Weka: A Tool for Data preprocessing, Classification, Ensemble, Clustering and Association Rule Mining

Shweta Srivastava
Assistant Professor (CSE Department)
ABES Engineering College, Ghaziabad

ABSTRACT
The basic principle of data mining is to analyze the data from different perspectives, classify it and recapitulate it. Data mining has become very popular in each and every application. Though we have large amount of data but we don't have useful information in every field. There are many data mining tools and software to facilitate us the useful information. This paper gives the fundamentals of data mining steps like preprocessing the data (removing the noisy data, replacing the missing values etc.), feature selection (to select the relevant features and removing the irrelevant and redundant features), classification and evaluation of different classifier models using WEKA tool. The WEKA tool is not useful for only one type of application, though it can be used in various applications. This tool consists of various algorithms for feature selection, classification and clustering as well.

Keywords
Weka, feature selection, classification, clustering, evaluation of classifier models, evaluation of cluster models.

1. INTRODUCTION
First of all why do we require data mining? We require data mining because of big data that is the term used for collection of large and complex datasets. A large data has several problems like capturing it, storing, searching, sharing and most importantly. All the industries have huge amount of data but they don't have appropriate knowledge extracting tools to get benefit out of it. There are various algorithms in data mining to help these industries for their better decision making. WEKA is a tool with capabilities of performing many data mining tasks such as data preprocessing, attribute selection, classification, clustering and improving the knowledge discovery using various meta classifiers. In this paper we'll discuss what is facilitated by the WEKA tool and what the steps to perform any activity on WEKA are. There are some constraints in Weka as it doesn't accept data in every format.

2. BACKGROUND STUDY
This section illustrates the brief insight into the data preprocessing, classification, clustering, ensemble techniques and their working in WEKA. This section is organized as follows: section 2.1 explains the introduction of Weka 3-6-9 interface 2.2 gives the overview of datasets, 2.3 gives the introduction of preprocessing, 2.4 describes the classification, prediction and ensemble techniques 2.5 briefs about clustering and 2.6 illustrates association techniques.

2.1 Using Weka Tool
Weka is a very good tool used for solving various purposes of data mining. There are four weka application interfaces: explorer, experimenter, knowledge flow and simple command line. The task can be processed using any of these interfaces. Not only can the interfaces, the open source code of weka also be used.

2.2 Datasets in WEKA
WEKA accepts the data in ARFF format that is attribute relation file format, CSV format that is comma separated values, . Though it can accept data in CSV format also and can be converted into ARFF format. ARFF file consists of: @RELATION <relation_name> gives the relation declaration. @ATTRIBUTE <attribute_name> <datatype> depicts attribute declaration showing the name of the attribute and its datatype. @DATA illustrates the data declaration that is the start of the data segment in file.

@DATA
5.1,3.5,1.4,0.2

Shows the data values means the values of each attribute in every sample. Data types can be numeric, nominal, string and date. Numeric, string and date are case insensitive. Data can be accepted from a database using JDBC connectivity.
In Weka, there is option of importing the data as well as generating it automatically.

2.3 Data Preprocessing
Preprocessing is one of the important and prerequisite step in data mining. Feature selection (FS) is a process to select features which are more informative but some features may be redundant, and others may be irrelevant and noisy [3]. When the data set consists of meaningless data that is incomplete (missing), noisy (outliers) and inconsistent data, preprocessing of the dataset is required. Preprocessing step includes:

(i) Data Cleaning: Handling the missing values by ignoring that particular tuple, filling that value with some specific value and handling noisy data using binning methods, clustering, combined human & machine inspection and regression. Inconsistency may be handled manually.

(ii) Data Integration: Sometimes we have data from various sources in data warehouse and we may require to combine them for further analysis. Schema integration and redundancy are major problems in data integration.

(iii) Data Transformation: Data Transformation is to transform the data in given format to required format for data mining. Normalization, smoothing, aggregation and generalization are few methods to perform transformation.

(iv) Data Reduction: Data analysis on huge amount of data takes a very long time. It can be performed using data cube aggregation, dimension reduction, data compression, numerosity reduction, discretization and concept hierarchy generation.

For the first 3 ways of preprocessing we have option of "filter" in WEKA. In filter option itself there are two types of

filters: Supervised and unsupervised. In both the categories we have filters for attributes and instances separately. After data cleaning, integration and transformation the data reduction is performed to get the task relevant data. For data reduction we have "Attribute Selection" option. It consists of various types of feature selection programs for wrapper approach, filter approach and embedded approach.

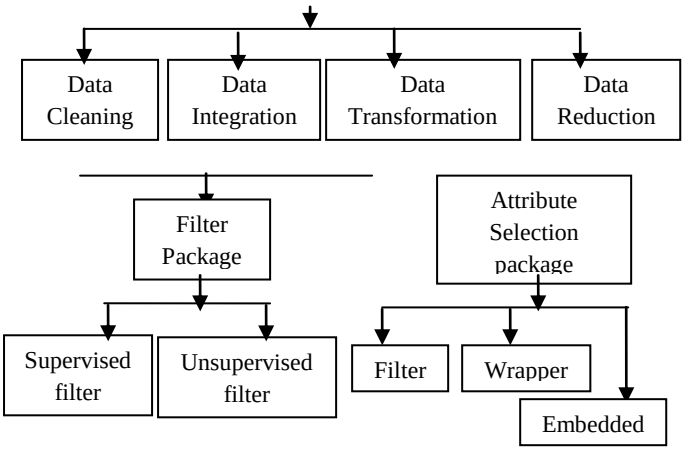

Fig.1. Data Preprocessing

2.4 Classification, Prediction and Ensemble techniques:

Classification is a data mining (machine learning) technique used to predict group membership for data instances [4]. It is the problem of finding the model for class attribute as a function of the values of other attributes and predicting accurate class assignment for test data. It can be divided in two types: supervised and unsupervised. Supervised is further divided in probabilistic and geometric. Probabilistic is further divided in parametric and nonparametric type. Classification is a two step process: first is model construction i.e. describing a set of predetermined classes and second is using that model for prediction i.e. classifying future or unknown objects.

For the Classification in Weka, we have supervised and unsupervised categories of classifiers. All the classifiers like lazy, tree, rules and naïve comes under these categories only. Meta classifiers are also there to enhance the accuracy of classifiers using various ensemblers.

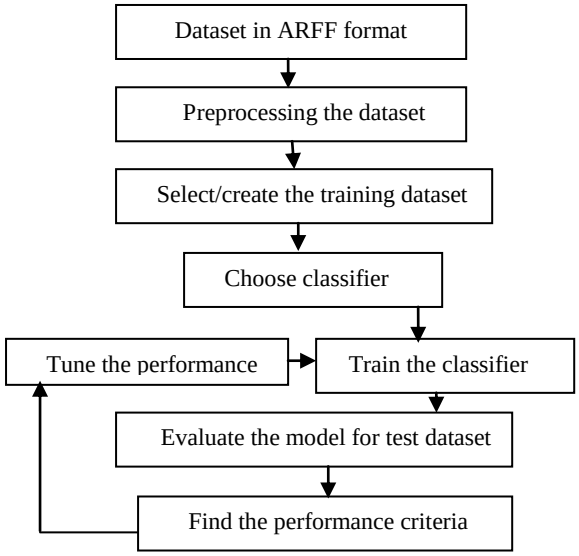

Fig.2. Classification steps

The performance criteria for evaluating the classifiers are: classification accuracy, specificity, sensitivity/recall, precision, AUROC curve, kappa statistics, mean absolute error, root mean squared error, Relative absolute error, root relative squared error, Time.

Classification accuracy: It is the ability to predict categorical class labels. This is the simplest scoring measure. It calculates the proportion of correctly classified instances.
Accuracy = (TP + TN) / (TP+TN+FP+FN)

Specificity: Specificity relates to the classifier's ability to identify negative results.
Specificity: TN / (TN+FP)
It is also called true negative rate.

Sensitivity/Recall: Sensitivity is the proportion of actual positives which are correctly identified as positives by the classifier.
Sensitivity = TP / (TP +FN)
It is also called true positive rate.

Precision: This is a measure of retrieved instances that are relevant.
Precision: TP/(TP+FP)

- True Positive (TP): If the instance is positive and it is classified as positive
- False Negative (FN): If the instance is positive but it is classified as negative
- True Negative (TN): If the instance is negative and it is classified as negative
- False Positive (FP): If the instance is negative but it is classified as positive

AUROC curve: It is the graph between false positive and true positive rate. The area measures discrimination, that is, the ability of the classifier to correctly classify the test data.

Kappa Statistics: The kappa measure of agreement is the ratio

$$K = \frac{P(A) - P(E)}{1 - P(E)}$$

where P(A) is the proportion of times the k raters agree i.e. percentage agreement between classifier and ground truth, and P(E) is the proportion of times the k raters are expected to agree by chance alone i.e. the chance agreement. K=1 indicates perfect agreement and K=0 indicates chance agreement. The value greater than 0 means classifier is doing better. Higher the kappa statistic value betters the classifier result. [6]

Mean absolute error (MAE): Mean absolute error is a calculation of how close predictions are to reality.

Root mean squared error:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(X_{obs,i} - X_{model,i})^2}{n}}$$

where X_{obs} is observed values and X_{model} is modeled values for tuple i and n is number of predictions. Small value of RMSE means better accuracy of model. So, minimum of RMSE & MAE better is prediction and accuracy. [7]

Relative absolute error:

$$\frac{|p_1 - a_1| + \ldots + |p_n - a_n|}{|\bar{a} - a_1| + \ldots + |\bar{a} - a_n|}$$

Where p_i is predicted target value and a_i is actual target value.

Root relative squared error: It is the square root of:

$$\frac{(p_1 - a_1)^2 + \ldots + (p_n - a_n)^2}{(\bar{a} - a_1)^2 + \ldots + (\bar{a} - a_n)^2}$$

Time: Time is the amount of time required to build the classifier model.

Where p_i is predicted target value and a_i is actual target value.

Steps to use classifier in weka: [1]

1. Open the weka explorer.

2. Load the dataset using either of the four options:

(a) Open file (b) Open URL (c) Open DB or (d) Generate

3. Some data processing steps can be performed using "filter" option. [Optional]

4. If attribute selection is required, go to select attributes menu where "attribute evaluator" and "search methods" can be chosen by their respective dropdowns. [Optional]

5. Go to menu "Classifier" and choose the classifier of your choice. There are four options for taking the dataset:

(a) Use training set, (b) Supplied test set (c) Cross validation folds and (d) Percentage split.

6. A classifier model and other classification parameters will be obtained for the training dataset. Now this classifier model can be used for the test dataset to evaluate the model. The prediction about the test data set can be summarized on the basis of various performance criteria's.

Ensemble techniques are used for improving the performance of classifiers. Also known as meta classifiers. It learns a set of classifiers and combines the predictions of multiple classifiers. With this approach, the classifier model can be improved and prediction strength can be enhanced. For using the ensemble in Weka, we need to choose option "meta" under the menu "Classifier".

2.5 Clustering

Clustering is used for finding the similar type of objects and group them together in a single cluster. Different objects are kept in different clusters. Clustering is used for unsupervised dataset but sometimes it can be used for supervised dataset also. Clustering can be done for samples as well as clustering of attributes can also be done. Cluster analysis is a difficult problem because many factors 1. Effective similarity measures, 2.criterion functions, 3. algorithms come into role for devising a well tuned clustering technique for a given clustering problems. [5]

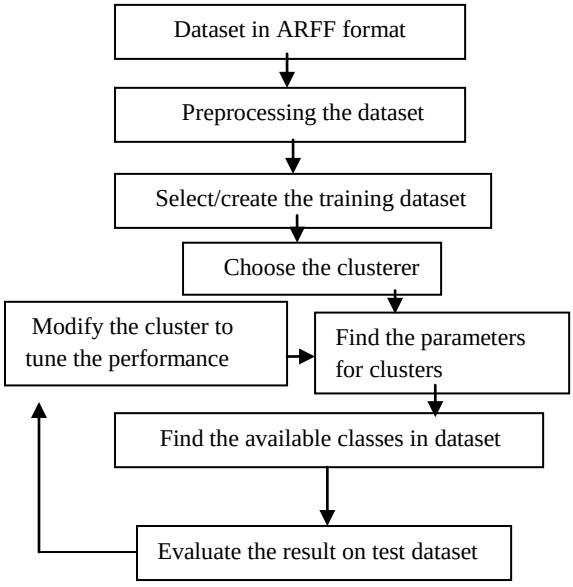

Fig.3. Clustering steps

Steps to use clustering in WEKA:[1]

1. Open the weka explorer.

2. Load the dataset using either of the four options:

(a) Open file (b) Open URL (c) Open DB or (d) Generate

3. Some data processing steps can be performed using "filter" option. [Optional]

4. If attribute selection is required, go to select attributes menu where "attribute evaluator" and "search methods" can be chosen by their respective dropdowns. [Optional]

5. Go to "Cluster" menu and select the clusterer of your choice. You can change the properties of the clusterer chosen such as number of clusters etc. by clicking on the clustering technique. An option of classes to clusters evaluation can be chosen if we have a class label in the dataset. Clusters can be stored for visualization.

6. A number of clusters will be obtained which shows the samples similarity in a single cluster and difference from other clusters.

2.6 Association Rule Mining Techniques

It is all about finding frequent patterns, associations, correlations or causal structures among sets of items or objects in transactional databases, relational databases, and other information repositories. The applications are market basket data analysis, cross-marketing, catalog design, loss-leader analysis, etc. Steps to use association techniques in WEKA:
Steps 1 to step 4 is common with classification and clustering. Next Choose the appropriate associator from "Associate" menu.

3. CONCLUSION AND FUTURE WORK

In this paper, working on WEKA tool is studied in detail. First we discuss about the data mining concepts and then different steps of data mining in Weka. The step involved in performing different concepts of data mining is discussed.

The API of WEKA is very useful as it can be used further to change any algorithm for some improvements.

4. REFERENCES

[1] http://weka.wikispaces.com/Use+WEKA+in+your+Java +code#Classification-Building a Classifier.

[2] Holmes, A. Donkin, I. H. Witten, WEKA: A Machine Learning Workbench, In Proceedings of the Second Australian and New Zealand Conference on Intelligent Information Systems, 357-361, 1994.

[3] C. Velayutham and K. Thangavel, "Unsupervised Quick Reduct Algorithm Using Rough Set Theory", JOURNAL OF ELECTRONIC SCIENCE AND TECHNOLOGY, VOL. 9, NO. 3, SEPTEMBER 2011

[4] Thair Nu Phyu, " Survey of classification techniques in data mining"; Proceedings of the International MultiConference of Engineers and Computer Scientists 2009 Vol I

[5] B.G.Obula Reddy[1], Dr. Maligela Ussenaiah[2] ,"Literature Survey On Clustering Techniques"; IOSR Journal of Computer Engineering (IOSRJCE) ISSN: 2278-0661 Volume 3, Issue 1 (July-Aug. 2012), PP 01-12

[6] Yugal kumar , G. Sahoo; "Study of Parametric Performance Evaluation of Machine Learning and Statistical Classifiers", I.J. Information Technology and Computer Science, 2013, 06, 57-64

[7] Yugal kumar , G. Sahoo; "Analysis of Parametric & Non Parametric Classifiers for Classification Technique using WEKA"; I.J. Information Technology and Computer Science, 2012, 7, 43-49