Comparative Analysis of Classification Algorithms using WEKA

Sheetal D'souza
Department of Computer Science and Engineering
AIT College, Chikmagalur

Shainy merin S
Department of Computer Science and Engineering
AIT College, Chikmagalur

Karthikeyan S M
Asst.Prof. Department of Computer Science and Engineering
AIT College, Chikmagalur

***Abstract*- Data mining is the process of obtaining knowledge from the large quantity of data. Generally, it is the process of interpreting data from different perspectives and summarizing it into useful information. It analyzes the most useful and familiar classification algorithms used by the Machine learning system, mainly in the artificial intelligent systems. It is mainly concerned with practical use of classification algorithm that is applied to the dataset of diabetes. In this we are analysing and comparing different classification algorithm. This paper predict the performance evaluation based on the correct and incorrect classified instances of data using the Naive Bayes, Random Forest and K* algorithm. The outcome of the paper describes which algorithm is more effective for a particular dataset. The analysis of classification algorithm is done by WEKA tool.**

Keywords:-Dataset, Weka Analysis, Classification; Algorithms;

I. INTRODUCTION

Data mining is developing vastly in various fields. Data mining is adopted for several use and designed for different database [8]. The functions of data mining include some of the terms like classification, clustering, association rule extraction, regression and visualization [7]. Classification is a data mining function that assigns items in a collection to target categories, classes. The goal of classification is to accurately obtain the target class for each case in the data [8]. For the execution of classification algorithm we have used WEKA tool. WEKA is a general collection of machine learning software written in Java, developed at the University of Waikato in Newzeland. Weka is a workbench that contains a collection of visualization tools and algorithm for data analysis and predictive modelling.

Throughout the discussion we try to understand some of the tests, analysis of classification algorithms. We make use of Weka tool to implement the different classification algorithms for the dataset diabetes. Here, we discuss about the content of data and fields which are related to the dataset. The discussion is followed as first we discuss the nature of different classification algorithm. The next discussion is all about analysis of classification algorithm based on diabetes dataset. To analyze the diabetes datasets we use algorithms, which include Naive Bayes [4], Random Forest and K*.

The Naive Bayes algorithm is a probabilistic classifier [5]. The Random Forest algorithm deals with the decision tree [2] and the K* algorithm is path that either save the model that is trained or load an already trained and saved model [1]. Finally the research results based on the algorithms used in the Weka tool are discussed.

II. METHODOLOGY

We have used Weka tool for the analysis of three different classification algorithms. The algorithms can be directly applied to a dataset. It is also well suited for developing new machine learning schemes. The data that is used for Weka should be made into the arff format and the field should have the extension dot arff (.arff). The arff works with three sections they can be categorised into Relational section, Attribute section and the data. Whereas the data types of arff are classified into Nominal and Numeric.

A. *WEKA GUI Chooser*

The WEKA GUI Chooser is the starting point for running the applications. There will be a choice between the command line interface (CLI), the Experiment, the Explorer and Knowledge flow. In this context we make use of explorer that gives access to all features of weka using menu selection and form filling.
The figure1 shows the WEKA GUI chooser for commencing the main application.

Fig1: WEKA GUI chooser

B. *Diabetes dataset*

The Diabetes dataset were selected from the UCI ML repository. The performance of a comprehensive set of classification algorithms has been analyzed. The dataset contains 768 instances and 9 attributes. The attributes specify the properties of a patient. This dataset is mainly used to differentiate the tested results that are number of patients tested positive and tested negative. The obtained data helps us to predict whether the person is affected by the diabetes or not. Some of the attributes used are preg (how many times a women has been pregnant), plas (the concentration of glucose in the month), pres (the diastolic pressing of the blood), skin (the skin width), insu (insulin), mass (weight Kg), pedi (the diabetes based race), age (the age), class.

The figure2 shows the diabetes dataset opened in WEKA tool.

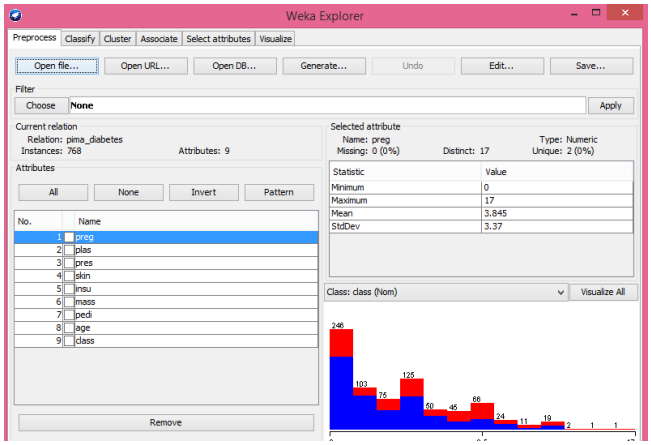

Fig2: Diabetes datasets open in WEKA

III. CLASSIFICATION ALGORITHMS USED

There are many classification algorithms available on WEKA tool but in our paper we have selected only three classification algorithms. Analysis of classification algorithm is the assembling of data in given classes.

A. *Naive Bayes*

The Naive Bayes algorithm is a simple probabilistic classifier that determines a set of possibilities by counting the constancy and combination of values in given data set [3]. This algorithm is also used in machine learning systems to conclude the new data or testing data, and it is based on the "Bayes" theory [4]. The application of this algorithm is performed by Weka tool, which provide opportunity to implement the above mentioned algorithm by using the estimator, for the numeric attributes and for using the Supervised Discretization to convert numeric attributes to the normal attributes [5].

B. *Random Forest*

Random Forest is a method for learning the classification and other tasks, which operate by developing a decision tree [2]. It contains some instances processed decision tree; otherwise this is a "forest" that contains some "trees" [6]. Weka tool which provides opportunity to this algorithm and it can be configured to improve the performance.

C. *K* Algorithm*

K* algorithm is a subdivision of "Lazy Learners" family [7].It use method of closeness neighbour with a distance vector function based on transformation. The implementation of K* algorithm in Weka tool is simple; it has a flag "entropic auto blend" to determine automatically global blend parameter. The K* algorithm can be faced with difficulties like phenomenon of Curse Dimensionality.

IV. DIABETES DATASET ANALYSIS

A. *Analysis using Naive Bayes Algorithm*

The initial analysis of this dataset is performed by default parameters that are provided by weka tool. The classification is performed by choosing "use training set" with training set of 66%, further with different testing options where 80% of data were used for training and remaining 20% for testing. Which specifies that out of 768 instances the correctly classified instances are 586 and incorrectly classified instances are 182. The final analysis is done by changing the parameter to "Supervised Descretization". The correctly classified instance is changed to 756 and incorrectly classified instance to 12. The table1 represents the Naive Bayes Algorithm for two different parameters.

Table1: Naive Bayes Algorithm table

Table1	Default parameter		Supervised descretize	
Correct instance	586	76.30%	756	98.43%
Wrong instance	182	23.69%	12	0.15%
Total instance	768	100%	768	100%

B. *Analysis using Random Forest algorithm*

As in the above used algorithm Naive Bayes, the Random Forest is performed in three different cases. The first option is used 10 fold cross validation. This algorithm specifies that out of 768 instances the correctly classified instances are 756 and incorrectly classified instances are 12. As in the previous case the final analysis is done by changing the parameter to "Supervised Descretization". The correctly classified instance is changed to 568 and incorrectly classified instance to 200. The table2 represents the Random Forest Algorithm for two different parameters.

Table2: Random Forest Algorithm table

Table1	Default parameter		Supervised descretize	
Correct instance	756	98.43%	568	73.95%
Wrong instance	12	1.56%	200	26.04%
Total instance	768	100%	768	100%

C. *K* Algorithm*

The default parameter will be applied continuing with Global Blend which has value of 20% by selection of 10 fold option. As the number of instances is more the execution time is also more but the time taken to build the model is 0s. Another parameter that strike enormously on the classification result is Global Blend setting. When the default parameter is changed to supervised descretize there is no change in the obtained value of the instances. The table3 represents the K* Algorithm for two different parameters.

Table3: K* Algorithm table

Table1	Default parameter		Supervised descretize	
Correct instance	531	69.14%	531	69.14%
Wrong instance	237	30.85%	237	30.85%
Total instance	768	100%	768	100%

V. CONCLUSION

This approach analyses the application of three different algorithms in diabetes dataset. The results obtained from three different algorithms are not similar the result differ according to the algorithms that are used. When Naive Bayes algorithm is used it has produced 76.3%, after the actuation of Supervised Descretize parameter the accuracy increased to 98.43%. Similarly for the K* algorithm the accuracy relatively increases from 69.14% to 100%. Whereas in Random Forest the accuracy is reduced after the actuation of supervised descretize parameter that is from 98.43% to 73.95%. Overall, the results indicate that the performance depends on the classification algorithms that are adopted. From the above obtained result we can conclude that Naive Bayes algorithm after the activation of Supervised Descretize parameter is best suited classification algorithm which gives 98.43% accurately, correct classified instance.

ACKNOWLEDGEMENT

We would like to thank Mr. Karthikeyan S M, Asst. Professor, Department of Computer Science and Technology, Adichunchanagiri Institute of Technology Chikmagalur for assistance we have received in writing this paper.

REFERENCES

[1] Thangaraju, R. Mehla, Analysis of KStar Classifier over Liver Disease, International Journal of Advanced Research in Computer science Engineering & Technology (IJARCET) Volume 4 Issue 7,July 2015
[2] Ho Tin Kham. “Random subspace Methods for Constructing Decision Forest”
[3] George Dimytoglou, James Adam, and Carol M. Jhim, “Comparisons of C4.5 and Naive Bayes Classification for the Analysis of Lung Cancer Survivabilities”
[4] Olivier C. Fhran, kois and Philip Leray Study of the Tree Augmented Naive Bayes Classification from deficient datasets LITIS. Sain-Etienne-De-Rouary, France.
[5] Jangtao Ron*,Sau Dan Le, Xianlo Chn , Ben Ka, Renold Chenk and David Cheunk Naive Bayer Classification of incalculable Data Department of Computer Engineering, Son Yaat-son University, Guangzhou, China
[6] Lio Bhreman, Jerom Fridman, Richerd Olshan, Charle Stone “Regression Tree” (Wardsworth).
[7] Ghopi Gandi, Rohith Shreevastav Modified k-methods algorithms for analysis and application to increase scalabilities and efficiencies for larger datasets.