### ***Data Mining for Scientific Applications***

Course No. CSE-40770

***Laboratory Assignment III:***



***To download additional .arff data sets go to:***

<http://repository.seasr.org/Datasets/UCI/arff/>

or

<http://www.hakank.org/weka/>

**zoo.arff, wine.arff, soybean.arff, zoo2\_x.arff,**

**sunburn.arff, disease.arff**

**or**

**UCI data repository**

***You can find all of these files under the Resources section of the Blackboard as well!***

1. Use the following learning schemes to compare the training set and 10-fold stratified cross-validation scores of the disease data (in disease.arff):

|  |  |
| --- | --- |
| Decision table | - weka.classifiers.DecisionTable -R |
| C4.5 | - weka.classifiers.j48.J48 |
| Id3 | - weka.classifiers.Id3 |

1. What kind of information or indication does the training set evaluation score provide you?

|  |  |  |  |
| --- | --- | --- | --- |
| Under the PREPROCESS tab, attribute “class” was set as the class attribute using the EDIT-> “attribute as class” option. 10-fold cross validation divides the input data into 10 sets; 9 sets for training, 1 set for testing and repeat the process 10 times. Here are the training set results:   |  |  |  | | --- | --- | --- | | Decision Table  Screen Shot 2017-08-01 at 8.03.14 PM.png | J48  Screen Shot 2017-08-01 at 3.33.44 PM.png | ID3 (unpruned J48)  Screen Shot 2017-08-01 at 3.34.18 PM.png |     Evaluation of the training set results mainly tells us about the rules used and the time it takes to build the model. Here, we see that decision table uses the nuclei attribute as a rule, and J48 and ID3 also use the nuclei attribute as the rule for classification. |

1. What does the cross-validation score evaluate?

|  |
| --- |
| The decision table cross-validation result is shown as an example:  Screen Shot 2017-08-01 at 3.37.13 PM.png  Cross-validation results provide more information about the performance of the model:   * correctly and incorrectly classified instances * kappa statistic which essentially compares the observed accuracy with what may be expected by random chance. * different types of error measures like mean absolute error, relative absolute error, etc. that are more for evaluating numeric predictions. * confusion matrix which shows the true positives, false positives, true negatives and false negatives * There are also TP rate, FP rate, Precision, Recall, F-measure that are derived from calculations based on the numbers in the confusion matrix. * MCC evaluates how well the model is predicting with 1 indicating perfect prediction, 0 random, and -1 complete disagreement between prediction and true observation. * area under the ROC and PRC that look at the accuracy of the classifier. It seems PRC values are more sensitive to how many instances are in each class or data balance? * we have these above mentioned measures for each class and as a weighted average for the overall model. |

C) Which one of these models would you say is the best? Why?

|  |
| --- |
| The cross-validation results for J48, pruned and unpruned (ID3) are the same:  Screen Shot 2017-08-01 at 4.01.36 PM.png  This is what the tree looks like:  Screen Shot 2017-08-01 at 4.02.32 PM.png  Both models used the nuclei (=1 or =2) as the rule to classify the normal and manic disease classes.  None of the models are good at correctly classifying the normal class. Using the experimenter to perform t-test on the weighted average F-measure, precision, recall, and ROC area, no significant differences were found among the three models. I feel we need more samples to help us obtain a better and more meaningful model. However, if I really had to choose, I would pick the decision table model because the weighted average F-measure, ROC area, precision and recall are slightly higher than the other two models. |

1. Use the following learning schemes to analyze the wine data (in wine.arff).

|  |  |
| --- | --- |
| C4.5 | - weka.classifiers.j48.J48 |
| Decision List | - weka.classifiers.PART |

1. What is the most important descriptor (attribute) in wine.arff?

|  |
| --- |
| I used 10-fold cross validations for both classifiers. The numeric attributes in the wine dataset were discretized using filter -> supervised -> attribute -> discretize. The “OD280/OD315\_of\_diluted\_wines” attribute is the most important descriptor because it is at the top in both models. |

1. How well were these two schemas able to learn the patterns in the dataset? How would you quantify your answer?

|  |
| --- |
| **Decision list cross-validation result:**  Screen Shot 2017-08-14 at 9.13.48 PM.png  **C4.5 cross-validation result:**  Screen Shot 2017-08-14 at 9.14.32 PM.png  I think these two schema performed well in learning the pattern in this data. My conclusion is based on several measures. First, both models correctly classified >93.8% of the instances. Kappa statistics of both models are around 0.90, indicating that both models are doing much better than just randomly guessing. The weighted average ROC area indicates that the overall accuracy of the C4.5 model is 0.965, and the decision list model, 0.961. The weighted average TP rate is all over 0.90 for all three classes in both classification models. FP rates are low (<0.1) for all classes. The weighted average precision is also over 0.90 in both models. The confusion matrices of both models support the high TP rates and low FP rates. |

1. Compare the training set and 10-fold cross-validations scores of the two schemas.

|  |  |  |
| --- | --- | --- |
| **Training set result:**   |  |  | | --- | --- | | **Decision list:**  Screen Shot 2017-08-14 at 9.24.21 PM.png | **C4.5:**  Screen Shot 2017-08-14 at 9.25.34 PM.png |   Both models consider “OD280/OD315\_of\_diluted\_wines” to be a very important attribute. Both model also consider the attribute values of “color\_intensity”, “magnesium”, and “hue”. However, the decision list model takes into account of the values of “alcalinity\_of\_ash” and “proline”, while J48 takes into account of “alcohol”. Decision list model makes two classification errors at two nodes in the training model. No errors in the C4.5 training model.  Results of the 10-fold cross-validations are shown in B). Running t-test in the Experimenter shows that, even though the C4.5 seem to have slightly higher weighted average accuracy measures, and slightly lower error, the two models do not differ significantly in F-measure, precision, ROC area, etc. |

1. Would you trust these two models? Did they really learn what is important for proper classification of wine?

|  |
| --- |
| I think, to classify wine, one might need more information, such as the type of grapes, the season the wine is made, the demographic area of production and harvest, the sugar content, etc. It seems the 13 attributes in the raw data is not enough. Wine classification models built on just these 13 attributes do not seem very realistic.  I checked the document regarding this dataset on the UCI Machine Learning Repository, and learned that the dataset is the result of a chemical analysis of wines produced in three different cultivars in the same region of Italy, and there are originally 30 attributes. This makes me wonder if the resulting model would perform differently had all of the attributes been included. Another question is, how would the resulting model perform if we have test instances from other regions. Would the model be overfitted to these three particular Italian cultivars? |

1. Which one would you trust more, even if just very slightly?

|  |
| --- |
| I would trust the decision list model slightly more because it considers more attributes in wine classification. |

1. Perform the same analysis of sunburn.arff as in Question #2. Instead of 10-fold cross-validations use 5-fold cross-validation. Answer the same questions as in A)-E) in the question #2.

|  |  |  |
| --- | --- | --- |
| The most important attribute is “lotion”.  **Training result:**   |  |  | | --- | --- | | **Decision list**  Screen Shot 2017-08-01 at 11.07.12 PM.png | **C4.5**  **Screen Shot 2017-08-01 at 11.08.22 PM.png** |   **5-fold cross-validation result:**  **Decision list**  Screen Shot 2017-08-01 at 11.13.30 PM.png  C4.5  Screen Shot 2017-08-01 at 11.16.07 PM.png  The two models are not very good at learning the pattern in the data:   * in the training scenario, both models misclassified two of the “burned = no” instances as “burned = yes”   In the cross-validation:   * kappa statistic is negative, indicating that classification is in very poor agreement with the actual class. * Both models correctly classified only 37.5% of the instances. * ROC area indicates corrected accuracy is at 0.333, which is worse than tossing a coin. * TP rate, precision, and F-measures are all very low.   Both models use the same rule and got the same performance evaluation results.  I would not trust these models because   * sample size is low--may not be a good representation of the greater population (other data) * classification performance is very poor * various factors, such as skin color, ethnicity, diet, and protective clothing, determine whether one gets sunburn or not. These are not present as attributes in the data. Perhaps data collection needs to be reviewed.   Since the models use the same rule, and they both perform poorly, I cannot trust either. |

F) Why didn’t we use 10-fold evaluation in this example? What would happen if we us the 10-fold?

|  |
| --- |
| We would need to have at least 10 instances in order to conduct a 10-fold evaluation. We only have 8 instances in this data. We cannot get a whole number from 8/10. We would get an error if we run this in WEKA. |

1. Choose one of the following three files: soybean.arff, zoo.arff or zoo2\_x.arff and use any two schemas of your choice to build and compare the models. Describe in details the process of building (data set, parameter settings/changes, etc) and evaluation of each individual model and comparison of the different models.

|  |  |
| --- | --- |
| data | zoo.arff |
| changes to data | changed the legs attribute from numeric to nominal  @ATTRIBUTE legs INTEGER [0,9] → @ATTRIBUTE legs {0,2,4,5,6,8} |
| classifiers | Decision table, C4.5 |
| parameter setting | default  Decision Table  Screen Shot 2017-08-02 at 12.07.27 AM.png  C4.5  Screen Shot 2017-08-02 at 12.09.19 AM.png |
| Test mode | 10-fold cross validation |

**Comparisons**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Decision Table**  **Screen Shot 2017-08-02 at 12.21.33 AM.png**  **C4.5**  Screen Shot 2017-08-02 at 12.15.53 AM.png  **Compare the trained model:**  The decision table model’s rules are based on the “milk”, “fins”, “legs”, and “tail” attribute values. C4.5 model considers “feathers” as the most important attributes. While C4.5 also incorporates the attribute values of “milk”, “fins”, “tail” and “legs” to construct the tree, it also considers the “backbone”, “airborne”, and “predator” attributes for classification.  Based on the visualization of the C4.5 model training result, the model is doing pretty well with only one misclassified instance in the tree. Interestingly, C4.5 tree considers more attributes than the decision table, but the latter does not perform that poorly compared to C4.5. Does that mean some of the attributes that C4.5 use to construct the tree are not necessarily that useful? Also note that the decision table model takes slightly longer to train.  **Compare the cross-validation result:**  **Decision Table**  Screen Shot 2017-08-02 at 12.32.29 AM.png  C4.5  Screen Shot 2017-08-02 at 12.34.36 AM.png  It seems that the C4.5 model is slightly better than the decision table model.   |  |  |  | | --- | --- | --- | |  | decision table | C4.5 | | % correctly classified instances | 88.1188% | 92.0792% | | confusion matrix | poorer classification performance on reptile and invertebrate classes | poorer classification performance on insect classes | | weighted average TP rate | 0.881  (TP rate is 0 for reptile class) | 0.921  (TP rate more evenly distributed across different classes) | | weighted average FP rate | 0.058 | 0.008 (weighted FP rate  is significantly lower as shown by running t-test in Experimenter) | | kappa statistics | 0.8378 | 0.8955 | | ROC area | 0.980 | 0.975 | | weighted average MCC | 0.824 | 0.914 (significantly higher as shown by running t-test in Experimenter) | | weighted average precision | 0.837 | 0.922 |   **Conclusion**  Overall, the C4.5 model has better performance on classifying the zoo.arff dataset. It makes fewer errors during classification, scored higher on most performance measures, and moreover, has significantly better TP rate and MCC. However, I also noticed that even though decision table uses fewer attributes to construct the model, it still did fairly well. For instance, even without using the feather attribute for classification, decision table still classifies birds very well. Maybe the C4.5 model does not need these many attributes. Having these many attributes are ok for this dataset, but may decrease the model’s flexibility in making good predictions with other data.  Out of curiosity, I only used the attributes selected by the decision table to construct a new C4.5 model, and the tree ends up having even better performance. Perhaps we can use the decision table’s selected features to build a better decision tree. |