CSF415 DATA MINING

WEKA

Objective:

- An introduction in to the WEKA – Explorer - Environment
- Understanding the Unsupervised Attribute/instance Filters for preprocessing the input data

Tasks:

1. Creating the ARFF (Attribute Relation File Format) – from an excel sheet data given in data sheet 1 and reading into WEKA.
   Steps :
   i. Save the data set as CSV Format
   ii. Open it with a word processor
   iii. Format it appropriately according to ARFF specifications & Save as data1.arff
   iv. Open this data1.arff into WEKA EXPLORER
   v. Analyze each attributes WEKA EXPLORER interface.

2. Try each of the following **Unsupervised Attribute Filter**
   (All the filters are available under the preprocess tab. Click Choose and select the filters specified. Attributes listed below are found from ***Choose->WEKA->Filters->Unsupervised->Attributes***)

   i. Use the attribute ***Add*** and perform the following:

   - Add the attribute *Average*

   Note: Once you selected ***Add*** from the dropdown list, click on ***Add*** to get the menu to set the necessary values for the attributes, like the name of the attribute, nominal values etc.

   ii. Use the attribute ***AddExpression*** and add an attribute which is the average of attributes M1 and M2. Name this attribute as *AVG*

   iii. **Add** an attribute class with "4" classes

   iv. Understand the purpose of the attribute ***Copy.***

   v. Use the attribute ***Discretize & PKIDiscretize*** to discretize the M1 and M2 attributes into 5 equi width bins

   vi. Use the attribute ***FirstOrder*** to convert the M1 and M2 attributes into a single attribute represents the first differences between them.

   vii. Use the attribute ***MakeIndicator*** to convert the class attribute with classes c1 and c2 as binary positives and the remaining as negatives. The type of this attribute must be nominal after the application of this attribute.

   viii. Try if you can accomplish the task in the previous step using the attribute ***MergeTwoValues***.

   ix. Try the following transformation functions and identify the purposes of each

   1. ***NumericTransform***
   2. ***NominalToBinary***
   3. ***NumericToBinary***
   4. ***Remove***
   5. ***RemoveType***
   6. ***RemoveUseless***
   7. ***ReplaceMissingValues***
   8. ***SwapValues***

x. Perform ***Normalize*** and ***Standardize*** on M1 and identify the difference between these operations

3. Try the following **Unsupervised Instance Filter**
(All the filters are available under the preprocess tab. Click Choose and select the filters specified. Attributes listed below are found from
Choose->WEKA->Filters->Unsupervised->Instances)

i. Refer the support material supplied with this sheet and understand how a sparse dataset is represented with ARFF. Create a sparse dataset with the attributes similar to the data in the sheet 1 in ARFF. Apply the following instance filters to the file just created
   1. ***NonSparseToSparse***
   2. ***SparseToNonSparse***

ii. Apply to the dataset supplied with this sheet the following filters and observe the results.
   1. ***Randomize***
      Perform Randomize on the given dataset, try to correlate the resultant sequence with the given one.
   2. ***RemoveFolds***
      Set the folds as 3 and view the instances in each of the folds.
   3. ***RemovePercent***
      Remove 25 percent of the instance in the given dataset.
   4. ***RemoveRange***
      Remove from 5th to 12th instances of the given dataset.
   5. ***RemoveWithValues***
      Apply this filter to a nominal and a numeric attribute
   6. ***ReSample***
      View the result of Sampling with replacement, setting a valid percent for the sample percent.

Note:

1. Observe the results of each filters with the graphical visualizer and view the resultant instances of the given set with the edit option
2. Click 'More' button near each filter options learn more about the corresponding filter.
3. Refer the file supplied herewith on ARFF to better interpret the results

4. Use j48 (C4.5) algorithm to the Data_3 (after preprocessing). Generate 2 trees (one for unpruned and another for pruned).
   - Compare the classification accuracy. Also, try out different testing options.
   - Divide the data into 2 sets. First apply the classifier on training data set and then on test data set for both the above algorithms and compare the results.