Data Mining II: Advanced Methods and Techniques

Assignment #4: Using Filters and Meta Learners

Please provide sufficient support so that your models could be reproduced and provide enough discussion to illustrate your understanding of the methods used and evaluate the model

1. What is a stratified holdout procedure? Describe how it works. What is its advantage over the holdout procedure without stratification?
2. Use Decision Table or any feature selection method of your choice (under "Select Attributes" tab in Weka) to get a 'relevant' subset of data. Run Decision Tree method on just those attributes and compare it to the output (evaluation) of the Decision Tree method ran on all of the attributes. You can use either Weather data set or data set of your choice.
3. Compare at least 2 different discretization techniques on the same learning method and same data set and compare results with the baseline model. Choose a relevant dataset and create a reasonable model.
4. What are similarities and differences between Bagging and Boosting?
5. Ensemble methods can be found under the Classify tab - Meta folder. Apply Bagging and Boosting(AdaBoost) to Naïve Bayes using 10-fold cross validation. First use iris.arff data set and then use a larger data set of your choice from the UCI data repository or Weka Data folder. Compare performance with the single Naïve Bayes (without Bagging and Boosting).
6. Train decision tree, forests of randomized trees and Boosting trained on the Hypothyroid Data Set. Don't forget to perform and describe data prep and cleaning steps. Perform feature importance analysis before the training. Several different configurations of each of the models/parameters should be explored, analyzed and plotted. Demonstrate how changes in parameters influences accuracy for different algorithms. Describe your process of parameter tuning and provide in detailed discussion of the results.
7. Perform Stacking Ensemble analysis on the Boston Housing Data set. Include interesting plots and attribute importance analysis to support the choice of the final model configuration chosen. Any combination of any of the Machine Learning algorithms we have covered in the class so far or you are already familiar with is acceptable to be used in the Stacking Ensemble (choose difference algorithms as level one and level two learners). Compare the results with bagging, boosting and the single algorithms.