Name of The Laboratory: Data Mining Lab

Course Objectives:

(a) Identify methodologies, and models for data warehousing.

(b) Design various business domains for data warehouses.

(c) Identify of hidden predictive information from large databases by data mining techniques.

(d) Describe different algorithms of data mining by taking example.

(e) Use data mining tools to predict future trends in various domain.

1) List all the categorical (or nominal) attributes and the real-valued attributes separately.
2) What attributes do you think might be crucial in making the credit assessment? Come up with some simple rules in plain English using your selected attributes?
3) One type of model that you can create is a Decision Tree -train a Decision Tree using The complete dataset as the training data. Report the model obtained after training.
4) Suppose you use your above model trained on the complete dataset, and classify credit good/bad for each of the examples in the dataset. What % of examples can you classify correctly? (This is also called testing on the training set) Why do you think you cannot get 100 % training accuracy?
5) Is testing on the training set as you did above a good idea? Why or Why not?
6) One approach for solving the problem encountered in the previous question is using cross-validation? Describe what is cross -validation briefly. Train a Decision Tree again using cross - validation and report your results. Does your accuracy Increase/decrease? Why
7) Check to see if the data shows a bias against "foreign workers" (attribute 20), or "personal status" (attribute 9). One way to do this (perhaps rather simple minded) is to remove these attributes from the dataset and see if the decision tree created in those cases is significantly different from the full dataset case which you have already done. To remove an attribute you can use the preprocess tab in Weka's GUI Explorer. Did removing these attributes have any significant effect? Discuss.
8) Another question might be, do you really need to input so many attributes to get good results? Maybe only a few would do. For example, you could try just having attributes 2, 3, 5, 7, 10, 17 (and 21, the class attribute (naturally)). Try out some Combinations. (You had removed two attributes in problem 7. Remember to reload the arff data file to get all the attributes initially before you start selecting the ones you want.)

9) Sometimes, the cost of rejecting an applicant who actually has a good credit(case might be higher than accepting an applicant who has bad credit (case 2). Instead of counting the misclassifications equally in both cases, give a higher cost to the first case (say cost 5) and lower cost to the second case. You can do this by using a cost matrix in Weka. Train your Decision Tree again and report the Decision Tree and cross -validation results. Are they significantly different from results obtained in problem 6 (using equal cost)?
10) Do you think it is a good idea to prefer simple decision trees instead of having long complex decision trees? How does the complexity of a Decision Tree relate to the bias of the model?
11) You can make your Decision Trees simpler by pruning the nodes. one approach is to use Reduced Error Pruning -Explain this idea briefly. Try reduced error pruning for training your Decision Trees using cross -validation (you can do this in Weka) and report the Decision Tree you obtain? Also, report your accuracy using the pruned model. Does your accuracy increase?
12) (Extra Credit): How can you convert a Decision Trees into "if –then -else rules". Make up your own small Decision Tree consisting of 2 - 3 levels and convert it into a set of rules. There also exist different classifiers that output the model in the form of rules -one such classifier in Weka is rules. PART, train this model and report the set of rules obtained. Sometimes just one attribute can be good enough in making the decision, yes, just one! Can you predict what attribute that might be in this dataset? OneR classifier uses a single attribute to make decisions (it chooses the attribute based on minimum error). Report the rule obtained by training a one R classifier. Rank the performance of j48, PART and oneR.

KG Reddy College of Engineering & Technology

(Approved by AICTE, New Delhi, Affiliated to JNTUH, Hyderabad)
Chilkur (Village), Moinabad (Mandal), R. R Dist, TS-501504

Accredited by NAAC