SURVEY PAPER ON PREDICTION OF DISEASES USING DATA MINING TECHNIQUES

Ranjana Joshi[1] , Shilpa Sethi[2]
Research scholar[1]
Assistant Proffessor[2]
Department of Computer Application, YMCA University of Science & Technology, Faridabad

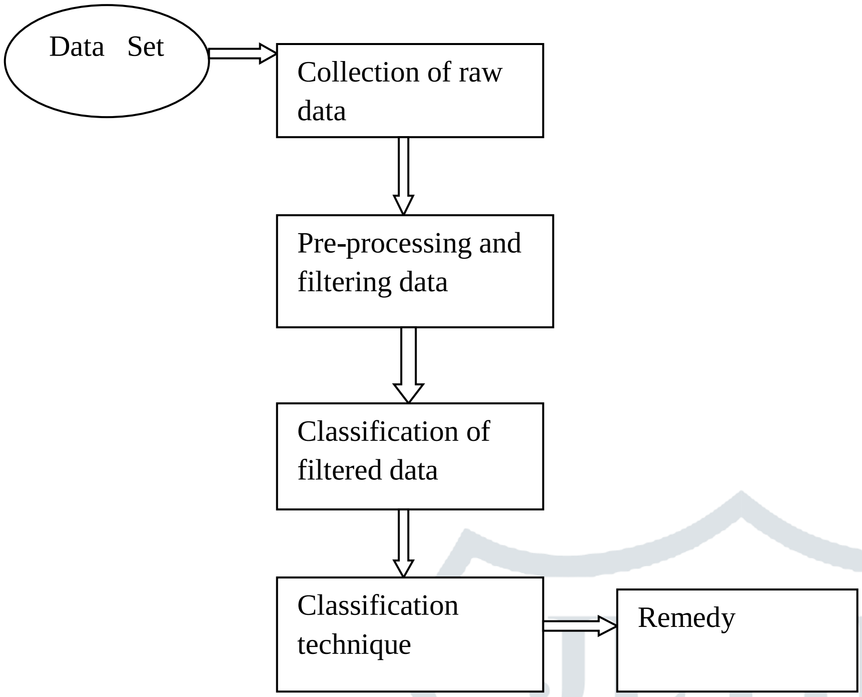
Abstract:- The medical field is dealing with huge amount of data regularly. Handling that large data by traditional way may affect the results. Data mining provides the methodology and technology to transform these mounds of data into useful information for decision making. In today's modern world cardiovascular disease is the most lethal one. Diagnosis of heart disease is a significant and tedious task in medicine. The detection of heart disease from various factors or symptoms is a multi-layered issue which is not free from false presumptions often accompanied by unpredictable effects. The massive amounts of data generated for prediction of heart disease which is too difficult and baggy to be processed and analysed by conventional methods. Data mining provides the methodology and technology to transform these data into useful information for decision making. Use of data mining algorithms will result in quick prediction of disease with high accuracy.

Keywords:-Data mining, traditional, methodology, technology, presumptions, algorithms.

1. INTRODUCTION

Data mining is a novel field for exploring the hidden information patterns from huge raw data sets. In a medical organization like hospitals and medical centres, generates a large amount of data which contains a wealth of hidden information, but these data are not used properly. Hence, that unused data can be converted into useful information by using different data mining techniques. In the modern world, cardiovascular diseases are the highest flying diseases and in every year more than 12 million deaths occur worldwide due to heart problems. Cardiovascular Diseases also cause maximum casualties in India and its diagnosis is a very complicated practice. Health Informatics is a rapidly growing field that is concerned with evolving Computer Science and Information Technology to medical and health data. Medical Data Mining is a domain of challenge which involves a lot of misdiagnosis and uncertainty.

Data Mining is about learning from existing real-world data rather than data generated particularly for the learning tasks. In Data Mining the data sets are large therefore efficiency and scalability of algorithms is important. As mentioned earlier the data from which data mining algorithms learn knowledge is already existing real-world data. Therefore, typically the data contains lots of missing values and noise and it is not static i.e. it is prone to updates. However, as the data is stored in databases efficient methods for data retrieval are available that can be used to make the algorithms more efficient. Also, Domain Knowledge in the form of integrity constraints is available that can be used to constrain the learning algorithms search space. This is also true in the engineering and medical fields. Data mining predicts the future of modeling. A general framework proposed for medical data mining is shown in Figure.1.


Data Set
Collection of raw data
Pre-processing and filtering data
Classification of filtered data
Classification technique
Remedy

1. Data Sets: A data set is a collection of data. Most commonly a data set corresponds to the contents of a single database table, or a single statistical data matrix, where every column of the table represents a particular variable, and each row corresponds to a given member of the data set in question.

2. Pre-processing and filtering data: *Data preprocessing* includes cleaning, Instance selection, normalization, transformation, feature extraction and selection, etc. Filtering of raw data. Filtering refers to the process of defining, detecting and correcting errors in raw data, in order to minimize the impact on succeeding analyses.

3. Classification of filtered data: A filter is a device or process that removes some unwanted components or features from a signal. Filtering is a class of signal processing, the defining feature of filters being the complete or partial suppression of some aspect of the signal. Most often, this means removing some frequencies or frequency bands. However, filters do not exclusively act in the frequency domain; especially in the field of image processing many other targets for filtering exist.
There are many different bases of classifying filters and these overlap in many different ways; there is no simple hierarchical classification. Filters may be:

1. Linear or non-linear

2. Time-invariant or time-variant, also known as shift invariance. If the filter operates in a spatial domain then the characterization is space invariance.

3. Causal or not-causal

4. Analog or digital

4. Classification techniques: There are different classification techniques. Some of these are as follows:

4.1. Naive Bayes: - Bayesian classifiers can predict class membership probabilities such as the probability that a given tuple belongs to a particular class. Naive Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of other attributes this assumption is called class conditional independence. The Bayes theorem is as follows: Let X={x1, x2……... xn} be a set of 'n' attributes. In Bayesian, X is considered as evidence and H is some hypothesis means, the data of X belongs to specific class C. We have to determine P (H|X), the probability that the hypothesis H holds given evidence i.e. data sample X. According to Bayes theorem the P (H|X) is expressed as

P (H|X) =P (X|H) P (H)/P(X)

4.2. WAC (Weighted Associated Classifier): Weighted Associative Classifier (WAC) is a new concept that uses Weighted Association Rule for classification. Weighted ARM uses Weighted Support and Confidence Framework to extract Association rule from data repository. The WAC has been proposed as a new Technique to get the significant rule instead of flooded with insignificant relation.

4.3. K-Nearest neighbor: This classification technique is called a memory-based technique, since the training samples should be stored in memory during runtime.

If a is the first sample denoted by (a1, a2,…, an), and b is the second sample denoted by (b1, b2,…, bn), the distance between them is calculated by relation 2-3.

$$\sqrt{(a_1-b_1)^2+(a_2-b_2)^2+(a_n-b_n)^2}$$

4.4. Decision tree: There are different types of decision trees. They only differ in the mathematical model they use to select the class of attribute during rule extraction. Gain ratio decision tree is the most common, successful type. It is a relationship between entropy (information gain) and classified information.

In entropy technique, the attribute which minimizes entropy and maximizes information gain is selected as the tree root. To select tree root, it is first necessary to calculate the information gain of each attribute. Then, the attribute maximizing information gain should be selected. Information gain, or entropy, is derived from relation.

$$E=-\sum_{i=1}^{k}p_i\,log_2^{p_i}$$

Where k is the number of response variable classes, pi is the ratio of the number of the i^{th} class events to total number of samples (occurrence probability of i)

4.5. Support vector machine: Given availability of support vectors, Support Vector Machine (SVM) is the boundary determining the best data classification and separation. In SVM, only those data lying inside support vectors are used as the base data for machine and building a model. This means that this algorithm is not sensitive to other data. It aims to find the best data boundary with the farthest possible distance from all classes (their support vectors). SVM transfers data to a new space with respect to their predetermined classes so that data can be classified and separated linearly (using hyper planes). Then, it searches for support lines (or support planes in multi-dimensional space) and tries to determine the equation of a straight line that maximizes the distance between each two classes. Each support vector is characterized with an equation describing the boundary line of each class.

2. LITERATURE REVIEW

This paper [1] authors had presented the feasibility study and the progress of heart disease classification embedded system. It provides a time diminution on electrocardiogram – ECG signal which can be practiced by decreasing the amount of data samples, without any significant loss. The objective of the urbanized system is the study of heart signals. The ECG signals are subjected onto the system that executes a preliminary filtering, and then utilizes a Gustafson–Kessel fuzzy clustering algorithm in order to exert for signal organization and correlation. The classification denotes usual heart diseases such as angina, myocardial infarction and coronary artery diseases. The system could also be used sudden "on duty" physicians, of any area of expertise, and could afford the first, or initial diagnose of any cardiopathy. If any system detects a heart problem, this system endows with better disease diagnose PPV evaluated to other testimonies, and therefore it tenders elevated assurance than other methods. Another foremost contemplation is the reality that this system was analogous to many other systems by accessing full data set, and this system exercised fuzzy clustering algorithm in order to diminish the data set, thus mitigating its use.

Data mining [2], as a resolution to haul out hidden pattern from the scientific dataset is projected to a database in this research. The database consists of 209 occurrences and 8 attributes. The system was employed in WEKA and MATLAB software and prophecy accuracy within Apriori algorithm in just 3 steps, are compared. MATLAB is pioneer as better performance software. Wide ranges of Apriori algorithms" sturdy system in data mining were evaluated to predict heart disease. A sole model consisting of one filter and appraisal methods are evolved. Three strong rules, as well as different estimation methods, are applied to find the superior software. Apriori rules are measured concerning their actual number of support, better accuracy, and considering strong rules. The high-performance software was introduced. The experiment can serve as a realistic tool for physicians to in effect predict uncertain cases and recommends consequently.

Authors in [3] presented a proficient advance for the forecast of heart attack from the heart disease database. Initially, the heart disease database is huddled using the K-means clustering algorithm, which will extort the data appropriate to heart attack from the database. This approach permits expertise the number of fragments through its k parameter. Consequently the frequent patterns are excavated from the extracted data, relevant to heart disease, using the MAFIA (Maximal Frequent Item set Algorithm) algorithm. The machine learning algorithm is modeled with the selected major patterns for the effectual prediction of heart attack. They have engaged the ID3 algorithm as the training algorithm to prove level of heart attack with the decision tree. The results showed that the designed prediction system is competent of forecasting the heart attack effectively.

In this paper [4] authors described about a prototype using data mining techniques mainly Naïve Bayes and WAC (Weighted Associated Classifier). The dataset is composed of important factors such as age, sex, diabetic, height, weight, blood pressure, cholesterol, fasting blood sugar, hypertension, disease. The system indicates whether patient had a risk of heart disease or not.

In this paper [5], authors proposed confidential scheme for predicting heart disease using two different models, Naive Bayes and Logistic Regression. As identified through survey, it is a need to have combinational approach to increase the accuracy of prediction for heart disease.

In this paper [6] authors proposed that, heart disease is one of the major causes of demise in the region of the world and it is essential to forecast the disease at a precipitate phase. The computer aided systems assists the doctor as a gizmo for forecasting and establishing heart disease. The intention of this paper is to extend about Heart related cardiovascular disease and to brief about accessible decision support systems for the computation and study of heart disease continued by data mining and hybrid intelligent techniques. Many DSS remains to predict the heart disease with several methodologies. The World life expectation statistics involve that heart disease has extended more in number. So it is essential to construct an efficient intelligent trusted automated system which predicts the heart disease precisely based on the symptoms according to gender/age and province knowledge of experts in the field at the lowest cost.

Authors in this paper [7] explicate that figures reveal that a heart disease is one of the foremost factors behind deaths throughout the world. Data mining techniques are pretty effectual in manipulative scientific support systems and having the capability to determine hidden patterns and relationships in medical data. Till now, Data mining classification techniques is applied to examine the various kinds of heart based problems. This paper is intended at mounting a heart disease prediction system using data mining clustering methods. This paper crews the various clustering techniques, k-mean, EM and the farthest first algorithm for the prophecy of heart disease. End result proves that farthest first clustering algorithm is the finest algorithm as evaluate to other algorithms. Since the ratio of correctly classified occurrences to the cluster is highest and the time taken to construct the model is minimum. This system can be further extended. More number of input attributes can be used and it can be further expanded by escalating the no. of the clusters. The same experiment can also be performed on other data mining tool such as R. And also the ensemble of classifiers can also be done to estimate their performance with the unique classifiers. Above algorithms can be subjected to other datasets in order to scrutinize whether the identical algorithm gives the highest precision or not.

Authors in this paper [8] proposed the incorporation of accessing a clustering approach and regression methodology. The clustering approach used is DBSCAN and for regression, multiclass logistic regression is subjected. By executing DBSCAN clustering algorithm, the entire dataset is fragmented into disjoint clusters. Resulted clusters were found to enclose fewer occurrences are then taken for consideration .These clusters are focused to multiclass logistic regression. This result is due to the clustering approach acquired by an unsupervised process. Once regression is achieved, we have accomplished at a termination, about actual variety of cardiac arrhythmia it is. The projected method accomplishes an overall accuracy of 80%, when evaluated with various other existing approaches. It projects a method for the prophecy of type of cardiac arrhythmia by assembling the use of DBSCAN clustering and multi class logistic regression algorithms. By balancing PCA-CRA with other methods, this method is found to be 80% accurate.

This paper [9] intends that large data existing from medical diagnosis is scrutinized by means of data mining tools and valuable information known as knowledge is hauling out. Mining is a method of investigating colossal sets of data to acquire the patterns which are hidden and formerly unknown associations and knowledge detection to facilitate the enhanced understanding of medical data to thwart heart disease. There are several DM techniques available namely Classification techniques concerning Naïve bayes (NB), Decision tree (DT), Neural network (NN), Genetic algorithm (GA), Artificial intelligence (AI) and Clustering algorithms like KNN, and Support vector machine (SVM). Numerous studies have been conceded out for mounting prophecy model by accessing entity technique and also by coalescing two or more techniques. This paper offers a rapid and simple evaluation and perceptive of obtainable prophecy models by means of data mining from 2004 to 2016. In this paper, a survey conducted from 2004 to 2015 gives the scheme of various models obtainable and the various data mining methodologies used. The exactness gained with these models is also specified. It is pragmatic that all the techniques accessible have not use big data analytics. Exploiting of big data analytics along with data mining will offer talented results to get the finest precision in manipulating the prophecy model.

Authors in this paper [10] recommends that heart disease is one of the diseases due to that fatality will occur mostly, and according to the world health organization the percentage is high for that. So Heart disease is determined for the big Data approach, and as Big Data is measured so use Hadoop Map diminish platform. For clustering improved K-Means and for the classification principle decision tree algorithm i.e. ID3 can be accessed in the hybrid approach. As second estimation is too better, the system is very helpful for the facilitating the forecast methods, based on the some restrictions like chest pain, cholesterol, age, resting Bp, Thalac and many more. Due to this system medical decision making will be enhanced as well as being rapid. It‟s also will impact on the humanizing the treatment process. In such way it will be very helpful in the prophecy of the heart disease. In such way authors had cultured about the big data and its properties, with its disputes and concerns. In the medical field the various parameters individuals are affecting to the heart. Improved K-Means is the algorithm which is viewing the precision in the centroid assortment more than the simple K-Means.

Authors in paper [11] intended that the medical doings examination plays a significant role in present trend. Discovery and study of medical doings is the most vital concern in real time scenario, since the requirement of training samples and adequate data‟s formulate these procedures much complex. This medical data analysis can be executed by effectual data mining methodology and advances. There are numerous unlike methods to diagnosis and prognosis diabetes mellitus. This paper had presented diverse techniques of the data mining methodologies to resolve the diabetes disease diagnosis problem. From the analysis, the discovery of several problems has been mentioned and locates in clinical datasets handling process.

In paper [12], a survey on a range of methodologies and algorithms for efficient classifier in premises of two issues such as class imbalance and dimensionality reduction has implemented. In the research it is noticed that numerous work categorizes under the class imbalance problem reduction and dimensionality reduction problems, but there is not a bit of the accesses were determined on both issues. So creating a classifier for the high-dimensional data with class imbalance problem will be a fascinating region for the future research. Numerous class imbalance problems are still not adequate for multiple class imbalance problems. This survey presents an outline on dimensionality reduction and class imbalance classification with the probable issues and outcomes. It portrays the major issues that slow down the classifier conduct due to these two problems.

Paper Number	**Techniques used**	**Advantages**	**Disadvantage**
1	Filtering process and fuzzy cluster algorithm.	Validates the principles of embedded system, and promotes the maximum possible efficiency.	Hardware limitations i.e., lack of memory management, absence of operational system.
2	Apriori algorithm applying WEKA and MATLAB software.	High performance which offers better accuracies.	Very slow and desires candidate generation every time.
3	K-means clustering algorithm with MAFIA method.	Easy to implement with a large number of variables,	Difficult to predict the number of clusters (K- Value), The order of

		K- Means may be computationally faster than hierarchical clustering.	the data has an impact on the final results.
4	Naïve Bayes and WAC (Weighted Associated Classifier).	Very simple and easy to use, highly scalable, make probabilistic predictions.	Features in the output class are independent, scarcity of data.
5	Logistic Regression.	Solve privacy violation problem with high accuracy.	Takes more speed and time for training and testing, discrete data will lead to some other problems.
6	Data mining algorithms such as Neural networks, naïve bayes, decision tree, genetic algorithm.	Effectual way to tackle the risks, integrates the strengths of various techniques.	Only 80% of accuracy can be achieved, lack of certain data security.
7	Clustering techniques such as k-mean, EM and the farthest first algorithm.	Automatic recovery from failure.	Complexity and inability to recover from database corruption.
8	Clustering approach with DBSCAN methodologies.	Different link connection is minimized, robust.	Not entirely deterministic, border points that are reachable from more than one cluster can be part of either cluster; quality depends upon the distance of the function region query.
9	Classification techniques involving Naïve bayes (NB), Decision tree (DT), Neural network (NN), Genetic algorithm (GA), Artificial intelligence (AI) and Clustering algorithms like KNN, and Support vector machine (SVM).	User friendly and scalable with 90% of accuracy.	All the techniques available have not used big data analytics.
10	Improved K-means and ID3.	Comparison of time complexity and accuracies is possible, speed and ease of use.	Sensitive to the selection of initial cluster centre, usually end without global optimal solution, but suboptimal solution; Sometimes the result of cluster may lose balance.
11	Support Vector Machine	Maximizes the prediction accuracy, avoids over-	Needs more tuning parameters and deep study is necessary, Can't

	(SVM).	fitting problem, classifies and diagnosis effectively.	perform using statistical analysis, Group attention selection process needs more attention.
12	Principal Component Analysis (PCA).	Low noise sensitivity, decreased requirements for capacity and memory, and increased efficiency.	Only be used if the original variables are correlated and homogeneous.

Table1: Review summary

3. CONCLUSION

In this paper, different research paper on prediction of diseases detection has been reviewed. Heart disease is the leading cause of death for both men and women. Know the warning signs and symptoms of a heart attack so that you can act fast if you or someone you know might be having a heart attack. This paper mainly focuses on the study of various approaches of heart attack disease prediction research papers are analyzed and studied. In survey we found different results of different techniques applied in the paper. This comparison helps in better future work. In survey we got to know about various disadvantages and benefits of techniques applied in different research papers. This helps in finding the better technique for future.

4. REFERENCES

[1]. de Carvalho Junior, Helton Hugo, et al. "A heart disease recognition embedded system with fuzzy cluster algorithm." Computer methods and programs in biomedicine 110.3 (2013): 447-454.

[2]. Mirmozaffari, Mirpouya, Alireza Alinezhad, and Azadeh Gilanpour. "Data Mining Apriori Algorithm for Heart Disease Prediction." Int'l Journal of Computing, Communications & Instrumentation Engg (IJCCIE) 4.1 (2017).

[3]. Khaing, Hnin Wint. "Data mining based fragmentation and prediction of medical data." Computer Research and Development (ICCRD), 2011 3rd International Conference on. Vol. 2. IEEE, 2011.

[4]. Patel, Ajad, Sonali Gandhi, Swetha Shetty, and Bhanu Tekwani. "Heart Disease Prediction Using Data Mining." (2017).

[5]. Wghmode, Mr Amol A., Mr Darpan Sawant, and Deven D. Ketkar. "Heart Disease Prediction Using Data mining Techniques." Heart Disease (2017).

[6]. Vijayashree, J., and N. Ch SrimanNarayanaIyengar. "Heart disease prediction system using data mining and hybrid intelligent techniques: A review." Int. J. Bio-Sci. Biotechnol 8 (2016): 139-148.

[7]. Singla, Meenu, and Kawaljeet Singh. "Heart Disease Prediction System using Data Mining Clustering Techniques."

[8]. Cp, Prathibhamol, Anjana Suresh, and Gopika Suresh. "Prediction of cardiac arrhythmia type using clustering and regression approach (P-CA-CRA)." Advances in Computing, Communications and Informatics (ICACCI), 2017 International Conference on. IEEE, 2017.

[9]. Banu, NK Salma, and Suma Swamy. "Prediction of heart disease at early stage using data mining and big data analytics: A survey." Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT), 016 International Conference on. IEEE, 2016.

[10]. Mane, Tejaswini U. "Smart heart disease prediction system using Improved K-means and ID3 on big data." Data Management, Analytics and Innovation (ICDMAI), 2017 International Conference on. IEEE, 2017.

[11]. Senthil Kumar, B., and Dr Gunavathi R. "A Survey on Data Mining Approaches to Diabetes Disease Diagnosis and Prognosis." IJARCCE 5 (2016): 463-467.

[12]. Kumar, B. Senthil, and R. Gunavathi. "Comparative and Analysis of Classification Problems." Journal of Network Communications and Emerging Technologies (JNCET) www. jncet. org 7.8 (2017).

[13] C. S. Dangare, S S. Apte Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques , International Journal of Computer Applications (0975 – 888) Volume 47– No.10, June 2012.

[14] J . Soni, U. Ansari, D. Sharma, S. Soni. Predictive data mining for medical diagnosis: An overview of heart disease prediction. International Journal of Computer Applications (0975 – 8887)Volume 17– No.8, March 2011

[15] K. Srinivas, B. Kavihta Rani, Dr A. Govrdhan. Application of data mining techniques in healthcare and prediction of heart attacks

[16] N .Aditya Sundar, P. Pushpa Latha, M. Rama Chandra, Performance analysis of classification data mining techniques over heart disease database, IJESAT volume-2, Issue-3, 470 – 478

[17] Nilakshi P. Waghulde, Nilima P.Patil, Genetic Neural Approach for heart disease prediction, International Journal of Advanced Computer Research, Volume-4 Number-3 Issue-16 September2