A Comparison of Data Mining Classifiers in Weka

Shruti Shishir Gosavi
Assistant Professor (Computer Science)
Tilak Maharashtra Vidyapeeth, Pune, India

Shraddha Suhas Kavathekar
M.E. (Computer Engineering)
PCCOE, Pune, India

Abstract

Data mining is a process of extracting knowledge from a huge set of data. Data mining has three major components classification or clustering, association rules and sequential analysis. Classification is an important data mining technique with immense applications to classify the different kinds of data used in nearly every field of our life. Classification is a data mining (machine learning) technique used to identify group membership for data occasion. In this paper we present the basic classification techniques which are Naïve Bayes, Random Forest, J48. The aim of this study is to provide an interdisciplinary review of different classification techniques in data mining and use of variety of datasets such as iris, bank marketing, cancer, super market, weather and soybean for experimenting accuracy of divergent classifiers in weka.

Keywords- Classification, j48, Naive Bayes, Random forest classifier.

1. Introduction -

1.1 Data Mining–

Data mining is an approach to perceive interesting knowledge, such as associations, patterns, anomalies, changes and significant structures from tremendous amount of data stored in information repositories. Data mining is a major elevation in the type of analytical tools. Data mining is a multi-disciplinary field which is an integration of machine learning, statistics, database technology and artificial intelligence. This technique includes number of phases: Business understanding, Data understanding, Data preparation, Modelling, Evaluation, and Deployment. There are 5 data mining techniques such as Association, Classification, Clustering, Neural Network and Regression.

1.2 Classification -

Classification is used to stratify the item according to the features of the item with respect to the predefined set of classes. Classification is a data mining (machine learning) technique used to envisage group membership for data instances. A classification task begins with a data set in which the class assignments are known. For example, a classification model that predicts credit risk could be enrooted based on scrutinize data for many loan applicants over a period of time. We compared different classifiers with their accuracy and error rate.

1.3 Literature survey-

a. Naive Bayes-

The Naive Bayes algorithm is a simple contingency classifier that calculates a set of probabilities by counting the frequency and consolidation of values in a given data set. The algorithm uses Bayes theorem and estimate all attributes to be independent given the value of the class variable. The simulation as Naive yet the algorithm tends to perform well and learn speedily in various supervised classification. It performs different applications such as sentiment analysis, document categorization and email span filtering Naive Bayesian classifier are deployed on Bayes theorem and the theorem of total probability.

$$P(C|X) = \frac{P(X|C).\ P(C)}{P(X)}$$

Where P(C|X) is the posterior probability, P(C) is class prior probability, P(X) is predictor prior probability [2]

b. J48-

J48 classifier is a simple C4.5 decision tree for stratification. In the Weka tool, it is an open source java implementation of C4.5 algorithm. With this technique, a binary tree is composed to model the classification process. Once the tree is created, it is applied to each tuple in the database and results in categorization for that tuple. The auxiliary features of J48 are handling missing values, decision trees pruning, continuous attribute value ranges, extraction of rules, etc [1]

c. Random forest-

Random forest (or random forests) is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the class's output by individual trees, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean forecasting (regression) of the individual trees. It is one of the most accurate learning algorithms available. For many data sets, it produces a highly accurate classifier. It runs efficiently on large databases and can handle thousands of input variables without variable deletion. It gives estimates of what variables are important in the classification. It uses a Bagging approach to create a bunch of decision trees with random subset of data. The output of decision tress in the random forests is combined to make the final prediction. the final of the random forest algorithm is extracted - by surveying the results of each decision trees and just by going with prediction that appears the most times in decision trees.

2. Data sets-

We have used different datasets and tested in Weka to check their accuracy for different classifiers. Datasets were taken from UCI Machine Learning Repository.

2.1. Iris dataset-

This dataset consists of three different types of irises (Sentosa, versicolour, verginika) petal and sepal length stored in 150*4 numpy. ndarray. This is feasibly the best-known database to be found in the pattern recognition works. In the iris dataset the Number of attributes is 4, Number of instances are 150 and the Attribute characteristic of whether dataset is real. The iris dataset has highest accuracy of 95.53% for Naïve Bayes classifier, which is visualized in table.

2.2. Bank Marketing-

This dataset is taken from UCI Repository. The bank direct marketing data set contains (45211) number of instances with attributes without missing values, Dataset attributes are Multivariate, Attribute Characteristics are Real, Number of Attributes are 17[3]

2.3. Lung Cancer-

Lung cancer dataset was taken from UCI repository and explored in weka tool. The Number of Instances mare32, Number of Attributes are 57 (1 class attribute, 56 predictive) and attribute 1 is the class label. Attribute characteristic is Integer also the Missing Attribute Values for lung cancer dataset Attributes 5 and 39[4]. Accuracy for Lung cancer dataset is same for all classifiers except for random forest it showed 74.14%.

2.4. Super market dataset-

Super market dataset was explored in Weka. In this dataset the number of instances is 4627, and Number of Attributes are 217 also the Attribute characteristic is integer. The accuracy for supermarket dataset is same for all selected classifiers which is showed in further table.

2.5. Weather dataset-

Weather dataset was experimented in Weka. In the weather dataset the Number of attributes is 5, Number of instances are14 and the Attribute characteristic of whether dataset is nominal. The weather dataset has highest accuracy of 90% for Random forests classifier, which is visualized in table.

2.6. Soybean dataset-

In the soybean dataset the Number of attributes is 35, Number of instances are 47 and the Attribute characteristic of whether dataset is categorical. The iris dataset has highest accuracy of 93.32% for Random forest classifier, which is visualized in table.

3. Experimental Study-

Demonstration in this paper are based on forecasting functionalities provided by classification techniques. Classification is a data mining task that outlines the data into predefined groups and classes. First step after the dataset creation and loading in Weka pre-process panel is model construction. Model construction comprises of set of predetermined classes. Each tuple is assumed to belong to a predefined class. The set of tuples used for model construction is mentioned in training set. For model creation the most important is to choose best classification algorithm. On all training datasets same techniques were applied, and technique with best performance was selected for model creation. Classification results are presented in Table respectively. The model can be represented as classification rules, decision trees, or mathematical formulae. Created model is used for classifying future or unknown objects [5]. For all classifiers presented in the tables we perform 10-fold cross-validation, without percentage split.

3.1. Weka-

We have used Weka version 3-9-1.

WEKA is a comprehensive open source Machine Learning toolkit, written in Java. WEKA contains many inbuilt algorithms for data mining and machine learning. It is open source and freely available platform-independent software. The people who are not having much knowledge of data mining can also use this software very easily as it provides flexible facilities for scripting demonstration. As new algorithms appear in research literature, these are updated in software.

The steps performed for data mining in WEKA are:

- Data pre-processing and visualization
- Attribute selection
- Classification (Decision trees)
- Prediction (Nearest neighbour)
- Model evaluation
- Clustering (Cobweb, K-means)
- Association rules

4. Result and discussion-

We have studied different classifiers in weka and tested different datasets for above algorithms. The results show that classifiers give different accuracy for varied datasets due to their different features. Following tables described their accuracy and error rate. The result shows that Iris dataset works well with Naïve Bayes as compared to Random forests and J48. The overall observation of this paper gives, Accuracy for Naïve Bayes is highest 95.53% in comparison with other classifiers as showed in below table. This paper has explored different classifiers in weka tool.

The formula to calculate accuracy is:

1) Accuracy = $\frac{TP+TN}{P+N}$

2) Error Rate = $\frac{TP+TN}{P+N}$

In the equations above

(i) Accuracy represents Total Accuracy, TP is True Positive, TN is True Negative, FP is False Positive and FN is False Negative.

(ii) Error Rate: False number of occurrences in the dataset.

Table shows the comparison of different algorithms on the base of accuracy and error.

Table- 1.1 Measuring accuracy for datasets.

Following graph shows visualization of different datasets for different algorithms. Naïve Bayes Classifier has highest accuracy in Iris dataset. The results show that iris dataset gives maximum accuracy compared to others. Naive Bayes classifier resulted in higher accuracy which is 95.53%. We focused on iris datasets by comparing classifiers to see which classifiers gave maximum accuracy and minimum error rate as shown in the table 1.2. The results show that error rate for bayes_na is 5.02, and for J48 was 4.95, which is minimum amongst all.

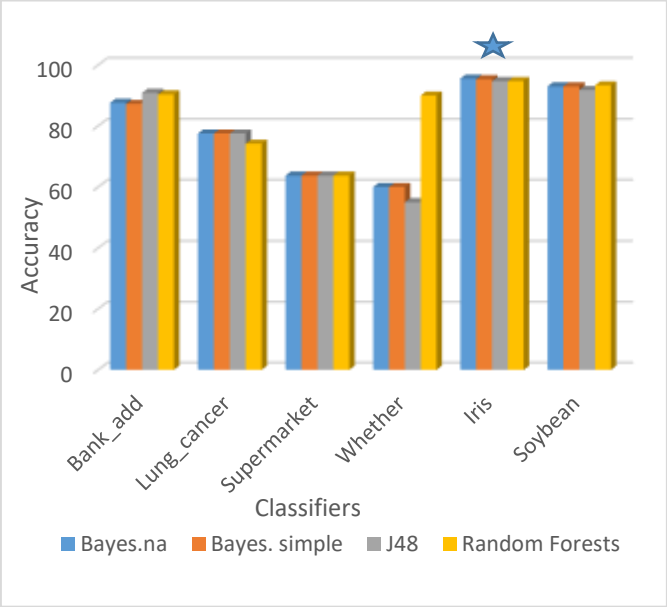
Datasets	Bayes.na	Bayes. simple	J48	Random Forests
Bank_add	87.62	87.21	90.87	90.36
Lung_cancer	77.5	77.5	77.5	74.17
Supermarket	63.71	63.71	63.71	63.71
Whether	60	60	55	90
Iris	95.53	95.33	94.73	94.73
Soybean	92.94	92.94	91.78	93.32

Table1.1- Comparing Accuracy for classifiers on varied datasets

Classifiers	Accuracy	Error Rate
Bayes.na	87.62	12.38
Bayes-simple	87.21	12.79
J48	90.87	9.12
Random-forest	90.36	9.12

Table1.2 - measuring accuracy for Bank dataset

Fig1.1-Accuracy for different classifiers



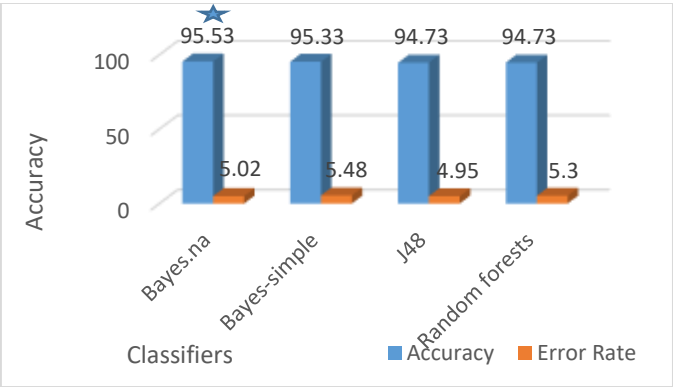


Figure 1.2- Comparing Accuracy and Error rate for Iris dataset

Classifiers	Accuracy	Error Rate
Bayes.na	95.53	5.02
Bayes-simple	95.33	5.48
J48	94.73	4.95
Random forests	94.73	5.30

Table1.3 - measuring accuracy for Iris dataset

5. Conclusion-

The results showed the classification performance of three different data mining techniques models - Naive Bayes, J48, Random Forest on different classifiers to see which classifiers give maximum accuracy. The classification performances of the four models have been using three statistical measures; Classification accuracy, sensitivity, error rate and specificity. Experimental results have shown the effectiveness of models. Naïve_Bayes has highest accuracy for Iris dataset 95.53%. we also calculated error_rate for different classifiers for iris dataset. we discovered that the two parameters accuracy and error rate for iris dataset gave different results for different classifiers. highest accuracy was given by naïve_bayes and lowest error rate was given by J48 4.95. Using weka tool we learned that due to different features of varied datasets the accuracy and error-rate changes for different classifers.

6. References -

1. Gaganjot Kaur Amit Chhabra, "Improved J48 Classification Algorithm for the Prediction of Diabetes", International Journal of Computer Applications (0975 – 8887) Volume 98 – No.22, July 2014
2. Guo, Yang, Guohua Bai, and Yan Hu. "Using Bayes Network for Prediction of Type-2 Diabetes." In Internet Technology and Secured Transactions, 2012 International Conference For, pp. 471-472. IEEE, 2012.
3. Hany A. Elsalamony, Helwan University, Cairo, "Bank Direct Marketing Analysis of Data Mining Techniques", Saudi Arabia International Journal of Computer Applications (0975 – 8887Volume 85 – No 7, January 2014
4. A. Floares., A. Birlutiu. "Decision Tree Models for Developing Molecular Classifiers for Cancer Diagnosis". WCCI 2012 IEEE World Congress on Computational Intelligence June, 10-15, 2012 - Brisbane, Australia.
5. Tina R. Patil, Mrs S. S. Sherekar "Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification", International Journal of Computer Science and Applications Vol. 6, No.2, Apr 2013 ISSN: 0974-1011.
6. Decision Tree Induction: An Approach for Data Classification Using AVL-Tree. Devi Prasad Bhukya1 and S. Ramachandram2.
7. Milos Ilic, Petar Spalevic and Mladen Veinovic, Wejdan Saed Alatresh, "Students' success prediction using Weka tool", in INFOTEH-JAHORINA Vol. 15, March 2016.
8. https://archive.ics.uci.edu/ml/machine-learning-databases/lung-cancer
9. https://archive.ics.uci.edu/ml/machine-learning-databases/00222