CSF415 DATA MINING

WEKA

Data for Q1-Q3: Data_1

Data for Q4-5: Data_2

Data for Q6-9: Data_3

1. Convert Test_Data to .arff file
2. Convert attribute type (e.g. Numeric to Binary) (Age)
3. Introduce a new attribute?

    3.1. Try to insert an attribute of type nominal, having values red, green, and blue. Some tuples should contain missing values as well.

    3.2. Try to insert an attribute for the batting average of the cricketers.

    Hint: This can be done by dividing the total number of runs scored by total number of matches played. You need to write an expression for that.

4. Convert Data_2 in presence-absence table and then apply Apriori (after pre-processing).

    Hint: apply appropriate pre-processing filters like *copy*, *nominaltobinary*, *numerictobinary*, *numerictonominal* etc.

5. Apply Apriori on Data_2. Try different combination of confidence and support and observe the differences in the results.

    Hint: You need to apply appropriate pre-processing steps (e.g. replace missing values, binning/discretization etc.) before applying Apriori.

6. Use j48 (C4.5) algorithm to the Data_3 (after preprocessing). Generate 2 trees (one for unpruned and another for pruned).

    6.1. Compare the classification accuracy. Also, try out different testing options.

    6.2. 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.

7. Use naive Bayes classifier on the given data set and study the results.
8. Use nearest neighbour classification on the data set and determine the appropriate value for k.
9. Use “Jrip” (Ripper Algorithm) to apply rule based classification and compare the results of Q6, Q7 and Q8.