International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 12 | Dec 2018 www.irjet.net p-ISSN: 2395-0072

Comparative Analysis of Data Mining Classification Techniques for Heart Disease Prediction

Dhara Mehta[1], Nirali Varnagar[2]

[1]Student, Dept. of Computer Engineering, Atmiya Institute of Technology and Science, Rajkot, Gujarat, India
[2]Assistent Professor, Dept. of Computer Engineering, Atmiya Institute of Technology and Science, Rajkot, Gujarat, India

---

Abstract - *Data Mining is the most advanced technology in the field of computer science. It is used to discover the fruitful information from the large amount of data, by extracting the frequent patterns generating on database. Medical Science is also using the techniques of data mining. In medical field, data mining techniques can be used to predict the risk level of a particular disease. In this research work, we have compared the accuracy levels of data mining classifiers. We compared most trending classifiers like Naïve Bayes, SVM (Support Vector Machine), K-Star, J48 and Random Forest. In order to examine these classifiers, we have used Data Mining tool WEKA. We used the heart disease dataset of UCI repository. Among all the classifiers, we got the highest accuracy in SVM classifier by 84.07%. This paper contains the introduction of data mining technology, its techniques and classifiers used to predict the accuracy.*

***Key Words*: Data Mining, Heart Disease, Classification Techniques, WEKA, Predication.**

1. INTRODUCTION

The cardiovascular disease is the biggest cause of heart failure. Heart is an important constituent to live in human body. If heart does not work properly, it may cause the death of a patient. . High blood pressure, abnormal cholesterol, diabetes, smoking are the major causes for heart disease[3]. The doctors and experts generally use Electrocardiogram (ECG) for diagnosis of a patient. The hospitals and organizations keep the records of all the patients in the databases which remain unusable. But by applying data mining techniques to extract knowledgeable information can help in checkup and diagnosis of a patient. This ultimately helps in prevention and detection of cause of health disease.

Data Mining is one of the sensational fields of machine learning which extracts hidden information from the large databases [7]. So researchers can use data mining techniques in solving these kinds of problems. The database of any hospitals keeps records of patient's age, sex, Blood pressure, cholesterol, ECG result etc. By applying data mining classification techniques, we can predict the risk level of Heart Disease. In this paper, there are five classification techniques like Naïve Bayes, SVM, K-star, J48 and Random Forest are used to analyze the accuracy.

2. Knowledge Discovery from Database

KDD refers to the process of generating useful information from the large amount of database [1, 10]. KDD is a step by step process of deriving knowledge.

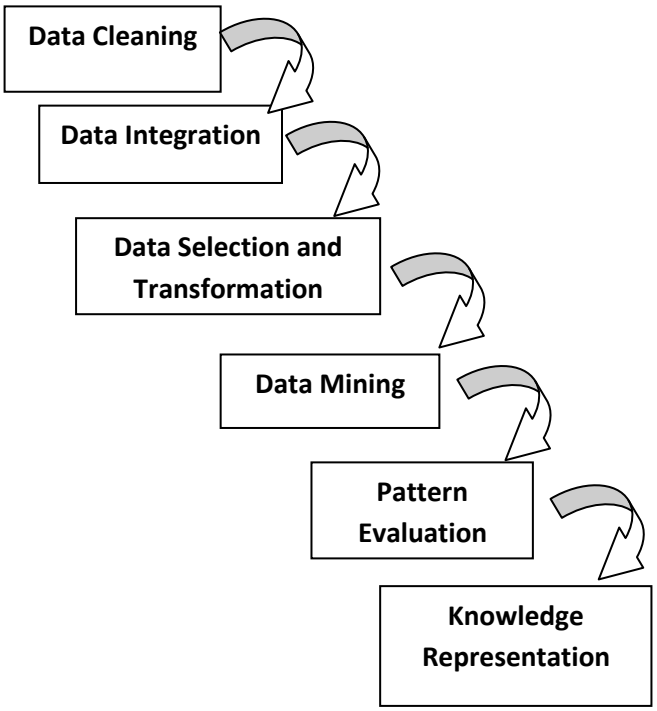

Fig -1: Procedure of KDD

2.1 Data Cleaning

It refers to the ability to understand and correct the data to get accurate analysis.

2.2 Data Integration

Data integration is the data preprocessing technique to merge the data from multiple data sources into a single database.

2.3 Data Selection and Transformation

Data selection is the process where the data necessary are selected from database. The data is to be transformed in

appropriate form in which data mining operations is going to be done. Sometimes, transformation process is performed before data selection.

2.4 Data Mining

It is an important phase where the data mining algorithms is applied on dataset to generate the pattern and knowledge.

2.5 Pattern Evaluation

The final step of data mining process includes deriving relevant information hidden in the dataset.

2.6 Knowledge Representation

Knowledge representation is about representing the knowledge retrieved from the data mining steps. It can be represented in terms of trees, tables, charts, matrix etc.

3. Data Mining Techniques

Data Mining techniques are use to get meaning information from the raw data. One of the most important steps in data mining is to choose the appropriate technique. There are many data mining techniques that can be used for different requirements. Some important techniques are listed below.

3.1 Association

Association is one of the data mining techniques which is use to unwrap the hidden relationship between unrelated data of relational database [1]. Association rule works on the basis of occurrence of one particular object based on occurrence of another object. It helps in analyzing the customer's behavior on shopping basket. Multilevel association rule, multidimensional association rule and quantitative association rule are the types of association.

3.2 Classification

Classification is a technique used to assign an item/items to the collection of target classes. The main goal of classification techniques is to predict accurately. The aim is to predict each class accurately for each and every class of data. In our case classification techniques is used to predict the risk level of heart disease by high, medium and low.

3.3 Clustering

Clustering is the process of grouping a set of objects into classes with similar objects. A cluster is a set of objects having similar characteristics to one another. It means the objects of one cluster can be treated as a whole. However it can be said as a data compression. Clustering analysis is a usual activity like as we learn to differentiate between cat and dog, by repeatedly improving schemes of clustering. Clustering can be done by different methods like partition, hierarchical agglomerative, density based, grid based and model based.

3.4 Regression

Regression and classification are used for prediction analysis but the difference is that regression is used to predict numeric and continuous values whereas classification can also be used for discrete data categories. The types of regression techniques are simple regression, multiple regression, hierarchical regression and set wise regression.

3.5 Prediction

Prediction is a data mining techniques about predicting the future behavior of an item based on current and past data to generate a prediction model. Types of prediction techniques are linear regression, multivariate linear regression.

4. Related Work

In this analysis task, we have used WEKA 3.8 tool for comparative analysis. WEKA contains machine learning algorithms for data classification. WEKA contains tools for classification, clustering and association. In this experiment there is an analysis for the accuracy of different classification algorithms and find the best suitable algorithm for heart disease prediction. We have used the Naïve Bayes, Support Vector Machine, K-Star, J48, Random Forest classifiers on the heart disease dataset.

The experiment has been conducted on the Heart disease dataset available on the UCI Repository. The dataset contains 270 instances and 14 attributes [5]. Table 1 shows number of attributes and its description available on UCI's Heart Disease dataset.

Table -1: Attributes Affected for Heart Disease

No.	Attribute Name	Type of Attribute	Attribute Description
1	**age**	Numerical	Age
2	**Sex**	Numerical	Sex
3	**chest**	Numerical	Chest pain type
4	**resting_blood_pressure**	Numerical	Resting blood pressure
5	**Serum_cholestoral**	Numerical	Serum Cholestoral in mg/dl
6	**Fasting_blood_s**	Numerical	Fasting blood sugar> 120

	ugar		mg/dl
7	**Resting_electrocardiographic_results**	Numerical	Resting electrocardiographic results(0,1,2)
8	**Maximum_heart_rate_achieved**	Numerical	Maximum heart rate achieved
9	**Exercise_induced_angina**	Numerical	Exercise induced angina
10	**Oldpeak**	Numerical	Oldpeak = ST depression induced by exercise relative to rest
11	**slope**	Numerical	The slope of the peak exercise ST segment
12	**Number_of_major_vessels**	Numerical	Number of major vessels(0-3) colored by flourosopt
13	**Thal**	Numerical	3 = normal; 6 = fixed defect; 7 = reversible defect
14	**class**	Nominal	class

4.1 Classification Techniques Used in Experiment

4.1.1 Naïve Bayes Classifiers

It is based on Bayes' theorem classification technique which having assumption of independence among predictors [3]. This technique assumes that the presence of one feature in a particular class is not related to the presence of any other feature. Naïve Bayes model is very easy and quick to build and can be used for a large dataset [11].

$$P(c/x) = \frac{p(x/c)p(c)}{p(x)}$$

$p(c/x)$: The posterior probability for *class* (c, *target*) given *predictor* (x, *attributes*).
$p(c)$: The prior probability of *class*.
p (x/c) : The probability for *predictor* given *class*.
$p(x)$: The prior probability of *predictor*.

4.1.2 Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning algorithm used for classification as well as regression. In this algorithm, we plot data item as a point in n-dimensional graph where each point is being a value of a particular coordinate. Then classification algorithm is being used to fine hyper-plane that differentiates 2 classes with maximum margin.

4.1.3 K-star

K-star algorithm work on the problem," Each new instance is being compared with the existing one by using a distance metric, and the closest instance will be used to assign a class to the newer one"[6].

4.1.4 J48

It is a one type of decision tree algorithm used as a classification algorithm. This algorithm is based on divide and conquers approach [3]. J48 divides the data into a sub range, based on existing attributes values available in training dataset.

4.1.5 Random Forest

Random forest is a supervised machine learning algorithm which creates forest and makes it random. The forest is an ensemble of decision trees, mostly trained with "training method" [12]. In simple word, Random forest algorithm generates multiple decision trees and merges them together to get the maximum accuracy and immutable prediction.

4.2 Performance Measures

4.2.1 Accuracy

Accuracy refers to the quality of prediction in data mining. Accuracy and be predicted by the formula:

$$\text{Accuracy}=\frac{Number\ of\ correct\ prediction}{total\ prediction}$$

$$\text{Accuracy}=\frac{(TP+TN)}{(TP+TN+FP+FN)}$$

4.2.2 ROC

ROC is receiver Operating Characteristics. ROC curve is a graphical representation of classifier's performance.

4.2.3 Kappa Statistics

The Kappa Statistics is used to compare the accuracy of the current system with the random system.

k= P (a)-P (e)/ (1-P (e))

P (a) = agreement percentage
P (e) = agreement chances

If the value of k is 0then it means that there is a chance of agreement. If the value of k is 1 then we can say that agreement is in a tolerable range of classifier.

4.2.4 RMSE

Root Mean Squared Error is defined as the difference between predicted value and observed value.

$$E_i=\sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(\frac{P_{(i-j)}-T_j}{T_j}\right)^2}$$

$P_{(i,j)}$ = predicted value
i = fitness
T_j = fitness applicability target value for j

4.3 Result Analysis

In this experiment, there are five classifiers are used in WEKA tool namely Naïve Bayes, Support Vector Machine, K-Star, J48 and Random Forest. The results generated by these classifiers are shown in table 2. The maximum accuracy is derived by the classifier Support Vector Machine in 0.24 seconds.

Table -2: Result Analysis from WEKA

Classifier	Accuracy	ROC	Kappa Statistic	RMSE	Time(sec)
Naïve Bayes	83.70	0.898	0.6683	0.3598	0.02
SVM	**84.07**	0.837	0.6762	0.3991	0.24
K Star	75.18	0.8939	0.4937	0.4438	0
J48	76.67	0.744	0.5271	0.4601	0.14
Random Forest	81.48	0.899	0.6244	0.3587	0.35

4.3.1 Classifiers Accuracy

Chart 1 shows the graph of accuracy for different classifiers. It shows that Naïve Bayes and SVM gives the best accuracy by 83.70% and 84.07% respectively.

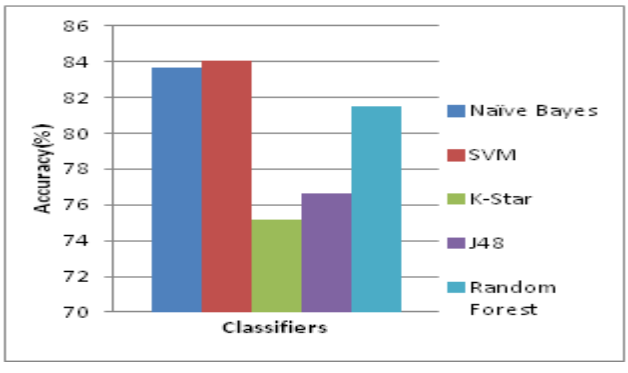

Chart -1: Classifiers Accuracy

4.3.2 ROC and Time

ROC curve shows good results for SVM and Random Forest whereas poor results for K-Star and J48. However the time taken by K-Star and J48 is lesser than SVM and Random Forest.

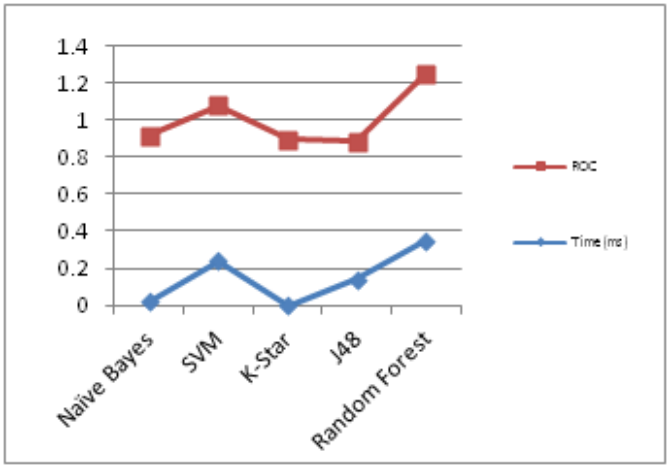

Chart -2: ROC and Time to build Model

4.3.3 K Statistic and RMSE

Chart 3 shows the better values of Kappa statistics for Naïve Bayes and SVM with lesser error as RMSE.

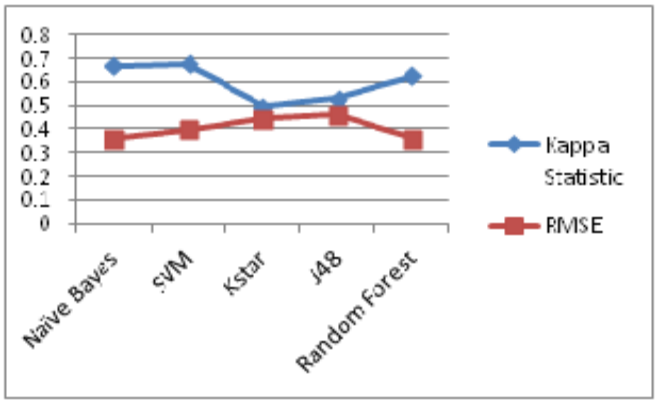

Chart -3: Kappa Statistic and RMSE

4.3.4 Accuracy Measures

Some other measures are also considered for prediction or classification. TP Rate, FP Rate, Precision, Recall and F-measure are majorly used measures.

TP Rate is the true positive rate which denotes the number of instances correctly classified and FP Rate is the false positive rate which denotes the number of instances that are wrongly classified.

Precision only considers positive cases as P=tp/ (tp+fp). Where tp denotes true positive rate and fp denotes false positive rate.

The recall represents the proportion of real positive values which is denoted by R=tp/ (tp+fn).

F-measure is the combination of precision and recall. F-measure =2*((precision*recall) / precision+recall).

Table -3: Major Values for Accuracy Measure

Classifier Name	TP Rate	FP Rate	Precision	Recall	F-Measure
Naïve Bayes	0.847	0.172	0.837	0.837	0.837
SVM	0.841	0.167	0.841	0.841	0.840
K-star	0.752	0.262	0.751	0.752	0.751
J48	0.767	0.240	0.766	0.767	0.767
Random Forest	0.815	0.191	0.815	0.815	0.815

5. CONCLUSION

In this research work, we have examined data mining classifiers on heart disease dataset. In comparison of 5 classifiers namely Naïve Bayes, SVM, K-Star, J48 and Random Forest; we got the highest accuracy by SVM classifier with 84.07%. The classifiers have been also compared by ROC, K-statistics and RMSE. In future better classifiers can be developed in order to get higher accuracy for more datasets of any particular disease.

REFERENCES

[1] Narender Kumar, Sabita Khatri, "Implementing WEKA for medical data classification and early disease prediction", In 3rd IEEE Conference on Computational Intelligence and Communication Technology (IEEE-CICT 2017)

[2] Meenal Saini, Niyati Baliyan, Vineeta Bassi, "Prediction of Heart Disease Severity with Hybrid Data Mining" 2017 2nd International Conference on Telecommunication and Networks(TEL-NET 2017)

[3] Theresa Princy. R, J. Thomas, "Human Heart Disease Prediction System using Data Mining Techniques" 2016 International Conference on Circuit, Power and Computing Technologies [ICCPCT]

[4] Salma Banu N.K and Suma Swamy, "Prediction of Heart Disease at early stage using Data Mining and Big Data Analytics : A Survey" 2016 International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques.

[5] Amita Malav, Kalyani Kadam and Pooja Kamat, "PREDICTION OF HEART DISEASE USING K-MEANS and Artificial NEURAL NETWORK as HYBRID APPROACH TO IMPROVE ACCURACY " International Journal of Engineering and Technology(IJET) Vol 9 No 4-Sep 2017

[6] Rashmi Madhukar Jadhav and Ms. Roshani Ade, "Analytical Approach to Predictive Disease Diagnosis using K-nn and Kstar" International Journal of Computer Science and Information Technology, Vol. 6(2), 2015, 1876-1879

[7] S. M. M. Hasan, A. Mamun, M. P. Uddin and M. A. Hossain, "Comparetive Analysis Of Classification Approach for Heart Disease Prediction" IEEE-2018

[8] K. Prasanna Lakshmi and Dr. C. R. K. Reddy, "Fast Rule-Based Heart Disease Prediction using Associative Classification Mining" IEEE International Conference on Computer, Communication and Control (IC4-2015)

[9] Mrs.S.Radhimeenakshi, "Classification and Prediction of Heart Disease Risk Using Data Mining Techniques of Support Vector Machine and Artificial Neural Network" IEEE-2016

[10] Hamidrza Ashrafi Esfahani, Morteza Ghazanfari, "Cardiovascular disease detection using a new ensemble classifier" 2017 IEEE International conference on knowledge-Based Engineering and Innovation (KBEI)

[11] M. A. Jabbar, Shirina Samreen, "Heart Disease Prediction System based on hidden naïve bayes classifier" IEEE-2016 International Conference on Circuits, Controls, Communications and Computing (I4C)

[12] Theresa Princy R and J. Thomas, "Human Heart Disease Prediction System using Data Mining Techniques" 2016 International Conference on Circuit, Power and Computing Technologies (IEEE-March 2016)