, Khaja Moizuddin Mohammed and Lakshmi Rajamani. "Framework for Surveillance of Instant Messages in Instant messengers and Social networking sites using Data Mining and Ontology", IEEE- Students' Technology Symposium, vol. 4, issue 1, pp. 23-48, 2014.

[12] Sushant Bharti, Ashutosh Mishra. "Prediction of Future possible offender's network and role of offender's",Fifth International Conference on Advances in Computing and Communications, vol. 8, issue 1, pp. 23-48, 2015.

[13] Dahlia Asyiqin Ahmad Zainaddin and Zurina Mohd Hanapi, Hybrid of Fuzzy Clustering Neural Network over Nsl Dataset for Intrusion Detection System, Journal of Computer Science, Volume 9, No. 3, pp. 391-403, 2013.

[14] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Din, "Data Mining With Big Data",IEEE Transactions on Knowledge and Data Engineering, Vol. 26, issue 1, pp. 23-34, 2014

[15] L. Parthiban and R. Subramanian, "Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm", International Journal of Biological and Life Sciences, vol. 3, no. 3, pp. 157-160, 2007.

[16] Anupama Chadha, Suresh Kumar,"An Improved K-Means Clustering Algorithm: A Step Forward for Removal of Dependency on K", 2014 International Conference on Reliability, Optimization and Information Technology -ICROIT 2014, vol. 8, issue 1, pp. 6-8, 2014