MACHINE LEARNING (CSI 5155)

(Assignment 1)

Student(s): Wei Li

1. INTRODUCTION

Assignment 1 requires exploring and analysing an imbalanced data set regarding seismic bumps using 4 different machine learning algorithms: decision tree, a rule‐based learning, a Naïve Bayesian classifier and a k‐ nearest neighbours classifier. After applied these algorithms to the dataset and have the outcome analysed from different metrics, a preference is given to the Naïve Bayesian classifier since it can give the class of interest (the positive class) the best account.

The data set to be analysed is about forecasting high energy seismic bumps in a coal mine. 18 different input features are given, either real number data or categorical data. The objective of the classifier is to predict whether high energy seismic bumps will occur in the next shift, output ‘0’ means no high energy seismic bumps occur, while output 1 mean high energy seismic bump occurs.

The data set is imbalanced, says, 93.4% of the data are class 0, while only 6.6% of the data are class 1.  As a result, the data may make it difficult for some classifiers to find the class of interest -- the true positive instance.

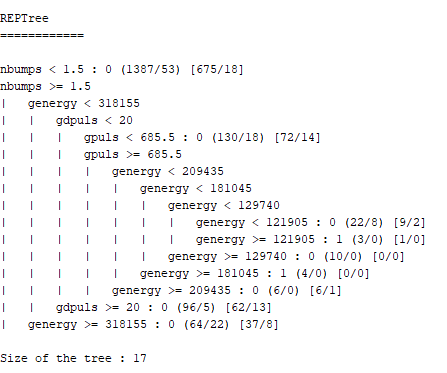
In the experiment, 4 algorithms: a decision tree, a rule‐based learning, a Naïve Bayesian classifier, and a k‐ nearest neighbor classifier were used, each gave a confusion matrix from the outcome, and the recall rate and precision rate of the positive class is calculated. Then, based on the prediction and ground truth, a ROC curve of the positive class was drawn in each of the algorithms. AUC was also calculated, from which we decided the best model to be used.

WEKA is used to complete the experiment.

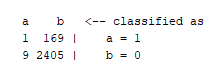
1. Data analysis and outcome
2. Question 1: Utilize Machine Learning

10-fold cross validation is used in all model to alleviate the overfitting, as default in WEKA.

* + 1. Decision tree(REPtree)
* model:



* confusion matrix:

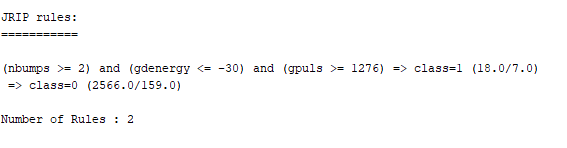


* Recall, Precision and other metrics

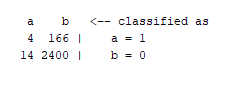


The class 1 recall equals 0.006, while its precision equals 0.100

* + 1. Rule based learning(JRip)
* Model:



* confusion matrix:

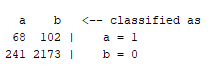


* Recall, Precision and other metrics:



The class 1 recall equals 0.024, while its precision equals 0.222

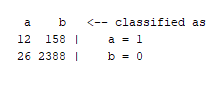
* + 1. Naïve Bayesian classifiers
* Model: no explict model representation
* confusion matrix:



* Recall, Precision and other metrics:



* + 1. k‐ nearest neighbor(lazy.IBk, K=5)
* Model: no explict model representation
* confusion matrix:



* Recall, Precision and other metrics:



* 1. Question 2: :ROC Curve

The ROC curve of these 4 models are given and compared as below, their respective AUC is also given:

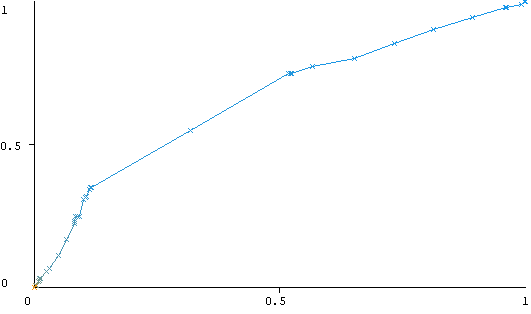
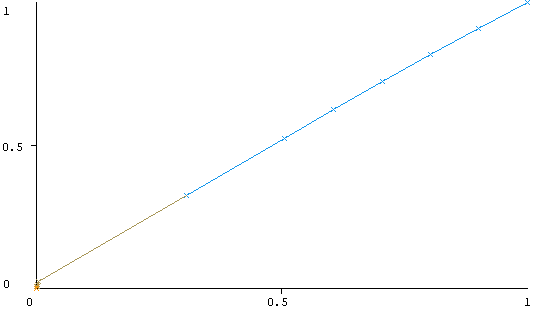
 

Figure 1-1 ROC of tree model, AUC =0.6554 Figure 1-2 ROC of rule-based model, AUC =0.5149

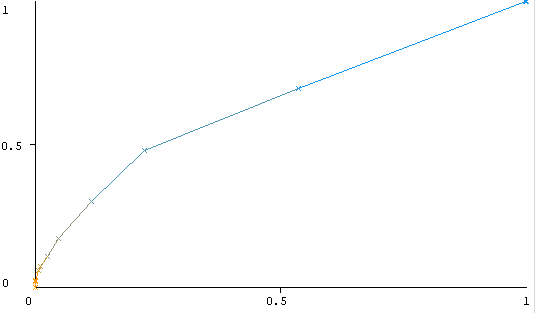
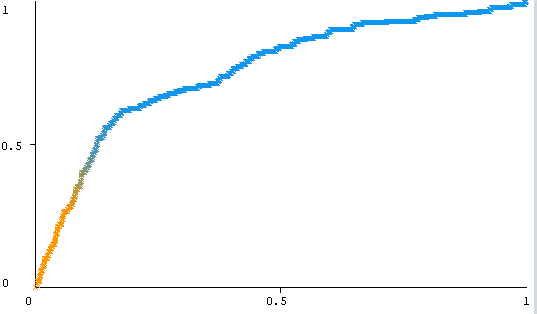


Figure 1-3 ROC of Naïve Bayesian model, AUC =0.7553 Figure 1-2 ROC of K-nearest model, AUC =0.6393

1. Conclusion

Question 3: Lessons Learned

In the experiment above, the four algorithms all obtained a relative satisfactory classify accuracy, Naïve Bayesian by 86.726%, k nearest neighbour by  92.8793%, rule-based by 93.0341% and tree model by 93.1115%.

However, since the data set is highly skewed to the negative class, their ability to forecast the positive class is not as good: while the Naïve Bayesian got its True Positive Rate and Recall of 0.4, the three other models only got true positive rates less than 0.1, which means that they are almost unable to predict the hazardous state that we really interested in. The Naïve Bayesian also do relatively good at positive class Precision Rate, which is 0.22, so compared to the others, it not only can retrieve more positive instances, and they are relatively reliable.

Examining the ROC curve, the Naïve Bayesian also shows superiority. Its Area Under Curve equals 0.7553, which is the highest among the four. On the other hand, the ROC curve of the rule-based model is almost diagonal, indicating that it has almost no ability to distinguish class 0 and 1.

Given the outcome and calculated information above, I would choose Naïve Bayesian as my model. Although may not has the highest accuracy, it accounts for the positive class the best, and thus suits our demand to predict a hazard the best.

Through the case, I learned that accuracy is not the only criteria to evaluate a model. When the data is imbalanced or when we are interested in particular class, other metrics such as Recall, Precision and ROC curve can reveal more information.

1. References

Marek Sikora^{1,2} (marek.sikora '@' polsl.pl), Lukasz Wrobel^{1} (lukasz.wrobel '@' polsl.pl)  
(1) Institute of Computer Science, Silesian University of Technology, 44-100 Gliwice, Poland  
(2) Institute of Innovative Technologies EMAG, 40-189 Katowice, Poland

Sikora M., Wrobel L.: Application of rule induction algorithms for analysis of data collected by seismic  
hazard monitoring systems in coal mines. Archives of Mining Sciences, 55(1), 2010, 91-114.