EasyChair Preprint

№ 3504

Comparative Analysis of Classification Algorithm on Heart Disease Dataset Using WEKA

Arslan Raza

EasyChair preprints are intended for rapid dissemination of research results and are integrated with the rest of EasyChair.

May 30, 2020

ComparativeAnalysis of Classification Algorithm on Heart Disease Dataset Using WEKA

Arslan Raza
Riphah International University

1-ABSTRACT

The huge World number of information is offered in science, business, industry, and many other domains. These statistics can provide vital information that may be used by management for making significant decisions. Using data mining we could discover valuable details. Data mining is a convenient subject among researchers. There's immance research that needs to be done and researchers found ease in data mining to do their research. However this paper concentrates on the basic idea of this Data mining that's Classification methods. The operation of the classifiers examined with the support of correctly classified instances, wrongly classified instances and time required to create the model and the end result can be revealed statistical in addition to graphically. WEKA the data mining tool is taken for this paper. The heart disease ratio is the leading ratio among death cases worldwide. It's hard to inspact the predection of this disease for medical experts as it's complex task that needs experience and knowledge. The health sector today contains hidden information that can be important in making decisions. Data mining algorithms such as Naïve Bayes, KStar, J48 and Random Forest are applied in this research for predicting heart disease. The research result shows prediction accuracy of 83%. Data mining enable the health sector to predict patterns in the dataset.

2-INTRODUCTION

Illness or disease leaves a bad impact on our lives. To die with disease or illness can down the moral of infected patients. Heart disease is one of them. In our daily lives, there are many people who are infected by this disease worldwide. Medical experts generate data with a wealth of hidden information present, and it's not properly being used effectively for predictions. For this purpose, the research converts the new data into a dataset for modeling using different data mining techniques.

Data mining is a technique to analyze data and process it into valuable information. This research intends to predict the probability of getting heart disease given the patient dataset. Predictions and descriptions are the principal goals of data mining. Prediction in data mining involves attributes or variables in the dataset to find an estimate and future value. In Data Mining different algorithms can be applied to find the estimated predicted results there are many ways to do like Classification, Clustering, Association. In this paper classification algorithms are applied to classify the dataset on attributes. This analysis will be helping out to predict heart attacks in patients.

3-LITRATURE REVIEW

The researchers [1] proposed a layered neuro-fuzzy approach to predict occurrences of coronary heart disease simulated in MATLAB tool. The implementation of the neuro-fuzzy integrated approach produced an error rate very low and a high work efficiency in performing analysis for coronary heart disease occurrences [1]. The researchers [5] also proposed a new approach for association rule mining based on sequence number and clustering transactional data set for heart disease predictions. The execution of the approch waas designed in C language and decerese RAM requirment by concedring the small cluster at a time in the result of scaleable and efficient.

Weka was developed at the University of Waikato in New Zealand, the name stands for Waikato Environment for Knowledge Analysis The system is written in Java and distributed under the terms of the GNU (General Public License). It runs on almost any platform and has been tested under Linux, Windows, and Macintosh operating systems and even on a personal digital assistant. It provides a constant interface to many different learning algorithms, along with methods for pre and post processing and for evaluating the result of learning schemes on any given dataset. Weka provides implementations of learning algorithms that can be easily apply to dataset. It also includes a variety of tools for transforming datasets, such as the algorithms.

4-RESEARCH METHODOLOGY

The purpose of the prediction methodology is to implement a model that can show characteristics of data from a combination of other data. The task of data mining in this research is to build models for prediction of class based on selected attributes. This research applies to four algorithms that are Naïve Bayes, KStar, J48, and Random Forest classifies and builds the model to diagnose heart attacks in the patient's dataset from medical experts.

5-PATIENT DATASET

The dataset is the collection of patient data that are infacetd from heart attack by different age groups. Heart attack patients are being dignosis with different age groups. Mostly this disease found in the the people having the age 40 and above.
There are 14 attributes in this dataset that are tested and trated results of the patients. The following vales in dataset are considered as nominal: Age, Sex(0 as women and 1 as men), Chest Pain, Blood Pressure Rate, Cholesterol, Smoking, Alcohol consumption and Blood Sugar Level.

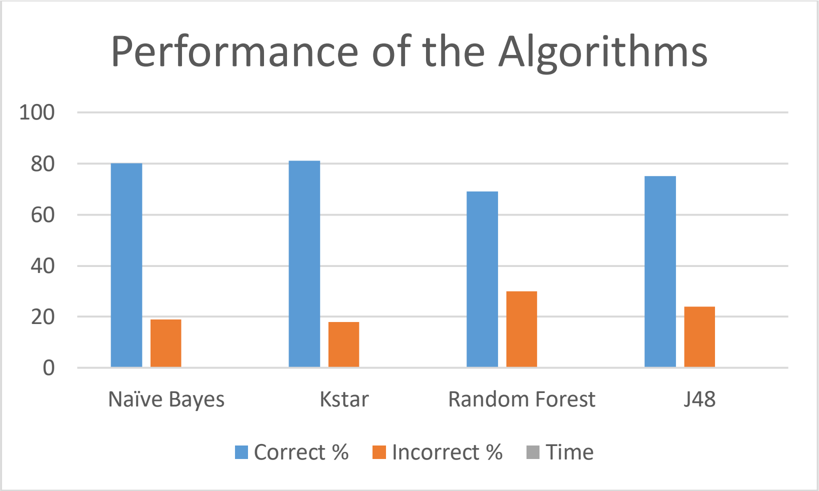
6-RESEARCH RESULTS

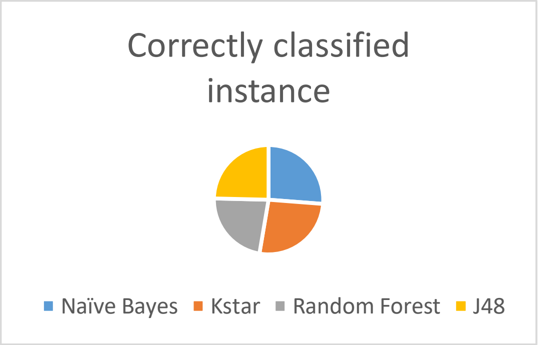
The algorithms are applied on the data set using stratified 10-fold validation in order to assess the performance of classification techniques for predicting a class.

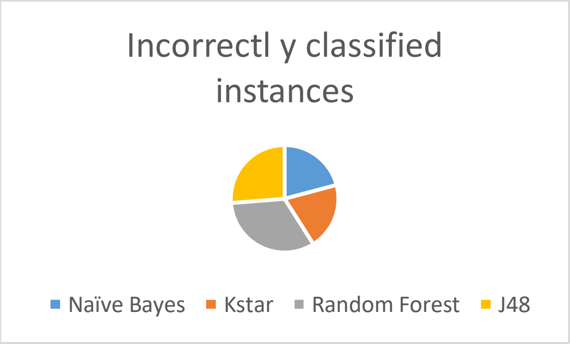
Evaluation Criteria	**Classifiers**			
	Naïve Bayes	KStar	Random Forest	J48
Correct %	80	81	69	75
Incorrect %	19	18	30	24
TP Rate	0.809	0.81	0.7	0.75
FP Rate	0.18	0.19	0.31	0.24
Time	0	0	0.09	0.02

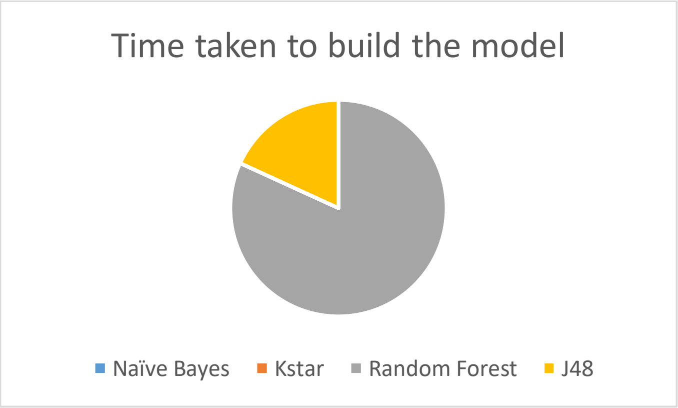
Accuracy Comparision of Classifiers

6.1-PREDICTED PERFORMANCE OF THE CLASSIFIERS









6.2-COMPARISON OF ESTIMATE

Evaluation Criteria	Classifiers			
	Naïve Bayes	KStar	Random Forest	J48
Kappa statistic	0.6123	0.63	0.3862	0.52
Mean absolute error	0.2043	0.23	0.4258	0.29
Root mean squared error	0.3776	0.37	0.4521	0.42
Relative absolute error	41.187	46.4	85.8304	58.9
Root relative squared error	75.8215	73.9	90.7756	84

6.3-CONFUSION MATRIX OF CLASSIFIERS

The confusion matrix consists of calculates the accuracy, sensitivity, and specificity measures [1]. The matrix denotes samples classified as true, others as false and others misclassified. Classification of the confusion matrix shows that J48, Random Forest, Naïve Bayes, and KStar show a prediction model of 354 cases with a risk factor positive for heart attacks. The techniques strongly suggest that data mining algorithms are able to predict a class for diagnoses. The confusion matrix clearly categorizes the accuracy of the model.

```
=== Confusion Matrix of Naïve Bayes ===

   a   b   <-- classified as

 105  33 |   a = 0

  25 140 |   b = 1
```

```
=== Confusion Matrix of  KStar ===

   a   b   <-- classified as

 106  32 |   a = 0

  24 141 |   b = 1
```

```
=== Confusion Matrix of J48 ===

   a   b   <-- classified as

 104  34 |   a = 0

  39 126 |   b = 1
```

```
=== Confusion Matrix of Random Forest ===

   a   b   <-- classified as

  81  57 |   a = 0

  34 131 |   b = 1
```

7-CONCLUSION

This research predict an experiment on application of different data mining algorithms to predict the heart diseases and compare the best algorithm. I applies the same experimental process as suggested by WEKA. The 75% data is used for traning and the rest for testing. In this tool all the resulults are based on instances and attributes. In the very first step the instances are partioned by the correcct and incorrect classification on the basis of numaric and percentage value that is shown in comparision of algorithms and the subsequently time taken is into next part. On the basis of comparision done on accuracy and FP rate of the algorithms the classification techniques with the highest accuracy obtained from KStar is 81%. The clearly shown in the above figure the highest accuracy that is obtained 81% and the lowest accuacy is 69% from Random forest. The best time taken by the algorithm is 0 sec and the worst time 0.09 sen. The best thing in weka is we do not need the depth knowledge of any algorithhm. This is the main reason the WEKA is suitable tool for data mining. This paper only shows the classification operation using WEKA.

8-REFERENCES

[1] Hlaudi Daniel Masethe, Mosima Anna Masethe, “Prediction System For Heart Disease Using Naive Bayes” in World Congress on Engineering and Computer Science.

[2] Ranjita kumari Dash, “Selection Of The Best Classifier From Different Datasets Using WEKA” in International Journal of Engineering Research & Technology (IJERT).

[3] Vikas Gupta, Prof. Devanand “A survey on Data Mining: Tools, Techniques, Applications, Trends and Issues” in International Journal of Scientific & Engineering Research(ISSN 2229-5518).

[4] Betim Cico, Vigan Raca, Rafet Duriqi “Comparative Analysis of Classification Algorithms on Three Different Datasets using WEKA” in 5th Mediterranean Conference on Embedded Computing.