Data Mining Assignment 1

Ankita Kulkarni 672369586

February 14, 2016

I declare that all material in this assessment task is my own work except where there is clear acknowledgement or reference to the work of others and I have complied and agreed to the UIS Academic Integrity Policy at the University website URL: http://www.uis.edu/academicintegrity

Student Name: Ankita Kulkarni **UID**: akulk2 **Date**: January 22, 2016

Abstract:

Data Mining is the extraction of the useful information of the knowledge from the data sets so as to use it for further processing. The extracted data is used for the future development. In this report I have used a Data Mining tool named WEKA for the classification of the zoo and bolts data set where the main aim for this classification was to test the accuracy and performance of the different classifiers.

I. INTRODUCTION

Data Mining is basically the extraction of the quality and important information that is hidden within the large data sets that are worked upon in Data Mining. Nowadays, as the information sizes collected from different fields are exponentially expanding, information mining methods that concentrate data from huge measure of information have gotten to be mainstream in business and exploratory spaces, including showcasing, client relationship administration, and quality administration.

There are many tools that have been developed over time to carry of the process of the classification of data. One such tool is WEKA, which is used to analyse the data and prepare this report. To carry out the research and complete this report the arff data sets Zoo and Bolts were used. Mostly arff data sets have the numeric and nominal class attributes. The classifiers used to classify the data were: Decision Stump, OneR, Decision table, C4.5, PART, Linear Regression and M5. While working with these classifiers for this research report I observed and inferred the working and functionality of these classifiers. Decision stump is generally

referred as a weak learner and can make only one level of tree which means that it can only take one decision at a time. For classifying the zoo data set the Decision stump classifier classifies each animal in the zoo in a single step such that initially the zoo data set will be divided (classified) into two, where one part will be a single kind of animal say reptile and the second being all the other animals except reptile. Once the classification is completely done, it then ends up creating an individual leaf for all the animals in the zoo. OneR classifier generally works on the algorithm according to which the one rule is generated for each predictor and the rule with the total minimum error is selected as the one rule for the classification. It is displayed in the form of a pie chart which makes it easier to understand the results. The Decision table displays the data in a hierarchical view and consists of two attributes at each level. The Decision Table inducer finds out the most important attribute that is required for classifying the data. The representation of the results in this classifier is done is a form of a pie chart. The C4.5 classifier can handle missing values and does a continuous attribute categorization. The Linear Regression is generally used where a relationship between the predictor and the target can be displayed along a straight line. It usually works with numbers. The M5P model combines the decision tree with the chances of linear regression. It depicts the trees of the regression models.

The intention behind this research report is to analyse the working and performance of the classifiers on the zoo and bolds data set. Also, this report mentions how the data is actually classified. It displays the result of and accuracy of the classifiers and how they work on different data sets.

II. METHODOLOGY

Before starting with the research I read a number of articles to get familiar with the procedure how research is carried out. I read about the evaluation of the classifiers and their performance to get a clear knowledge of how the above mentioned classifiers actually work.

Initially starting with the zoo data set, I first observed the attributes and values of this data sets and then started with the classification of the data. Starting with the Decision stump - weka.classifiers.rules.JRi, I tested the JRi classifier for the Cross Validation test option for following different values of the folds:

No. of Folds	Correctly Classified Instances	Incorrectly classified Instances
10	87.1287%	12.8713%

17	86.1386%	13.8614%
29	89.1089%	10.8911%
41	85.1485%	14.8515%

After observing the above values I tested the JRi classifier for some more random values. To my conclusion I inferred that the JRi classifier for the Cross Validation test correctly classifies the instances to somewhere around 29/30 folds. The correctly classified instances decreases below 29 and above 30. Which states that it produces the best result near 30 folds.

I then tested the JRi classifier for the Percentage split test option for different values of the split percentage. To my notice I found that when the percent split is 87%, the classifier correctly classifies the instances to 92.307% which is the maximum value when the instances are correctly classified.

Percent Split	Correctly Classified Instances	Incorrectly classified Instances
33%	75%	25%
55%	80%	20%
70%	90%	10%
87%	92.307%	7.6923%

The next classifier used to carry out the research was the OneR - weka.classifiers.rules.OneR. OneR classifier has the worst performance for the zoo data set when compared to the different classifiers. It gives a rate of 42.5743% for the correctly classified instances and 57.4257% for the incorrect classification for all values of the folds. The OneR classifier does not work for the zoo data set. Even for the Percentage Split test option the OneR classifier has a poor performance rate. It gives a result where the instances are correctly classified around 45-55 percent, which is a poor performance.

Another classifier used to classify the zoo data set was the Decision table. When classifying the data set with the Cross Validation test the classification is carried out correctly to an extent of 88.1188% where the number of folds are 23. Below are some of the reading I tested:

No. of Folds	Correctly Classified Instances	Incorrectly classified Instances
10	86.1386%	13.8614%
23	88.1188%	11.8812%
33	84.1584%	15.8416%
70	86.1386%	13.8614%

The performance of the Decision table classifier is best for the percentage split of 85%. It gives a correctness of 93.33%. For Decision table, the test Percentage Split gives a better performance when compared to the Cross Validation test. I tested for some more random values. However the best performance of Decision table for the Cross validation test was for 23 folds.

When working with the C4.5 classifier, I observed that it gives the best classification result when tested for the Cross validation. The instances are correctly classified to 92.0792% for any value of the folds. However when the test for Percentage split is done the correct classification of the instances is done to 92.3077% for a split of 87%. The zoo data set when classified with the C4.5 classifier produces good results as the C4.5 classifier performs a continuous attribute classification.

Apart for the above mentioned classifiers I have also tested the performance of the PART classifier for the zoo data set. When implementing the Cross Validation test it was observed that the Correctly Classified Instances fluctuates between the values 92.0729% and 93.0693% for different values of the folds. The PART classifier was tested for the Cross Validation for the following values:

No. of Folds	Correctly Classified Instances	Incorrectly classified Instances
10	92.0729%	7.9208%
20	93.0693%	6.9307%
30	92.0729%	7.9208%
50	93.0693%	6.9307%

The PART classifier for the Percentage split test option for different values of the split percentage. And it was observed that for the percent split of 66%, the classifier correctly classifies the instances to 94.1178%, which is the maximum value when the instances are correctly classified. Below are the tested values:

Percent Split	Correctly Classified Instances	Incorrectly classified Instances
33%	80.8824%	19.1176%
66%	94.1178%	5.8824%
77%	91.3045%	8.6957%
88%	91.6667%	8.3333%

When the overall classification for the zoo data set was studied for the different classifiers and for the different test cases, it was concluded that the best classifier that correctly classifies the data set was the PART classifier tested for the Percentage Split test mode. Also the performance of OneR could be easily differentiated from the performance of the other classifiers as the result produced by the OneR classifier was very low. To my thought it was because OneR classifier generally works on a single attribute and it takes that attribute as the predictor which has the total minimum error value. However for the zoo data set the minimum error value was not calculated correctly and hence its performance was not god. The OneR classifier classified all the animals as mammals and hence it performs badly.

Another data set which was examined was the bolts data set. For this data set the classifiers that were examined were the KStar, Decision table, Linear Regression and M5P. The KStar provides an average (not so good) classification for the bolts dataset with a Correlation Coefficient of 0.5347. When the Correlation Coefficient is near 1, then the classification is more accurate. However when the Correlation Coefficient is near -1, then the classification is not accurate at all. The performance of KStart classifier is not good as it shows extreme values for different folds when using the Cross Validation test mode. The bolts data set was also tested with the classifier Decision table, which produced the Correlation Coefficient of 0.9204 for the test mode Percentage Split which shows that the data set was correctly classified. The

Correlation Coefficient being near to 1. For the Cross Validation test mode the Correlation Coefficient was 0.81, which is also near to 1 but the Performance of the Decision table is much better in the Percentage Split mode.

The next classifier whose performance was examined for the bolts data set was the Linear Regression. The Linear Regression classifier works for numeric values. It is displayed with a straight line between the predictor and the targeted value. It works best for the test mode Percentage Split, where the split is around 88% with a Correlation Coefficient of 0.90. However for the Cross Validation test mode the performance of the Linear Regression classifier was not good, with the Correlation Coefficient of only 0.64. The M5P classifier has the best performance for the bolts data set amongst the above mentioned classifiers. It produce a Correlation Coefficient of .95 for the Percentage Split test mode with split of 85%. However in the bolts data set the amount of data is less which is why in the Cross Validation test mode it cannot tested for more number of folds. The performance of the classifiers for this data set hence is not always good. The classification of the zoo and the bolts data set for the different classifiers does makes sense as the

III. RESULT

To examine the execution of Zoo and Bolts dataset on the given classifiers, few analyses were performed. For the classification of the zoo and the bolts dataset the classifiers were executed in the two test modes: Cross Validation and Percentage Split and led the analyses by changing the number of folds in Cross Validation and the Percentage split for all the given classifiers. The results were examined for all the classifiers and hence the performance was concluded. It was found that in Zoo dataset, the OneR classifier does not perform well as it miss classifies the other animals also as mammals which were not actually mammals. The performance of the PART classifier for the zoo data set was the more accurate when compared to the other classifiers. However when the bolts data set was classified the results were not as good as the bolts data set is small. The best performance for the bolts data set was of the M5P classifier with the Correlation Coefficient of 0.95 which is near to 1 and is hence more accurate.

IV. CONCLUSION

After examining the data sets and their classification using different classifiers I was able to conclude that there is a great variation in the accuracy of the classifiers for different test modes and that the accuracy of the classifier also depends on the size of the data set, the type of data

contained in the data set and the number of attributes in the data set. Also the tool WEKA used for the classification of the data set plays an important role and it a good tool for Data Mining.

V. REFERENCES:

1. Census Data Mining and Data Analysis using WEKA
   Link: http://arxiv.org/ftp/arxiv/papers/1310/1310.4647.pdf
2. Performance analysis of Data Mining algorithms in Weka
   Link: http://www.iosrjournals.org/iosr-jce/papers/Vol6-Issue3/E0633241.pdf
3. Website Link: http://datastage4you.blogspot.com/2014/04/performance-tunings-in-datastage.html