CSI_5155_Homework_1

Student Name: Lingfeng Zhang
Student Number: 300134245
In this homework, I used Weka to analyze data.

- decision tree

  In Weka, I used J48 Algorithm as decision tree classifier.

  Confusion Matrix:

	Predicted minority class	Predicted majority class
Actual minority class	TP=0	FN=170
Actual majority class	FP=2	TN=2412

  Minority recall: 0

  Minority precision: 0

  The ROC curve is shown in Figure 1

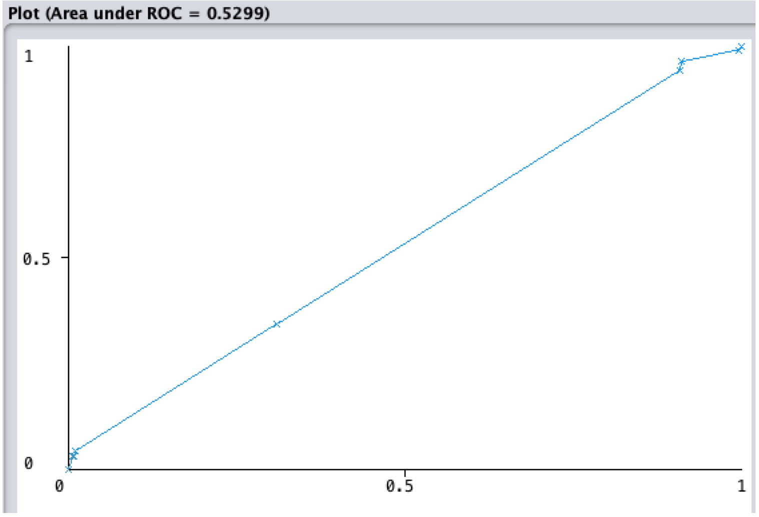

Figure 1: J48 ROC curve

  The AUC area is equal to 0.5299.

- rule-based learning

  In Weka, I used DecisionTable Algorithm as rule-based learning classifier.

  Confusion Matrix:

	Predicted minority class	Predicted majority class
Actual minority class	TP=0	FN=170
Actual majority class	FP=0	TN=2414

Minority recall: 0

Minority precision: 0/0

The ROC curve is shown in Figure 2

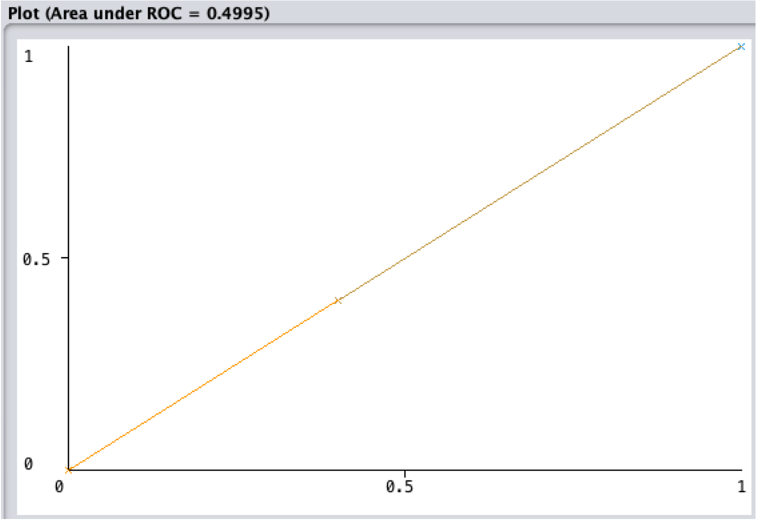

Figure 2: DecisionTable ROC curve

The AUC area is equal to 0.4995.

- Naive Bayesian classifier

  In Weka, I used NaiveBayes Algorithm as Naive Bayesian classifier.

  Confusion Matrix:

	Predicted minority class	Predicted majority class
Actual minority class	TP=68	FN=102
Actual majority class	FP=241	TN=2173

  Minority recall: 0.4

  Minority precision: 0.22

  The ROC curve is shown in Figure 3

  The AUC area is equal to 0.7553.

- k-nearest neighbor classifier

  In Weka, I used IBk Algorithm as k-nearest neighbor classifier.

  Confusion Matrix:

	Predicted minority class	Predicted majority class
Actual minority class	TP=28	FN=142
Actual majority class	FP=132	TN=2282

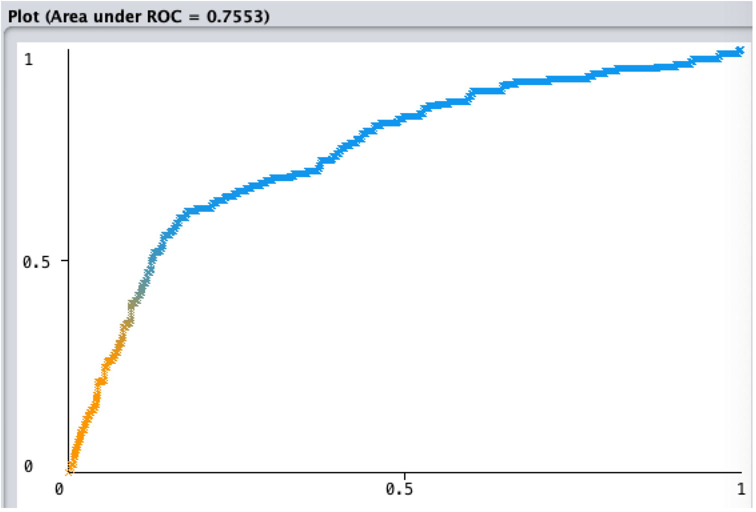

Figure 3: NaiveBayes ROC curve

Minority recall: 0.165

Minority precision: 0.175

The ROC curve is shown in Figure 4

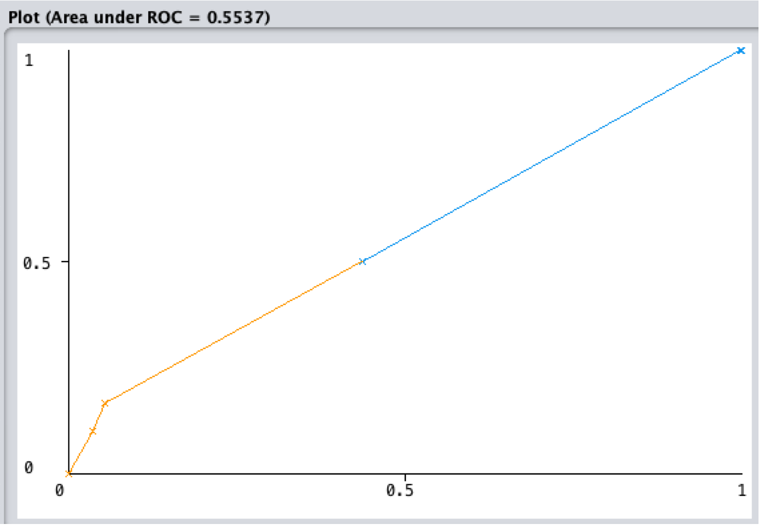

Figure 4: IBk ROC curve

The AUC area is equal to 0.5537.

Conclusion:

In the quite inbalanced data, the accuracy is meaningless because the accuracy is high if the data are always predicted as majority class. To deal with this issue, recall and precision is useful in inbalanced data. Recall represents how many minority class data are correctly predicted in all minority class data and precision represents how many minority class data are correctly predicted in all data predicted as minority class. If both minority class recall and precision are higher, this kind of classifier is better. In addition, ROC(verify classifiers' performance) curve can show which classifier is the best according to the AUC(area under the curve) size. The larger AUC, the better classifier. To summarize, the recall,precision and AUC size in Naive Bayes classifier are highest, so Naive Bayes classifier is the most suitable for detecting seismic bumps in these four classifiers.