Web-Weka1.0: Online Tool for Comparison of Classification Algorithms of Weka

Gopala Krishna Murthy N
Dept. of CSE,
Grandhi Varalakshmi Venkatarao Institute of Technology (GVIT),
Bhimavaram, W G District, Andhra Pradesh, India - 534207

VSN Bhushana Rao
Dept. of CSE,
Grandhi Varalakshmi Venkatarao Institute of Technology (GVIT),
Bhimavaram, W G District, Andhra Pradesh, India - 534207

Naga Raju Orsu
Dept. of Computer Science,
Government Degree College,
Macherla, Guntur District, Andhra Pradesh, India - 522426

Ramya Chiluvuri
Dept. of CSE,
Grandhi Varalakshmi Venkatarao Institute of Technology (GVIT),
Bhimavaram, W G District, Andhra Pradesh, India - 534207

Suresh B. Mudunuri
Centre for Bioinformatics Research & Software Development (CBRSD)
GVIT, Bhimavaram, W G District, Andhra Pradesh, India - 534207
sureshverma@gmail.com

ABSTRACT

Machine learning and Data mining are becoming increasingly important in the recent years and have been successfully applied to solve a number of problems, especially in the areas of science and engineering. A variety of algorithms exist in the popular data mining tool 'Weka' to classify a given set of records into different classes. However, choosing the right classifier among them is a tricky task as the performance of a particular algorithm depends on various factors such as the application domain and the data set. In order to aid the researchers in comparing the performances of various classification algorithms in Weka, we have developed an online resource named Web-Weka using which one can compare a set of classification algorithms on a single data set on the fly and choose the right classifier for their study. The tool is available for free at www.mcr.org.in/webweka.

1. INTRODUCTION

Machine Learning is a branch of artificial intelligence that gives computers the ability to learn with out being explicitly programmed [1]. It provides the basis for the field Data Mining. Machine learning is a collection of tools and techniques used practically in the field of data mining to find hidden patterns in data. It has a wide range of applications including stock market analysis, natural language processing, speech recognition, bio-informatics, fraud detection, robotics, tumour classification, computer vision, etc. [2].

WEKA (Waikato Environment for Knowledge Analysis) is a machine learning tool kit developed by Waikato University of New Zealand [3]. Weka is an open-source software distributed under GNU Public License and is widely used by researchers, statisticians, and business analysts. The software is written in Java programming language and runs in Windows, Mac and Linux environments. Weka is equipped with different modules that perform various tasks of data mining that include data preprocessing, clustering, classification, association rule extraction, regression, visualization, etc. Of these, Classification is one important task that can solve a variety of problems [4]. The goal of classification is to rightly predict a class based on a set of attributes in the given dataset. Classification has many applications such as drug discovery, pattern recognition, speech recognition, biometric identification, natural language processing, etc. In order to make the best use of machine learning in practical data mining problems, it is necessary to understand how well the algorithms have been doing.

2. MOTIVATION FOR THE STUDY

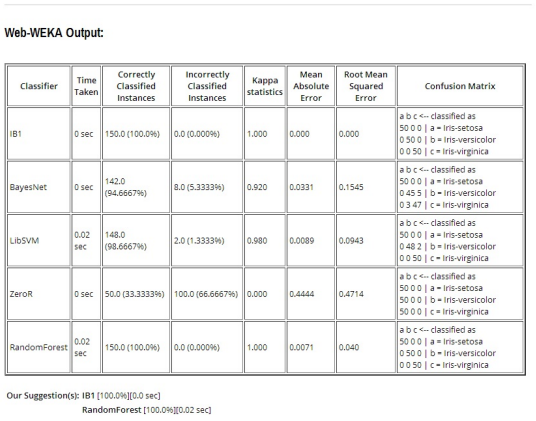
Many classification algorithms (also known as classifiers) have been proposed in the past, such as the neural networks, the decision trees, the linear discrimination analysis, the bayesian networks, etc. However, there is no single classifier that is superior over the other classifiers. Some of the methods work well only on two-class problems and are not extensible to multi-class problems. The performance of these classifiers differs according to the application domain and also due to the data set being used [5][6]. This makes the researchers to often get confused on choosing the right classifier for their study. In this paper, we tried to address this problem by developing an useful web application that compares the performance of various classification algorithms on their dataset.

The 19th International Conference on Management of Data (COMAD), 19th-21st Dec, 2013 at Ahmedabad, India.

In order to compare different classifiers, one needs to run Weka or other data mining tool on their dataset separately for each algorithm and manually compare the performances of these tools. Though Weka is being used widely through out the world, it is surprising that there is no availability of a web-based weka module that can perform the datamining tasks on the fly over web. As a part of our research, we have attempted to develop such a tool to perform one of the data-mining tasks over web.

3. WEB-WEKA

Web-Weka is a web-based tool developed to aid the naive users or beginners of machine learning and datamining to compare a set of classification algorithms on their small datasets. The classification module of 'Explorer' application of the GUI based Weka is mimicked (to an extent) in Web-Weka online tool. All the classification algorithms (classifiers) in Weka 3.6 have been categorized into different groups namely Bayes, Functions, Lazy, Meta, MI, Misc, Rules, and Trees. The users of Web-Weka can upload an Attribute-Relation File Format (ARFF - a standrard machine learning dataset format) file, set few training & test parameters and select a set of classification algorithms from the above groups. The tool will submit the ARFF file as input to each of the classifiers selected by the user and build a model each using the training and test options set. Finally, a summary table will be generated with the results of all the classifiers. A sample output summary table of Web-Weka1.0 can be found in Figure 1.



Web-WEKA Output:

Classifier	Time Taken	Correctly Classified Instances	Incorrectly Classified Instances	Kappa statistics	Mean Absolute Error	Root Mean Squared Error	Confusion Matrix
IB1	0 sec	150.0 (100.0%)	0.0 (0.000%)	1.000	0.000	0.000	a b c <-- classified as 50 0 0 | a = Iris-setosa 0 50 0 | b = Iris-versicolor 0 0 50 | c = Iris-virginica
BayesNet	0 sec	142.0 (94.6667%)	8.0 (5.3333%)	0.920	0.0331	0.1545	a b c <-- classified as 50 0 0 | a = Iris-setosa 0 45 5 | b = Iris-versicolor 0 3 47 | c = Iris-virginica
LibSVM	0.02 sec	148.0 (98.6667%)	2.0 (1.3333%)	0.980	0.0089	0.0943	a b c <-- classified as 50 0 0 | a = Iris-setosa 0 48 2 | b = Iris-versicolor 0 0 50 | c = Iris-virginica
ZeroR	0 sec	50.0 (33.3333%)	100.0 (66.6667%)	0.000	0.4444	0.4714	a b c <-- classified as 50 0 0 | a = Iris-setosa 50 0 0 | b = Iris-versicolor 50 0 0 | c = Iris-virginica
RandomForest	0.02 sec	150.0 (100.0%)	0.0 (0.000%)	1.000	0.0071	0.040	a b c <-- classified as 50 0 0 | a = Iris-setosa 0 50 0 | b = Iris-versicolor 0 0 50 | c = Iris-virginica

Our Suggestion(s): IB1 [100.0%][0.0 sec]
RandomForest [100.0%][0.02 sec]

Figure 1: Web-Weka Sample Output

The summary table contains the details such as Accuracy, Correctly/Incorrectly classified Instances, Time taken to build the model, Kappa Statistics, Mean Absolute Error, Root Mean Squared Error, and also the confusion matrix. Once the summary table is generated, the tool will provide a suggestion for the best possible classifier based on three parameters (in the order of their importance): 1) Better Accuracy of the model 2) Lesser Time taken to build the model 3) Lesser Error Values Reported.

User Options

The current version of Web-Weka allows the user to upload an ARFF input file of a maximum size of 5 MB and at most 5 different classification algorithms can be compared in a single run. As the tool is intended for beginners, we have provided with only a few options to change the default parameters. A preprocessing option has been provided to select only the important attributes and remove the unwanted ones from the dataset before building the model. Three different test options have been provided. The default testing option for classification has been set to 10 Folds Cross-Validation. However, the users can also have the choice to change this value. Moreover, the test options can also be set to 'use training set' that uses the same training set for testing also. An option to split the data set into training and test sets (based on user's choice of percentage) is also provided.

Technology Used

Web-Weka has been developed using the Java libraries of Weka version 3.6.6. The web-interface has been developed using HTML, CSS and Javascript. The server side scripts have been written in Java Server Pages (JSP). The Web-Weka has been hosted on a Linux Server with Apache Tomcat.

Advantages

As the tool is available online, it is platform independent and does not require installation of any specific software for using this tool. No login is required in order to use this tool. As many of the naive users are confused of what classifier is best suited for their study, this tool acts as a good resource to test their small datasets on various classification algorithms and choose the one that performs better.

4. FUTURE WORK

The Web-Weka 1.0 will further be extended to add more functionality and ability to compare more number of algorithms at once. The other forms of machine learning such as clustering, association mining, etc., are also to be added in the next versions of Web-Weka.

5. REFERENCES

[1] Ethem Alpaydin. *Introduction to machine learning*. MIT press, 2004.

[2] Usama M Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy. Advances in knowledge discovery and data mining. 1996.

[3] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.

[4] Donald Michie, DJ Spiegelhalter, CC Taylor, and John Campbell. Machine learning, neural and statistical classification. 1995.

[5] Eibe Frank, Mark Hall, Len Trigg, Geoffrey Holmes, and Ian H Witten. Data mining in bioinformatics using weka. *Bioinformatics*, 20(15):2479–2481, 2004.

[6] Gopala Krishna Murthy Nookala, Nagaraju Orsu, Bharath Kumar Pottumuthu, and Suresh B Mudunuri. Performance analysis and evaluation of different data mining algorithms used for cancer classification. *International Journal of Advanced Research in Artificial Intelligence*, 2(5):49–55, 2013.