### **Fundamental of Data Mining**

***Assignment #3:***

1. Use the Decision tree method (Classify Tab, “trees” folder, J48) to analyze the iris data (iris.arff can be found in Weka’s Data folder or in Blackboard under Resources):

* Give a brief description of the Decision Tree model
* Discuss what you learned about the Iris dataset from the J48 classifier.
* How did Decision tree method perform? (We will cover the evaluation techniques in more details later in the class. You can choose any of the available options for not. However, please specify what option you chose: training data set, cross-validation or % split was used).
* How did Decision tree method provide you with the insight into your data set/rules/patterns and why?

1. Data preparation is an essential step in data mining. How the training data set is presented to a method can drastically affect the produced model’s performance. Use the J48 Decision tree-learning scheme to analyze weather.numeric.arff and weather.nominal.arff (the data sets come with the Weka installation in Weka/data folder) data set. Make predictions for the ‘temperature’ attribute for both data sets.

* Try to use J48 on weather.numeric.arff with no modifications to the dataset. Did you get an error? The method only performs on nominal class data – use the DiscretizeFilter (Unsupervised-Attribute- Discretize) filter, in the preprocess tab, before applying the learning method. Be sure to note how you discretized the dataset and take a moment to consider why you made the choice? Did you discretize all the attributes? How many bins did you discretize each attribute into?
* Analyze the output of the model that learned the discretized attribute ‘temperature’? What was the performance, can you improve it? What did the model tell you about the data? (Hint: you can modify the number of bins in the discretize filter in an attempt to improve the model performance or mimic the nominal dataset)
* Analyze the output of the model that learned the nominal attribute ‘temperature’? What was the performance, can you improve it? What did the model tell you about the data? How do the results differ from the model produced on the discretized version of the same attribute?

1. Use the J48 Decision tree learning scheme to analyze the bolts data (bolts.arff without the TIME attribute). The dataset describes the time needed by a machine to produce and count 20 bolts. (More details can be found in the file containing the dataset, you can open the file using a file editor to read the comments)

* Why should you ignore the TIME attribute?
* Analyze the model produced. What adjustments (if you were to make any) would have the greatest effect on the time to count 20 bolts (attribute: T20Bolt) (i.e. what is the most important/selective attribute/value pair in the tree)?
* According to the classifier, how would you adjust the machine (the other attributes) to get the shortest time to count 20 bolts?

1. Decision tree model is a predictive modeling approach used in Data mining and statistics. It uses a decision tree (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves).

Correctly Classified Instances using training set is 98% while Correctly Classified Instances using cross-validation is 96%.

The decision tree has classified 50 Setosa objects as Setosa. The decision tree has classified 49 versicolor objects as Versicolor and 2 as Virginica, leading in 2 missclasification. The decision tree has classified 48 Virginica objects as Virginica and 1 as Versicolor, leading in 1 missclassification.

1. 444
2. After running several scenarios, the output below is the shortest and cleanest model I was able to produce, with a tree size of 4 and just 3 leaves.  It was structured by using 3 bins and added an additional layer of error pruning within the classifier tab. It is clear from this model that the speed setting that controls the speed of rotation (SPEED1) of the plate at the bottom of the dish has the most significant impact on the time to count 20 bolts (T20BOLT). The model is telling me that if SPEED1 is between infinity and 3.33 or 3.33 and 4.66, the average time to count 20 bolts is infinity – 34.66.  This is the range we want to be in for maximum efficiency.

The model below was produced using J48, 3 bins and no additional error pruning selected in the classifier. While slightly less effective; 80% vs. 82.5% of classifications are correct, it provides more detail for fine tuning the bolt counting process. Beyond the adjustments for the speed setting (SPEED1) noted above, you could also fine tune by adjusting the sensitivity of the electronic eye (SENS) to between 3.33 and 6.66 when NUMBER2 is set between infinity and .66, the total bolts to be counted (TOTAL) is set between infinity and 16.66, and SPEED1 is set between 4.66 – infinity.  You could also fine tune the setting for NUMBER2 under the same SPEED1 setting of 4.6 – infinity. See below for model output.

SPEED1 = '(-inf-2.4]'  
|   NUMBER2 = '(-inf-0.2]': '(-inf-15.522]' (8.0/4.0)  
|   NUMBER2 = '(0.2-0.4]': '(15.522-23.724]' (0.0)  
|   NUMBER2 = '(0.4-0.6]': '(15.522-23.724]' (0.0)  
|   NUMBER2 = '(0.6-0.8]': '(15.522-23.724]' (0.0)