Data Mining Assignment 3

1 Task 1:

1.1 Methodology for Weather Data (Nominal):

1.1.1 ID3:

It follows top-down approach for constructing decision tree. The process involves:

- Selecting attribute for root node and then creating branch for each possible attribute value
- Split the instances into subsets
- Procedure is repeated recursively, for only those instance who satisfies the conditions along the path from root node to the branch
- Stops when all instances belongs to the same class

Here, the attribute is selected which will produce smallest tree and which produces the 'purest nodes' i.e those who gives 'highest information gain' [1].

1.1.2 PRISM:

Commonly known as covering algorithm, because it covers all the instances in it and eliminating instances which are not present in the class. At every stage a rule is identified which will covers some of instances [1].

1.1.3 JRip:

Repeated Incremental Pruning to Produce Error Reduction (RIPPER) makes rules for classifying all class values except the majority ones. In this, last rule is usually the default rule for the majority class. It generates compact rule sets.

1.2 Results:

Table 1: Accuracy comparison of ID3, PRISM, JRIP on weather data (Nominal)

Weather Data (Nominal)		
Algorithms	10 fold-cross validation	Confusion Matrix
ID3	85.71%	$ID3 = \begin{bmatrix} 8 & 1 \\ 1 & 4 \end{bmatrix}$
PRISM	64.28%	$PRISM = \begin{bmatrix} 7 & 0 \\ 3 & 2 \end{bmatrix}$
JRip	64.28%	$JRIP = \begin{bmatrix} 7 & 2 \\ 3 & 2 \end{bmatrix}$

Table 2: Accuracy comparison of ID3, PRISM, JRIP on vote data

Vote Data		
Algorithms	10 fold-cross validation	Confusion Matrix
ID3	93.10%	$ID3 = \begin{bmatrix} 253 & 14 \\ 16 & 152 \end{bmatrix}$
PRISM	93.33%	$PRISM = \begin{bmatrix} 256 & 8 \\ 16 & 150 \end{bmatrix}$
JRip	96.09%	$JRIP = \begin{bmatrix} 257 & 10 \\ 7 & 161 \end{bmatrix}$

1.3 Discussion/Conclusion:

1.3.1 Weather Data:

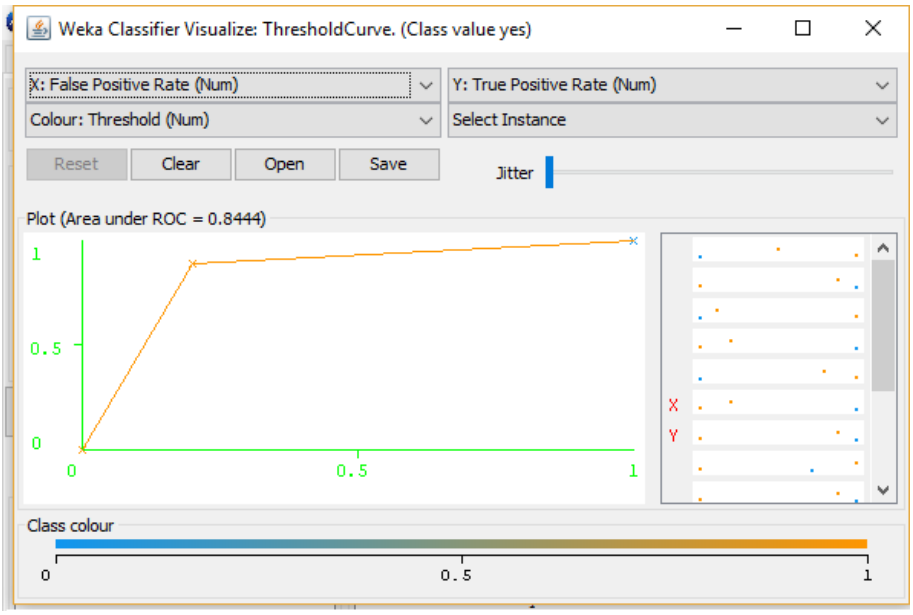

Figure 1: ROC of weather data with ID3

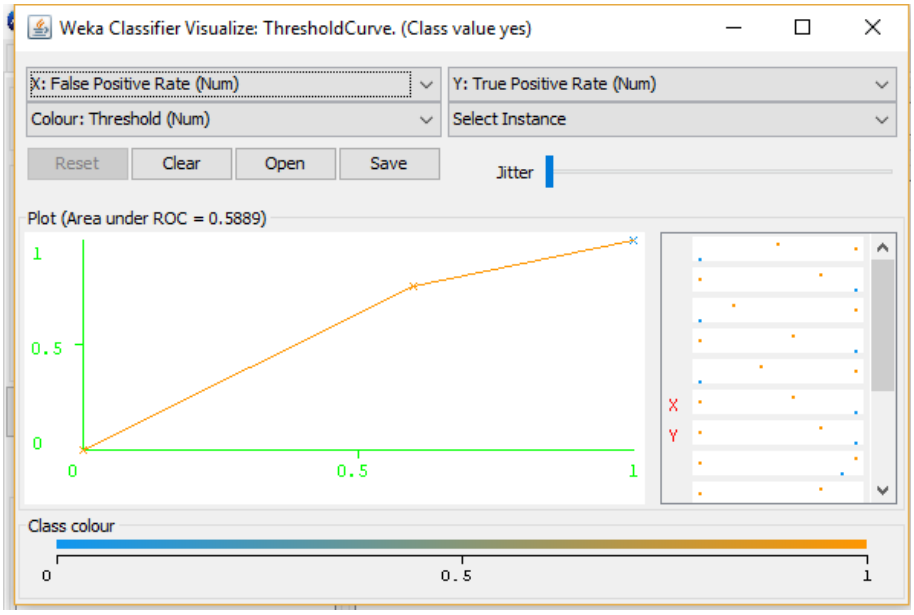

Figure 2: ROC of weather data with PRISM

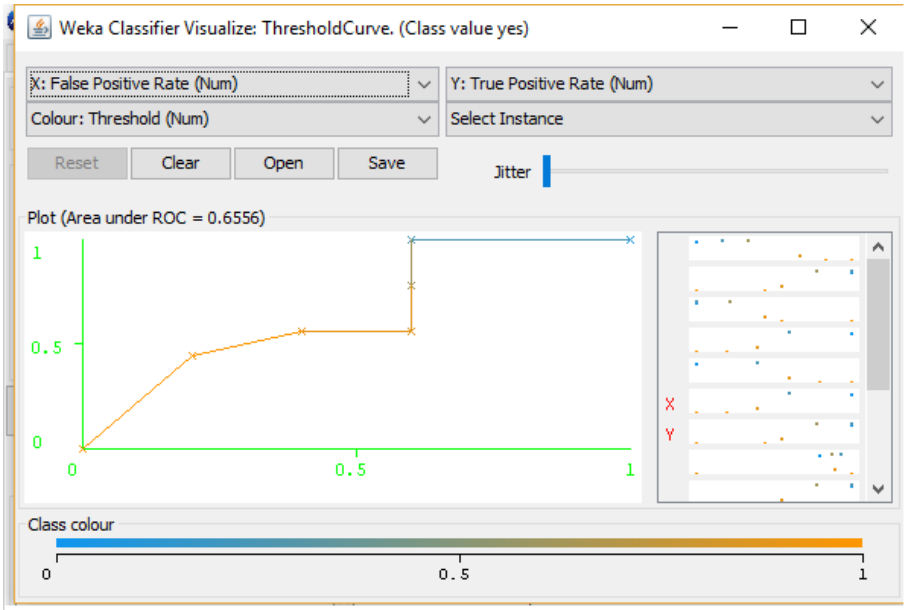

Figure 3: ROC of weather data with JRip

All the three classifiers work well on Weather data (Nominal). As seen in the Table 1, ID3 has the highest accuracy as compared to PRISM and JRIP in terms of correctly classified instances. I believe this is because, ID3 greedily selects the best attributes in terms of highest information gain to generate next branch. Because of this approach, ID3 has highest number of correctly classified instances which is equals to 12 with only 2 instances being incorrectly classified. PRISM classifies 9 instances correctly and 3 incorrectly. JRip also classifies 9 instances correctly but 5 instances are incorrectly classified. Same can be seen from the confusion matrix, ID3 has the minimum false positive (1,1) rate followed by PRISM as (0,3) while JRIP has the highest false positive rate (2,3) in this way decreasing its accuracy. On the other hand, major drawback of using ID3 is its decision tree approach which in turn increases redundancy. While rules; however, produces results which are easy to interpret. Among rules JRip produces more compact set of rules. Both ID3 and PRISM only works with nominal data. From ROC curve figures, it can be clearly seen that ID3 has the more accurate ROC curve followed by PRISM and JRip.

1.3.2 Vote Data:

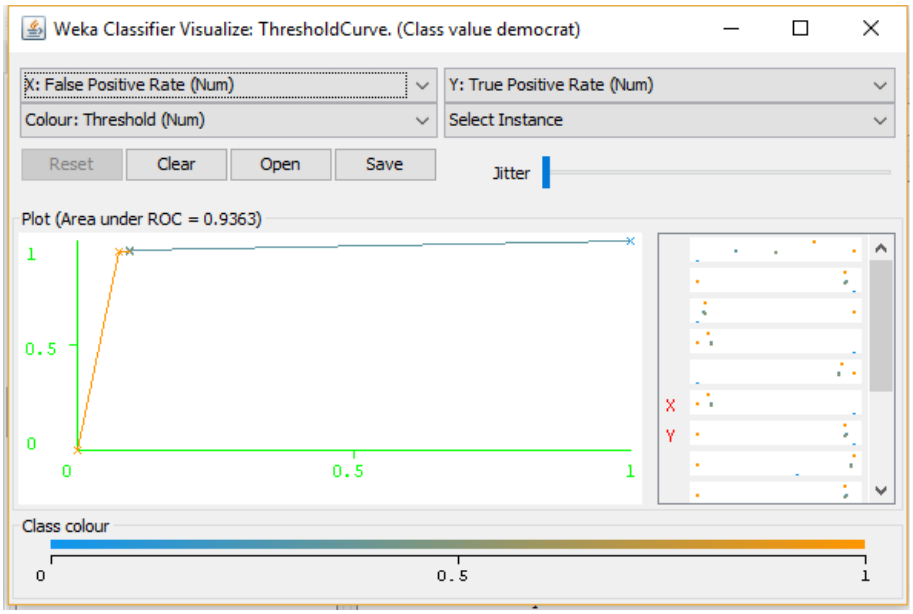

Figure 4: ROC of vote data with ID3

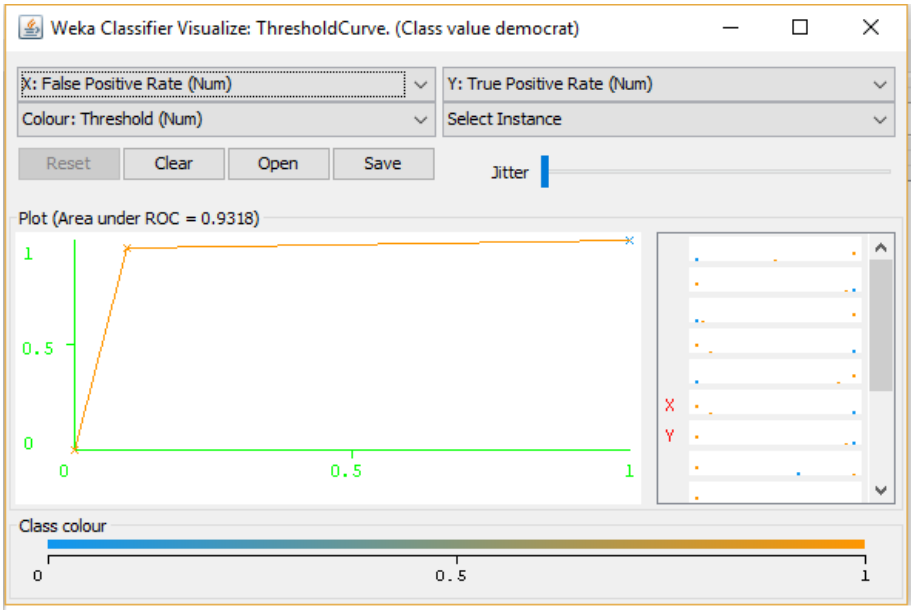

Figure 5: ROC of vote data with PRISM

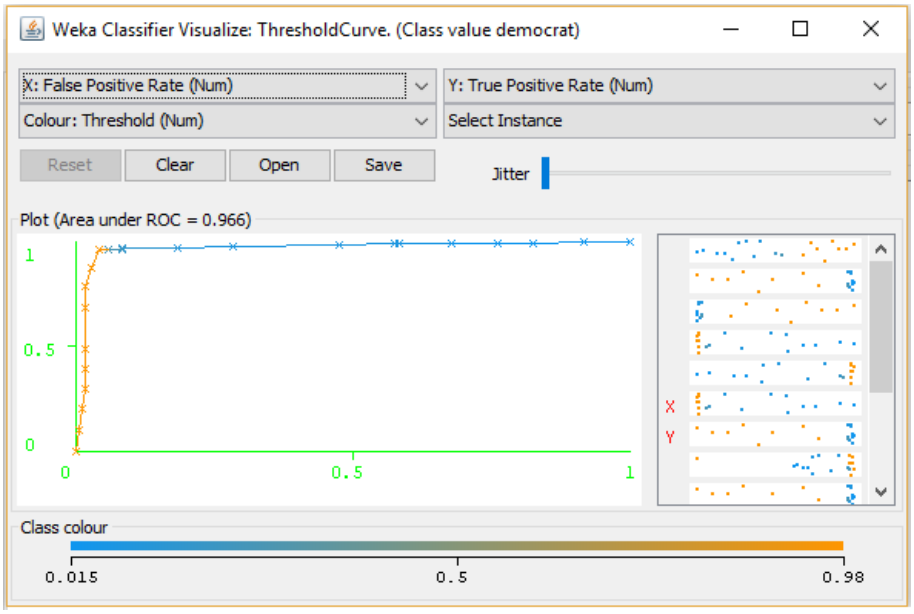

Figure 6: ROC of vote data with JRip

Classifiers ID3, PRISM do not works on vote data even though the data type is Nominal. The main reason for this is vote dataset contains Missing Values. To enable these classifiers with vote dataset I had applied **ReplaceMissingValue** filter. While JRip do not have any issues in handling missing values. It creates rules for missing values. As seen in Table 2, JRip has the highest accuracy followed by PRISM and ID3 which has lowest accuracy. I believe, the main reason for ID3's low accuracy here is the size of dataset. ID3 works well on small dataset. High information gain is not always the best strategy. JRip has the highest accuracy as well as it generates compact set of rules which are easy to interpret. JRip has the less number of incorrectly classified instances which is 17, then PRISM with 24 then ID3 with highest number equals to 30. The same observation can be verified with confusion matrix as well. Also we can see from ROC curves for all the three that JRip has the most accurate ROC curve followded by ROC curves for PRISM and ID3.

2 Task 2:

2.1 Methodology:

We can replace missing values in dataset by using weka's inbuilt ReplaceMissing-Value filter. ReplaceMissingValue filter replaces the missing value with attributes mean for Numeric data type and with Mode for Nominal data. So, I exported the breast-cancer.arff file to csv by using weka's ArffViewer. Attributes node-caps and breast-quad had missing values. I used **Imputation** method for replacing missing values with the mode of the attributes[2]. Mode is the value which repeated most often. Attribute node-cap has values as either yes or no. So for finding mode I used spreadsheets inbuilt function (**COUNTIF**) to count the total number of occurrence for no which is 222 times and for yes it is 56. As no has the maximum number of occurrence, so I selected no as mode for replacing all the missing values. In the similar manner, I chose breast-quad attribute's value, left_low which has the count of 111, while other attribute such as left_up, right_up, right_low and central has count equals to 97, 33, 24 and 21 respectively. Since, left_low has the highest count I chose left_low as mode value for breast-quad attribute.

2.2 Results:

Table 3: ID3 on Breast-Cancer data

Breast-Cancer Data		
Heuristic	With Filter	Without Filter
10 fold-cross validation	56.99%	56.64%

2.3 Discussion/Conclusion:

As we can see in Table 3, the accuracy with weka's inbuilt filter and with manual edition varies slightly but does not depicts much difference. However, I found accuracy with inbuilt filter slightly more. I believe the reason may be weka's inbuilt filter finds the mode values (default for missing values) more precisely than changing them manually. At the same time, I do not found much difference with accuracy in confusion matrix. However, ID3 performs bad on breast-cancer dataset. Being a tree classifier it leads to the overfitting of dataset. Again, high information-gain is not always a good approach.

3 Task 3:

3.1 Results:

Table 4: ID3 on soybean data

Soybean Data		
Split Criteria	Using Training Set	Supplied Test Set
60% for test set 40% for training set	99.75%	99.63%
70% for test set 30% for training set	99.79%	99.51%
75% for test set 25% for training set	99.80%	99.41%
80% for test set 20% for training set	99.81%	99.27%
90% for test set 10% for training set	99.83%	100%

3.2 Discussion/Conclusion:

As we can see in Table 4, ID3 performs fairly well on soybean dataset. I have applied RemovePercentage filter for dividing the dataset and divided the dataset with 40%, 30%, 25%, 20% and 10% split for training set and 60%, 70%, 75%, 80% and 90% split for test set. As we add more number of instances to training set we can clearly see increase in accuracy with test set. More number of instances in training set increases the ability of classifier to correctly classify instances.

4 Task 4:

4.1 Results:

Table 5: ID3 on soybean data with larger test set and smaller training set

Soybean Data		
No. of instances in training set	Accuracy of Classified Instances	Number of instances classified
68	Correctly: 19.02% Incorrectly: 80.97%	Correctly: 117 Incorrectly: 498
98	Correctly: 32.35% Incorrectly: 66.01%	Correctly: 199 Incorrectly: 406
128	Correctly: 46.05% Incorrectly: 41.47%	Correctly: 286 Incorrectly: 255
159	Correctly: 64.22% Incorrectly: 31.54%	Correctly: 395 Incorrectly: 194
190	Correctly: 76.09% Incorrectly: 19.02%	Correctly: 468 Incorrectly: 117

4.2 Discussion/Conclusion:

In the above experiment, I have divided the soybean data into smaller training set and larger test set. I applied 90% filter for training set leaving behind just 68 instances and 10% filter for test set with 615 instances. As seen in Table 5, the algorithm produces worst results and also trains large number of classifiers incorrectly. In the first experiment, ID3 had classified 498 out of 615 instances incorrectly thus by reducing its accuracy to 19.02%. In order to increase its accuracy, I added some of the incorrectly classified instances to training set. For finding which instances are miss-classified, I checked the Output Prediction option from Classifier Evaluation Options. Output prediction option displays all the instances in Classifier output panel and indicates error with + sign in front of those instances which are miss-classified. So I added 30 miss-classified instances to training set in each iteration. As the number of training instances increases so does the accuracy. As we have already seen from experiment performed in Task 3, more number of instances in training set increases the ability of classifier to correctly classify the instances. In the first iteration 498 instances were incorrectly classified. In 2nd iteration, when I

increased the number of training set to 98, 406 instances were incorrectly classified. This number was further decrease during 3rd and 4th iteration, when number of training instances were increased to 159 and 190 respectively. As the number of instances in training set increased, I also observer increase in number of correctly classified instances. So it can be clearly seen from the above experiment that the accurate decision trees (ID3) usually trained more quickly by iterative method than by training directly from the entire training data set.

5 References:

[1] "Data Mining: Practical Machine Learning Tool And Techniques" - Third Edition by Ian H. Witten, Eibe Frank, Mark A Hall
[2] https://pdfs.semanticscholar.org/28d0/a9cf62888120eddfd1c0014e1ba74261f01d.pdf