**WEEK 6&7**

1. We discuss this week that the quality of Rule-Based Classifiers can be evaluated using measures such as coverage and accuracy. Can you think of any particular case(s) that Rule-Based classifier may produce rules with high accuracy resulting in overfitting?
2. Can you use the SMO (Sequential Minimal Optimization) algorithm for training SVM classifier for a problem with multi-class? If yes, how was the SMO extended to handle such a case in Weka?
3. Which classification algorithm covered this week do you think resilient to overfitting? Identify one and explain what makes it resilient to overfitting.
4. What is a validation data in addition to training and test data which can be prepared from a large dataset for classification? Can you think of any specific classification algorithm which utilizes it and why?
5. This week’s video lecture on the Ensemble Method shows running the Random Forests algorithm, Bagging, and AdaBoostM1 in Weka. What is the basic concept behind Random Forests, which is different from the other two?

Q1: You are to discuss the Rule-Based classifiers in general. What is the Associate classifier? You must be thinking of CAR, which is ARM, where RHS is a class value.

1.  
Rule Based classifiers are algorithms that use If-Then rules for predicting the class of the set. One of the types of Rule based classifier if Associative classifier. It is a combination of associative rule mining algorithm and classification. It starts by making association classification rule from the training set keeping the specified frequency and confidence in its calculation and from all the rules generated it picks up the best associative rule to use for prediction. This leads to a higher accuracy as it generates a large number of rules however if the dataset is large and dense, this would result in even larger number of associative rules on the training data but would lead to overfitting and high error rate on the test dataset.

Rule Based Algorithm generates rules that tries to cover all the instances in the training set. This leads to development of rules that are complex and specific to the training data. As the rules continue to grow and become specific , they eventually stop targeting the possibilities that might occur in the test dataset. This leads to overfitting causing high error rate on test data.

2.  
Yes SMO can be used to train a SVM classifier for a multiclass problem. In case of weka Polynomial or RBF kernels are used by the SMO to train SVM. Issues with the multi-class are solved using Pairwise classification. Usually the machine learning algorithms are designed for binary decision problems i.e K=2. Pairwise classification transforms k-class problem into a k-1 binary problem by considering the instances of the class i as positive instances and the instances of class j>i as negative instances .

3.

Naive Bayes algorithm is resilient to overfitting as it uses a simple linear hypothesis function. And because the hypothesis function is so simple, it cannot accurately represent several of the complicated scenarios. Thus resulting in low variance and does not fail while generalizing the unseen data that is based on the training set it used as the simplicity of the hypothesis class prevents it from overfitting the data. However this doesn't mean Naive Bayes algorithm is completely immune to overfit , its outcome is directly proportional to the train sample if the feature selection of the train sample is poor it results in overfitting.

4.  
Validation data is similar to test dataset which is held back while training your model .Khun and Johnson explained it using a large dataset where a set of samples are kept separate for evaluating the final trained model. After the dataset is trained for the classifier the validation set is used to test the performance of the classifier and observe the error rate. This error rate gives us an unbiased estimation or prediction for what to expect while testing for the new data.  
Artificial Neural Network uses validation set . It splits the data into training set, validation set and test set. Since ANN is a non-linear model, using just the training set and test set would result in 100% accuracy and lead to overfitting. Thus validation set is used in the cross validation technique for parameter selection.

5.  
Tree bagging is the simpler version of Random Forest , it basically builds decision trees from the dataset and the process is repeated several times. The tress built are just slightly different from each other however the dataset may contain some distinctive feature or characteristic which in turn would affect the tree building algorithm as all the tress would consider these peculiar characteristics while developing and thus all the tress would be related to one another.  
This issue is mitigated by the Random Forest algorithm which is the generalization of the Bagging method. RF randomly selects a feature subset while building the tree node and thus some trees would completely ignore the peculiar characteristics present in the dataset. Thus each tree built would be different from one another as they would only be considering a subset of the factors resulting in Random Subset Selection. This will help mitigate the sampling bias by de-corelating the trees which was a limitation in the Bagging method.