Data Mining Assignment 4 Evaluation Report

(SHARMASHI)

A. The classification methods I implemented are as follows:

Decision Tree (C4.5) as basic method.

Random Forest as the ensemble version of the classification method.

Following is the set of evaluation measures in tabular form.

	Algorithm	**Accuracy ((TP+TN)*100)/(P+N)**	**Error Rate (FP+FN)/(P+N)**	**Sensitivity TP/P**	**Specificity TN/N**	**Preceision TP / (TP+FP)**	**F1 Score**	**FB Score (B=0.5)**	**FB Score (B=2)**
Breast Cancer	**Basic**	52%	0.471698113	0.172413793	0.662337662	0.17241379	0.172413793	0.172413793	0.172413793
	Ensemble	73%	0.273584906	0.034482759	0.987012987	0.5	0.064516129	0.135135135	0.042372881
Poker	**Basic**	65%	0.348082596	0.958605664	0.00913242	0.67381317	0.800727934	0.71942446	0.902749282
	Ensemble	68%	0.32300885	1	0	0.67699115	0.807387863	0.723746452	0.912887828
Led	**Basic**	69%	0.30952381	0	1	0	0	0	0
	Ensemble	69%	0.30952381	0	1	0	0	0	0

I ran the ensemble version (random forest) with following constant values:

K = 10000; //Number of training set samples and decision trees that needs to be made

F = 2; //Number of random attributes that needs to be used to determine the split at each node.

B. The ensemble method (Random Forest) significantly improved the performance of the basic classification for one data set (Breast Cancer) increasing the accuracy from 52% to 73% and It improved the performance for data set poker marginally from 65% to 68%.
   However it could not improve performance for data set led which already has an accuracy of 69% with basic method.

   Amount of improvement by Ensemble method for breast cancer is 21%.
   Amount of improvement by Ensemble method for poker is 3%.
   Amount of improvement by Ensemble method for led is 0%.

   So the breast cancer dataset gained the highest improvement and led gained minimum improvement.

   The reason why ensemble method was most effective on the breast cancer dataset is because the training set for breast cancer has all attribute values present in each tuple. So the decision trees generated are the most effective among all data sets.

C. The ensemble method on breast cancer dataset has highest accuracy of 73%. It has Specificity of 0.987012987012987 which means it was able to specify the negative value correctly for 98%.