Data Mining Assignment 1

February 11, 2017

1 Task 1 - Experimental Research:

1.1 Abstract:

The main objective of Assignment 1 is to analyze Zoo, Bolt dataset by using WEKA Workbench for Task 1 and Task 2. Task 3 deals with the theoretical understanding of Data Mining and how data mining concept can be applied to solve the issues involved in Marketing Department and how it is helpful for a researcher to design a digital document search tool by using various techniques of data mining.

1.2 Introduction:

In this assignment I am using WEKA tool for data analysis. Weka stands for Waikato Environment for Knowledge Analysis. Weka is a collection of machine learning algorithm. We can apply these algorithms directly to a dataset or else can call them from a java code. Weka contains tool for pre-processing, classification, regression, clustering, association rules, and visualization [1]. In this assignment I am concentrating on OneR, PART, M5P, Linear Regression, Decision Stump and Decision Table algorithms for data analysis and comparing their results.

1.3 Methodology:

1.3.1 OneR:

It comes under weka.classifiers.rule.OneR. It uses the minimum-error attribute for predicting, representing numeric attributes. It is called as OneR, because **One attribute does all the work**. It contains 1-level decision tree and one branch for each values and each branch assigns most frequent class. The logic is selecting the class which occurs most frequently in the dataset. Then it is easy to determine the error rate. It selects the rule that has smallest total error as its **one-rule**. It treats missing values as another attribute [3].

1.3.2 DecisionStump:

It comes under weka.classifiers.trees.DecisionStump. It is used for building and using decision stump. It is a decision tree which has only one internal node i.e root node which is connected to terminal node. Mainly used with boosting algorithm. Missing values are treated as separate values.

1.3.3 DecisionTable:

Comes under weka.classifiers.rules.DecisionTable. Used for building simple decision table. -R parameter in decision table is for displaying rules. When it is set to 'true' it will display the decision table in classifier output window.

1.3.4 C4.5:

Comes under weka.classifiers.trees.J48. It can generate pruned or unpruned C4.5 decision tree. It constructs tree by using top-down approach. It choose a node as its rood node which contains highest information. It produces a tree and output which is easy for human understanding. Missing attribute values are not used for information gain [2].

1.3.5 PART:

Comes under weka.classifiers.rules.PART. As the name suggest it builds partial decision tree. Uses separate-and-conquer algorithm for tree construction. It uses C4.5 for building partial decision tree in each iteration and chooses the 'best' leaf as its rule.

1.3.6 Linear Regression:

Comes under weka.classifiers.functions.LinearRegression. Uses linear regression function for output prediction and builds regression model for given dataset.

1.3.7 M5P:

Comes under weka.classifiers.trees.M5P. Uses M5Base for generating trees and rules.

1.4 Results:

1.4.1 OneR:

aardvark ->mammal
bass ->fish
chicken ->bird
clam ->invertebrate. It will apply the same criterion for all 101 instances for deciding types of animals.

1.4.2 J48:

feathers = false AND milk = true: mammal (41.0)
feathers = false AND milk = false AND backbone = true AND fins = true: fish(13.0)
feathers = true: bird(20.0)
feathers = false AND milk = false AND backbone = true AND fins = false AND tail = false: amphibian(3.0)
feathers = false AND milk = false AND backbone = true AND fins = false AND tail = true: reptile(6.0/1.0)
feathers = false AND milk = false AND backbone = false AND airbone = true: insect(6.0)
feathers = false AND milk = false AND backbone = true AND airbone = false AND predator = true: invertebrate(8.0)

1.4.3 PART:

feathers = false AND milk = true: mammal (41.0)
feathers = true: bird(20.0)
aquatic = true: amphibian (3.0)
backbone = true AND tail = true: reptile(6.0/1.0)

fins = true: fish (13.0)
backbone = true AND airbone = false AND predator = true: invertebrate(8.0)

1.5 Discussions/Conclusions:

Zoo dataset contains 101 instances and 18 attributes. I have re-run zoo data for 5 times and every time it creates the same result with correctly classified instances as 42.57% and Incorrectly classified instances as 57.42% for OneR. It creates the same result every time since it is very simple. It produces the same result even after setting the minimum bucket size to 10. J48 produces better result than OneR. It produces correctly classified instances to 92.07% and Incorrectly classified instances to 7.92%. However, after running J48 with percentage split set to 80% I found decrease in Correctly classified instances to 90% whereas Incorrectly classified instances are slightly increased by 2% i.e. to 10%. With maximum percentage split its accuracy decreases. After re-running J48 by setting the seed value to 2,3,4,5,6 produces the same result. PART produces the same accuracy as that of J48 i.e Corr classified instances as 92.07% and Incorr classified instances as 7.92%. After running it with percentage split to 66%, however accuracy is increased with Corr classified instances to 94.11% and Incorr classified instances to 5.88%. While in PART by setting minNumObj option to 3 it produced the same result and when minNumObj are set to the values 4,5,6,7 I found gradual decrease in accuracy. Decision table produces accuracy of 86% with cross validation and when I switch the test option to percentage split I found decrease in accuracy to 76%. Similarly, DecisionStump produces accuracy of 60.39% which is less as compared to DecisionTable and after changing the option to percentage split I have noticed further decrease in accuracy to 55%.
Among all OneR produces less accurate results, because it uses one attribute for result calculation. In the Zoo dataset it has found animal class with highest number. So it has done its predictions based on animal class. It does not uses other criterion such as feathers, milk, backbone, legs, predator like J48 for comparison purpose. Thus, I believe it performs bad as compared to others.
For the animals that has been not used in dataset, I found decrease in accuracy in results.

2 Task 2:

2.1 Q1

The greatest effect on the time to count 20 bolts can be achieved by reducing the speed of rotation (SPEED1) of the plate at the bottom of the dish.

2.2 Q2

For a machine to get the shortest time to count 20 bolts can be achieved by deleting the value of T20BOLT with values which are lesser than 10 from the dataset.

2.3 NumericToNominal:

However, I noticed that Bolt dataset has Numeric type. To enable the classifiers OneR and DecisionTable we have to convert it to Nominal type. This can be done by applying filter NumericToNominal.

3 Task 3:

3.1 Data Mining:

From all the above experiments performed, data mining is about finding patterns in data and gaining knowledge from that data which is beneficial particularly for economic gain. It can be achieved through various techniques such as machine learning, artificial intelligence. As seen from Zoo dataset, all the algorithms, such as PART, J48, OneR are predicting the type of animal i.e they are getting the information from dataset such as whether the animal gives milk, has feathers, tail, backbone and based on these criterion determining the type of animals. This is nothing but discovering the patterns among data and processing it.

3.2 Marketing Department Issue:

Setting a ratio attribute for recording customer complaints is a good criterion. Since, ratio attribute allows us to store values such as count, temperature. First of all, we have to consider the total number of product sold by the company. Lets say the total number of product sold by company be TOTAL. Then, our job is to find out number of complaints received, lets call it as NUMCOMP. Then we have to compare TOTAL values with NUMCOMP. Suppose, assume that company sold five product of cosmetics lipstick, powder, cream, nail polish, and eyeliner with lipstick being the best selling product among all, then we have to find out that out of these five products which product is receiving the maximum number of complaints. Suppose company sold 10 nail polish and just 6 lipstick. If NUMCOMP received for nail polish is 4 and for lipstick is 4 then while calculating the percentage complaint received by the product lipstick will always show the higher percentage, since TOTAL for lipstick is less as compared to that of nail polish. To solve this problem, it is very important to set the value for minimum number of product to be sold. This will help in improving the accuracy and avoid misleading scenarios.

3.3 Digital Document Search Tool:

Digital document search tool allows user to retrieve any files either by entering full text, or by combination of various patterns such as any notes, properties etc. We can apply data mining techniques to improve efficiency of this tool as follows:

3.3.1 Clustering:

Clustering is a process of grouping objects based on their similarity. The objects belonging to one cluster have some similarity in their properties. By applying clustering to digital search tool we can keep all the records, objects which are having the same patterns into one cluster. For example, as with the zoo dataset we can create the cluster of all the mammal, birds or reptile into separate groups accordingly. But no two different clusters can have similarity. For example, there is no similarity between mammal and bird.

3.3.2 Classification:

This techniques predicts the output by classifying the given dataset. For example, in weather dataset, the output is predicted based upon classifying the conditions such as if sunny and not windy then play particular game or not.

3.3.3 Association:

Association works well even if no specific pattern or class is known. It can establish relationship with any attributes and predicts the output. For example, association is finding peoples buying pattern. If a customer is buying a pasta with sauce then it will associates a relation of buying pasta with sauce and next time providing special discounts or giving coupons on sauce to attract them.

3.3.4 Anomaly Detection:

It helps to detect any anomaly in the dataset. Therefore, it very important to detect these anomaly and handle them. If a file location is changed and if user tries to enter the old path then it will show invalid path, so anomaly detection keeps track of all such abnormal instances.

4 References:

[1] URL:
http://www.cs.waikato.ac.nz/ml/weka/
[2] URL:
https://en.wikipedia.org/wiki/C4.5_algorithm
[3] "Data Mining: Practical Machine Learning Tool And Techniques" - Third Edition by Ian H. Witten, Eibe Frank, Mark A Hall

Last updated: February 11, 2017

Typeset using TeXShop