Data mining in customs risk detection with cost-sensitive classification

Xin Zhou

Abstract

To improve the efficiency and accuracy of risk management in Customs, this paper explores the data mining process for risk detection with decision tree and boosting algorithms. The data are characterised by high dimensionality, imbalance and cost sensitivity. In particular, misjudging a false declaration as truthful can be more harmful than misjudging a truthful declaration as false. Therefore, considering the different costs of misclassification, we suggest taking a cost-sensitive approach with cost matrix in data mining. The inspection results are set as the prediction target variable to train the classifiers and make predictions. A data mining model of binary classification is formulated after feature selection and rebalancing. We evaluate its performance with classic measures of classification and customs risk assessment. The results show that the performance has been significantly improved with boosting while the output is less sensitive to cost-ratio under boosting.

1. Introduction

To ensure trade facilitation and safety, most customs administrations have developed risk management systems to identify potentially high-risk cargo and transport conveyances for closer scrutiny and inspection. With the application and integration of automated systems, customs risk management is becoming more reliant on the in-depth analysis of massive data. Customs in many countries have explored and implemented big data initiatives (Okazaki, 2017). Predictably, machine learning from historical data will be increasingly helpful for effective risk assessments and accurate targeting decisions.

In recent years, big data has become a key basis of business competition, and meanwhile, risk analysis based on data mining and machine learning are widely adopted by many industries (Mikuriya, 2016). For example, credit card companies have taken advantage of classification algorithms to identify possible fraud. Historical data are processed to train the model with the risk rate of the transactions as the target variable. The input variables are the attributes of transactions, such as the location, frequency and sum, as well as the main features of the applicants, such as gender, income and job. Therefore, the main features of high-risk transactions are analysed to detect potential fraud.

Similarly, Customs also face potential cases of fraud in declarations. Many customs administrations have explored risk profiling with various data mining methods, such as clustering (Hua et al., 2006), classification (Yaqin & Yuming, 2010), association (Laporte, 2011) and statistical scoring (Coundoul et al., 2012). Data mining allows Customs to identify the key risk indicators, to summarise the parameters from large databases and increase the accuracy of targeting. Thus, it can incorporate human expertise into machine learning, which can then determine the rules, which would not be able to be detected by human intuition and experience alone.

The decision tree is one of the most widely used algorithms for classification in data mining. It can process both numerical and non-numerical data. Its outcomes are highly accurate and efficient and, importantly, easily interpreted, which is crucial for Customs. In this study, we apply the C5.0 classification algorithm with boosting to risk detection, recognising that single weak classifiers can be strengthened by 'boosting' to reduce the bias and variance of the model.

Generally, for classification modelling, two types of errors—false positive and false negative—are considered with the same impact. However, regarding risk detection in Customs, misclassifying fraudulent declarations as legitimate (FN, false negative) has more serious consequences than misclassifying legitimate declarations as fraudulent (FP, false positive). In view of this, this paper builds the data mining model with cost-sensitive learning, which allows the variation of costs for different misclassification types. The performance of the model is then discussed under varied cost ratios of the two types of misclassification. Note that the term 'classification' in this paper is used generically, not in the context of tariff classification.

2. Characteristics of customs data

2.1 High dimensionality

In most countries, there are many data elements to declare to Customs, such as consigner/consignee, loading/unloading port and cargo information. When these data are linked with inspection records and the enterprise's financial information, the resultant dataset is particularly high dimensional. While there is more information in high-dimensional datasets, it is not necessarily desirable for data mining, as high dimensionality may include irrelevant features and 'noise' that makes it difficult to understand and visualise the outcome of the model. Moreover, the amount of time and memory required for computing could be enormous (Tan et al., 2005).

Therefore, to reduce the high dimensionality, it is necessary to undertake a feature selection prior to data mining. Feature selection, also known as attribute selection, is the process of selecting a subset of relevant features (variables, predictors) to be used in model construction. Generally, feature selection chooses key fields and filters unrelated or repeated fields, relying on both selection algorithms and human expertise. The common selection algorithms include principal component analysis (PCA), linear regression and Pearson correlation coefficient.

2.2 Imbalanced class distribution

Most transactions are declared truthfully to Customs but others are declared falsely. For instance, the ratio of true and false declarations is 82.73 per cent to 17.27 per cent in the dataset of this study. This implies an imbalanced class distribution among the customs dataset. It means the likelihood that one class is represented by a large quantity of sample declarations, while the other one is represented by only a few. Standard classifiers generally perform poorly on imbalanced datasets and pay less attention to the smaller class. Classification rules that predict the small class tend to be fewer and weaker than those that predict the prevalent class (Sun et al., 2006).

For this reason, data rebalancing is indispensable if Customs is to avoid misclassification when detecting the false declarations, which are samples of a small class. The most common methods of rebalancing are oversampling and undersampling (Tan et al., 2005; Chawla et al., 2011; Sug, 2011).

Oversampling is a method to get more data by replicating existing data samples with fewer classes of data. Undersampling refers to balancing the number of different categories of data samples by reducing the number of classes of existing data samples. However, random undersampling and oversampling methods have their own shortcomings. The undersampling method can potentially remove certain important examples, while oversampling can lead to overfitting (Chawla, 2005). In practice, models after rebalancing are more likely to provide a higher identification rate on the rare category. With imbalanced class distribution, the data mining for customs risk profiling could rebalance the data at the beginning. However, the degree of rebalancing varies in different applications.

2.3 Cost-sensitive classification

In the two-class scenario, samples can be categorised into four groups after the classification process is denoted in the confusion matrix. This study adopts the two-class classification for customs risk detection, assuming that the predicted positive declarations are considered to be of high risk and inspected, while the predicted negative declarations are considered of low risk and released. The confusion matrix is presented in Table1.

Table 1: Confusion matrix of 2-class classification for customs risk detection

		Predicted class	
		Predict positive – inspected	**Predict negative – released**
Actual class	**False**	False declaration inspected (True positives, TP)	False declaration released (False negatives, FN)
	True	True declaration inspected (False positive, FP)	True declaration released (True negatives, TN)

In this two-class classification model, there are two types of errors: false negative (FN) and false positive (FP). False negative (FN) refers to the false declarations that are wrongly released. False positive (FP) refers to the true declarations that are unnecessarily inspected. Obviously, the actual losses of different types of misclassification are different. Take the bank’s loan business for instance, it will incur much higher costs when misjudging an ‘actual bad’ as an ‘actual good’ than misjudging an ‘actual good’ as an ‘actual bad’. Similarly, regarding risk detection in Customs, the consequences of misjudging a false declaration as legitimate are much more serious than misjudging a true declaration as a fraudulent one. Therefore, customs risk detection could be categorised into the cost-sensitive decision-making process, where different misclassification errors incur different costs.

In view of this, the cost-sensitive classification technique can be introduced to generate a model that has the lowest cost (Elkan, 2001). Therefore, the classifier can cover more positive examples, although at the expense of generating additional false alarms. The cost matrix for custom risk detection is provided in Table 2. The cost of committing a false negative error is denoted as Cost (A), and the false positive error is denoted as Cost (B). The cost of correct classifications—true positive and true negative—are both set to be zero.

Table 2: The cost matrix for custom risk detection

		Predicted class	
		Predict positive – inspected	**Predict negative – released**
Actual class	**False**	0	Cost (A)
	True	Cost (B)	0

According to the previous assumption that all the positive predictions are inspected, higher Cost (A) will lead to a larger proportion of positive predictions, that is, the rate of inspection will increase. So that the cost matrix could be set according to the target inspection rate and the detective rate (successfully seized rate). As a result, the ratio of Cost (A) and Cost (B) in the cost matrix in Table 2 is basically the trade-off between trade security and facilitation. For the purpose of detecting high-risk commodities such as drugs, the ratio should be significantly higher. In contrast, if it is for general risk profiling of regional declarations, the ratio could be adjusted under the constraints of limited inspection resources.

3. Decision tree and boosting

3.1 Decision tree

A decision tree is a classic learning method in machine learning. A decision tree is a tree structure in which each internal node represents a prediction about an attribute, each branch represents the output of a prediction, and each leaf node represents an output of classification with inference rules. Compared to other classification algorithms, the decision tree uses a white box model, so its rule set is simple to understand and interpret.

The decision tree belongs to supervised learning. In supervised learning, each example in the training data set is a pair consisting of an input object and a desired output value, and supervised learning analyses the training data and produces an inferred function, which can be used for mapping new examples. A decision tree is obtained by learning the input samples and determining the classification of the new data. Commonly used decision-tree algorithms include ID3, C4.5, C5.0, CART and Quest.

3.2 C5.0 Algorithm

C5.0 algorithm is a descendent of the C4.5 machine learning algorithm. It is derived from an earlier system called ID3. The C5.0 model works by splitting the sample based on the field that provides the maximum information gain. Each subsample defined by the first split is then split again, usually based on a different field, and the process is repeated until the subsamples cannot be split any further. Finally, the lowest level splits are re-examined, and those that do not contribute significantly to the value of the model are removed or pruned (Thombre, 2012).

Compared to the C4.5 algorithm, the advantages of the C5.0 algorithm are obvious: it is faster, and its memory usage is more efficient than C4.5. C5 gets smaller decision trees and generates more accurate rules (Pandya & Pandya, 2015). In particular, it supports boosting, which is a process of generating several decision trees, which are combined to improve the predictions (Pang & Gong, 2009).

3.3 Boosting

Boosting is an ensemble method that combines the performance of a set of weak classifiers to produce a single strong classifier. Boosting refers to a general and provably effective method of producing a very accurate prediction rule by combining rough and moderately inaccurate rules of thumb in a manner similar to that suggested above (Freund & Schapire, 1996; 1997). Boosting works by repeatedly running a given weak learning algorithm on various distributions over the training data, and then combining the classifiers produced by the weak learner into a single composite classifier. Generally, combining multiple classifications can reduce the bias error (Kittler, 1998).

The boosting approach starts with a method or algorithm for finding the rough rules of thumb. The boosting algorithm calls this 'weak' or 'base' learning algorithm repeatedly, each time feeding it a different subset of the training examples. Each time it is called, the base learning algorithm generates a new weak prediction rule. Boosting assigns a weight to each training example and may adaptively change the weight at the end of each boosting round. After many rounds, the boosting algorithm must combine these weak rules into a single prediction rule that, hopefully, will be much more accurate than any one of the weak rules (Schapire, 1999; 2002).

4. Evaluation measures

Evaluation measures play a crucial role in both assessing the classification performance and guiding the classifier modelling. In this study, the performance of the model is evaluated with both classic measures of classification and customs risk assessment, as follows:

(1) Inspection rate

Taking the positive prediction as high risk to be inspected, the inspection rate can be derived as the percentage of the positive predictions in all training samples

$$\text{Inspection rate}=\frac{TP+FP}{TP+TN+FP+FN}$$

(2) Accuracy

Accuracy is defined as the percentage of the number of correct predictions in total number of predictions.

$$\text{Accuracy}=\frac{TP+TN}{TP+FP+TN+FN}$$

However, for classification with the class imbalance problem as mentioned above, accuracy is no longer a proper measure since the rare class has very little impact on accuracy as compared to the prevalent class. For example, in a problem where a rare class is represented by only 10 per cent of the training data, a simple strategy can be adopted to predict the prevalent class label for every example. It can achieve 90 per cent accuracy.

(3) Precision-detection rate

Precision determines the fraction of records that turn out to be positive among the predicted positive class (Tan et al., 2005). The definition of precision is given below.

$$\text{Precision, } p=\frac{TP}{TP+FP}$$

In customs risk-detection modelling, if it is assumed that the false declaration can be seized after inspection, the detection rate of inspection is equivalent to the precision above.

(1) Recall

Recall measures the fraction of positive examples correctly predicted by the classifier (Tan et al., 2005). Classifiers with a large recall have very few positive examples misclassified as the negative class. Recall is also defined as true positive rate:

$$\text{Recall, } r = \frac{TP}{TP + FN}$$

In this study, we assume that all the predicted positive examples are targeted and inspected, thus recall means the fraction of inspected examples among all the false declarations.

(2) F_1 measure

In practice, there is a trade-off between the precision and recall values. For example, if we predict all the examples as positive, then recall will be perfect, but with a very poor precision value. Precision and recall can be summarised into a F_1 measure, which represents a harmonic mean between recall and precision

$$F_1 \text{ measure} = \frac{2rp}{r + p}$$

(3) AUC

The area under a ROC (receiver operating characteristic) curve (AUC) provides a single measure of a classifier’s performance for evaluating which model, on average, is better. The AUC value is equivalent to the probability that a randomly chosen positive example is ranked higher than a randomly chosen negative example.

In this study, we use the above evaluation measures to compare and improve the data mining model for custom risk detection.

5 Classification model for risk detection with cost sensitivity

5.1 Data understanding and preparation

We employ the C5.0 decision tree algorithm in IBM SPSS Modeler to analyse data in the study. SPSS Modeler provides an intuitive graphical interface to help visualise each step in the data mining process as part of a stream and offers multiple machine learning techniques, including classification.[1] In this study, the classification model is trained by the historical declaration data of China Customs. The dataset contains 30,000 records, and all these records are inspected declarations. According to the results of inspection, 82.73 per cent of the records are negative, referring to true declarations, and 17.27 per cent of records are positive, referring to false declarations. Remarkably, we suggest training the model with the data of inspected declarations instead of the whole dataset of all declarations including the declarations released without inspection. This reason is that, the declarations released without inspection are tagged as negative, but it may turn out to be actual positive instead.

Besides the inspection result, the dataset has 21 attributes, including the name and description of the goods, modes of transportation, country of origin, HS code, mode of trade, unit, quantity, gross weight, number of packages, unit price, total price, currency, as well as the information of the operator, such as credit class, province, region, type, industry, the registered capital, currency and registered time.

According to the inspection results, the declarations are assigned into two categories: positive and negative, tagged as type 1 and type 0. HS code and country of origin are transformed into HS chapter and continent. The name and description of the goods are excluded because they are both strings of text, and text mining is not involved in this study. After the transformation, the inspection result is set as target variable, and the other attributes are set as predictive variables.

5.2 Feature selection

Confronted with the high dimensionality of this modelling, we use Pearson Chi-square to select the main features from predictive variables. Pearson Chi-square tests for the independence of target variable and the predictive variables without indicating the strength or direction of any existing relationship. If the correlation between predictive variables and target variables is relatively strong, the impact of predictors on target variables will be significant and show high importance values.

After the independence between the target and the predictive variables was tested, the predictive variables were sorted with importance value. In this case, it turned out that the importance values of the 15 predictive variables were higher than 0.9[2], as shown in Table 3. We removed the unimportant variables with the importance value under 0.9 and remained the remaining 15 variables as input variables for data mining.

Table 3: The importance values of predictive variables

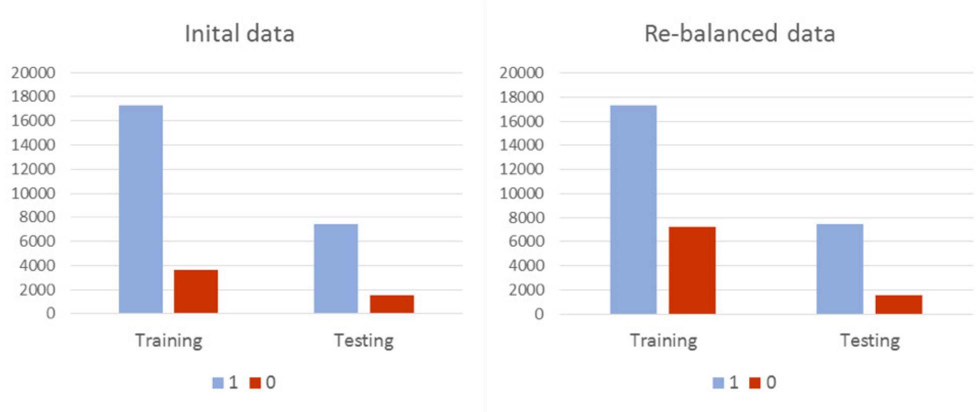
Predictive variables	Importance value	Input or not
Mode of trade	1	Y
Origin country	1	Y
HS chapter	1	Y
Mode of transportation	1	Y
Unit	1	Y
Province of the operator	1	Y
Industry of the operator	1	Y
Types of the operator	1	Y
Continent of the origin country	1	Y
Credit class of the operator	1	Y
Registered time of the operator	1	Y
Gross weight	1	Y
Quantity	0.999	Y
Unit price	0.921	Y
Total price	0.904	Y
Registered capital of the operator	0.611	N
Number of packages	0.315	N

5.3 Data partition and balance

The data were partitioned into training and testing data, with 70 per cent of the data set to train and the remaining 30 per cent to test. The proportions of the positive class (tagged as ‘1’) and the negative class (tagged as ‘0’) were17.4 per cent and 82.6 per cent in both the training and testing set, as is shown in Figure 1.

The distribution of the dataset explored (82.73% negative and 17.27% positive) indicated that a relatively balanced distribution attains a better result. However, it does not mean that the ratio of sample size of small class to the prevalent class should be 1:1. At what imbalance degree the class distribution deteriorates the classification performance varies in different applications. We used oversampling to balance the data and compared the classification performances of decision trees with different ratios. Considering the possible over fitting, we chose to double the sample size of positive class in the training dataset.[3] For the purpose of evolution, the class distribution in the testing dataset remained the same as the initial data. After balancing, the proportions of the positive class (tagged as ‘1’) and the negative class (tagged as ‘0’) were changed into 29.4 per cent and 70.4 per cent in training dataset. Meanwhile, the positive class remained at 17.4 per cent and the negative class at 82.6 per cent in the testing set for the sake of evaluation, as is shown in Figure 1.

Figure 1: Partitioned sample sizes before and after balancing



5.4 Primary decision tree model

After the data preparations above, we trained the primary decision tree model with the following parameter setting. The pruning severity was 75 per cent.[4] The ratio of Cost (A) and Cost (B) was 1:1, where the former was the cost of committing a false negative error and the latter was the cost of false positive error. Cross-validation with ten folders was applied to ensure the reliability of the model. Part of the decision tree generated is shown in Figure 2 and classification results of the primary model are shown in Tables 4–6.

Figure 2: The decision tree generated (partial)

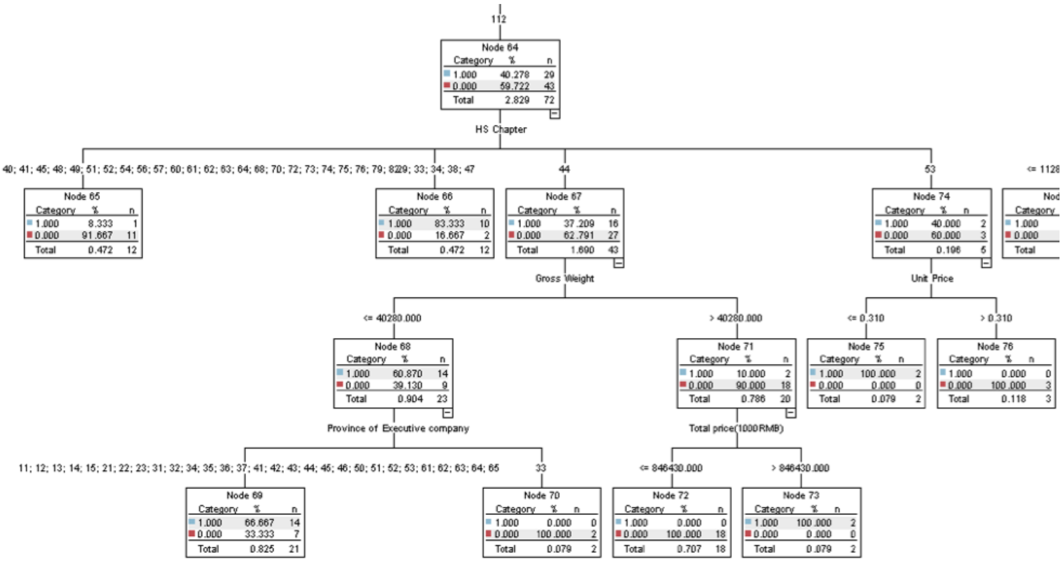


Table 4: The classification results of the primary model in training data

		Predictive class		Sum	Correct samples	Accuracy
		1	0			
Actual class	1	4,498	2,802	7,300	4,498	61.62%
	0	517	16,788	17,305	16,788	97.00%
Sum				24,605	21,286	86.51%

Table 5: The classification results of the primary model in testing data

		Predictive class		Sum	Correct samples	Accuracy
		1	0			
Actual class	1	775	800	1575	775	49.21%
	0	187	7283	7470	7283	97.50%
Sum				9045	8058	89.09%

Table 6: AUC and Gini of the primary model

	AUC	Gini
Training	0.882	0.764
Testing	0.836	0.673

Generally, the accuracy in training data is supposed to be higher than that in testing data; however, in our experiment it was not so. The accuracy in training data was 86.51 per cent while it was 89.09 per cent in testing data. However, this does not mean the performance was worse in training data as AUC, Gini and the accuracy of positive class were higher in training data. This result proves that AUC is a better measure than accuracy for evaluating and comparing classifiers.

The AUC value in testing data was 0.836, which was acceptable. As shown in the previous section, the accuracy of 89.09 per cent was less meaningful here, because in imbalanced data such as this, it could be 82.6 per cent if all the examples were predicted to be negative.

Overall, the model performed well in predicting the negative class with an accuracy of 97 per cent, but its performance was not satisfactory when dealing with the positive class because the accuracy dropped sharply to 49.21 per cent. The results suggested that the model needed to be optimised.

5.4 Boosting

In this study, we applied boosting with 10 trials. With the iteration of ten trials boosting, the model would generate ten trees and ten sets of rules. Each tree was a weak classifier and then ten trees were formed into a strong classifier after boosting. The classification results with boosting are shown in Tables 7–9.

Compared to the primary model in Tables 4–6, the performance of the classifier was significantly improved after boosting. The overall accuracy and AUC were increased from 89.09 per cent to 94.1 per cent respectively and 0.836 to 0.982 in testing data. For the prediction of the positive class, the accuracy was also raised from 49.21 per cent to 71.57 per cent, while the prediction of the accuracy in the negative class was 98.94 per cent. The classifier with boosting achieved satisfactory results.

Table 7: The classification results with boosting in training data

		Predictive class		Sum	Correct samples	Accuracy
		1	0			
Actual class	1	6536	764	7300	6536	89.53%
	0	229	17076	17305	17076	98.68%
Sum				24605	23612	95.96%

Table 8: The classification results with boosting in testing data

		Predictive class		Sum	Correct samples	Accuracy
		1	0			
Actual class	1	1120	455	1565	1120	71.57%
	0	79	7391	7470	7391	98.94%
Sum						94.1%

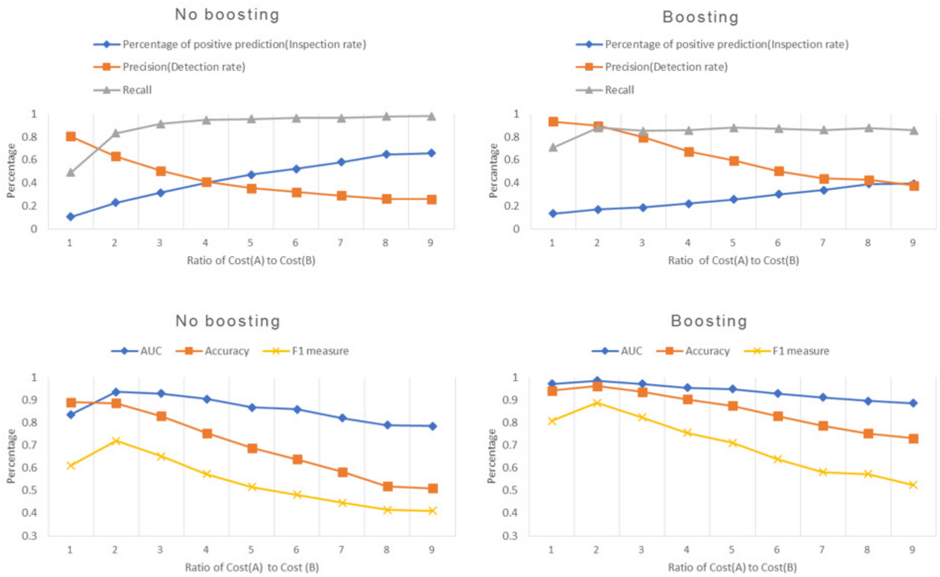
Table 9: AUC and Gini of the primary model

	AUC	Gini
Training	0.991	0.971
Testing	0.982	0.943

5.5 Cost-sensitive analysis

As discussed in the above section, risk detection in Customs is cost sensitive and, in this study, the cost of false positive, Cost (B), was set as the baseline. The ratio of Cost (A) and Cost (B) was set above one. The results of the decision tree models were compared with the variation of cost ratio, with and without boosting. We tried to explore (1) how the ratio change impacts the classifiers' performance; and (2) whether boosting is able to improve the performance on positive class with varied cost ratios.

Figure 3: The performance of the models with the variation of cost ratio



With the increase in the cost ratio, the percentage of positive prediction (inspection rate) increased with or without boosting, which proved the trade-off between trade security and facilitation, but precision (detective rate) decreased. Without boosting, the recall rate also increased with the growth of cost ratio, which indicated that more positive samples were targeted as the result of the higher inspection rate. This also showed the trade-off between recall and precision.

When boosting was applied, recall increased due to the change of cost ratio from 1 to 2. However, recall remained almost the same when cost ratio changed from 3 to 9. With similar inflection point, other metrics such as AUC, accuracy and F_1 measure improved when cost ratio increased from 1 to 2 (only except the accuracy without boosting and declined as cost ratio changed from 3 to 9).

In summary, we have come to the following conclusions:

(1) Overall, the performance of the classifiers is satisfactory regardless of the cost ratio. However, multiple rule sets will be generated with boosting, which can be too complicated to understand and interpret.

(2) The evaluation measurements with or without boosting, have different sensitivity to cost ratio. With boosting, the evaluation measurements are less sensitive to ratio variation. In contrast, without boosting, the evaluation measurements are more sensitive to ratio variation. In particular, the recall rate increases with the cost ratio without boosting. This could be applied to risk detection when a high recall rate is required, such as drug smuggling.

(3) The performance has an inflection point with the growth of cost ratio. The evaluation measurements are not changed linearly with the growth of cost ratio. After the inflection point, the performance of the classifiers will be significantly reduced as the cost ratio is raised.

The conclusions above are also reflected in Figures 4–5, which demonstrate the distributions of predictive classes under the same actual class, when the cost ratio varies in 1, 2, and 3 with or without boosting. Given a cost ratio, there are two rows that represent the sample size of actual class, tagged as ‘1’ and ‘0’. The left part in the row of actual 1 indicates true positive predictions (TP), while the rest indicates false negative predictions (FN). Similarly, the left part in the row of actual 0 indicates false positive predictions (FP), while the rest indicates true negative predictions (TN).

It also shows that when boosting is not applied, the positive predictions increase with the growth of cost ratio. More true positive predictions (TP) are covered, but meanwhile, false positive predictions (FP) also grow dramatically. In contrast, when the cost ratio builds up from1 to 2, the true positive predictions (TP) rises while false positive predictions (FP) slightly increased. As shown above, the predictions of the classifier with boosting are less sensitive to the change of cost ratio.

Figure 4: The distributions of predictive classes without boosting

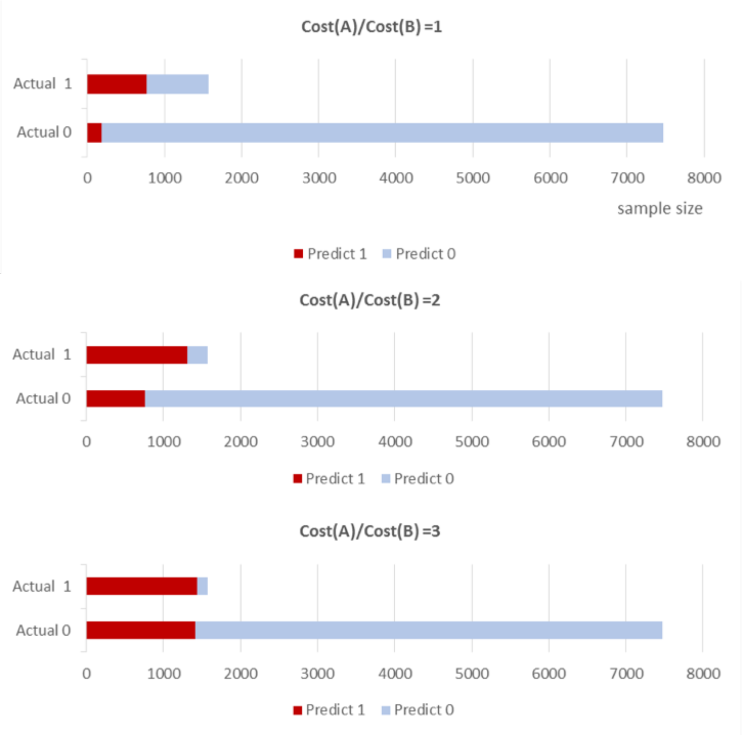
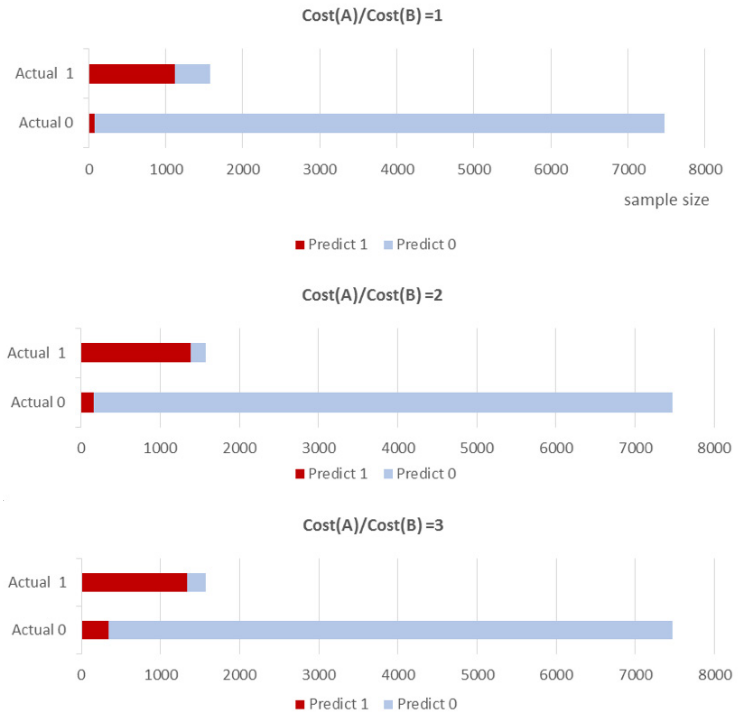


Figure 5: The distributions of predictive classes with boosting



6. Conclusion

This paper demonstrates the data mining process with a decision tree algorithm. We conclude that customs data have the characteristics of high dimensionality, imbalance, and cost sensitivity. In view of this, a data mining model of binary classification is investigated and the interactive influences of cost sensitivity and boosting on performance of the classifiers are discussed by comparing the output change with parameter adjustment. It has been proved that the model with boosting can achieve an ideal classification performance. Ultimately, this paper aims at providing a process of data mining modelling and the way of parameter adjustment, rather than the optimal values of parameters. The reason for this is that the optimal values of parameters may vary in different data applications, even for the same model.

The following research issues are open for future investigation:

(1) In this study, the positive records in the data set are combined with different types of non-compliance. If these records are segmented, according to the risk types, such as drug smuggling, price understating, or high-risk commodity, the model can generate a more specific rule set.

(2) The name and description of the goods are excluded in this study because they are both strings of text. Text mining can be explored in the future research, which can extract more information and cross-validate the data.

Overall, this study explored data mining in risk detection of Customs. With tremendous potential for applications, data analysis and machine learning will continue to receive more attention and play irreplaceable roles in customs administrations. It is worth pointing out that data mining for customs risk detection is not a one-time solution. In order to achieve a stable performance, the dataset should be updated regularly, and the parameters need maintaining constantly.

Acknowledgement

The author thanks the editors, Mr Thomas Cantens (Head of WCO Research Unit), Gael Raballand (World Bank), Cyril Chalendard (World Bank) and Christopher Grigoriou (University of Lausanne) for their helpful comments. This research is supported by MOE (Ministry of Education in China) Project of Humanities and Social Sciences (Project No.19YJC630235).

References

Chawla, N. V. (2005). Data mining for imbalanced datasets: an overview. In O. Maimon & L. Rokach (Eds), *Data mining and knowledge discovery handbook*, pp. 853–867. Boston, MA: Springer.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2011). Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*(1), 321–357.

Geourjon A. M., Laporte, B., Coundoul O., & Gadiaga M. (2012). Inspecting less to inspect better: The use of data mining for risk management by customs administrations. *Working Papers*. Fondation pour les études et recherches sur le développement international. Retrieved from http://www.ferdi.fr/sites/www.ferdi.fr/files/publication/fichiers/P46-eng_WEB_0.pdf

Elkan, C. (2001). The foundation of cost-sensitive learning. *Proceedings of the Seventeenth International joint Conference on Artificial Intelligence*, vol. 2, 973–978. San Francisco, CA: Morgan Kaufmann Publishers.

Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm. *Thirteenth International Conference on International Conference on Machine Learning*. San Francisco, CA: Morgan Kaufmann Publishers Inc.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, *55*(1), 119–139.

Hua Z., Li S., & Tao Z. (2006). A rule-based risk decision-making approach and its application in China's customs inspection decision. *Journal of the Operational Research Society*, *57*(11), 1313–1322.

Kittler, J., Hatef, M., Duin, R. P. W., & Matas, J. (1998). On combining classifiers. *International Conference on Pattern Recognition, 2*, 897–901. IEEE.

Kunio, M. (2016). Digital Customs, the opportunities of the information age. *WCO News*, 79, 9–10.

Laporte, B. (2011). Risk management systems: using data mining in developing countries' customs administrations. *World Customs Journal*, *5*(1), 17–27.

Okazaki, Y. (2017). Implications of big data for Customs – how it can support risk management capabilities. *WCO Research Paper*, No.39.

Pandya, R., & Pandya, J. (2015). C5.0 algorithm to improved decision tree with feature selection and reduced error pruning. *International Journal of Computer Applications, 117*(16), 18–21.

Pang, S. L., & Gong, J. Z. (2009). C5.0 classification algorithm and application on individual credit evaluation of banks. *Systems Engineering – Theory Practice, 29*(12), 94–104.

Schapire, R. E. (1999). A brief introduction to boosting. *Sixteenth International Joint Conference on Artificial Intelligence* (Vol.14, pp.1401–1406). San Francisco, CA: Morgan Kaufmann Publishers Inc.

Schapire, R. E. (2002). The boosting approach to machine learning—an overview, *MSRI Workshop on Nonlinear Estimation and Classification*, pp. 149–172. Berkeley, CA,.

Sug, H. (2011). An effective method to find better data mining model using inferior class oversampling. In: Lee G., Howard D., & Ślęzak D. (Eds). *Convergence and hybrid information technology. ICHIT 2011. Communications in Computer and Information Science*, 206. Berlin, Heidelberg: Springer.

Sun, Y., Wang, Y., & Wang, Y. (2006). Boosting for learning multiple classes with imbalanced class distribution. *International Conference on Data Mining* (pp. 592–602). IEEE Computer Society.

Tan, P. N., Steinbach, M., & Kumar, V. (2005). *Introduction to data mining* (1st Edn). Boston: Addison-Wesley Longman Publishing Co. Inc.

Thombre, A. (2012). Comparing logistic regression, neural networks, C5.0 and M5' classification techniques. *Machine learning and data mining in pattern recognition*. Berlin, Heidelberg: Springer.

Yaqin W., & Yuming S. (2010). Classification model based on association rules in customs risk management application. *International Conference on Intelligent System Design and Engineering Application* (Vol.1, pp. 436–439). IEEE Computer Society.

Notes

1. For more information on C5.0 node of IBSS SPSS modeler, please refer to the website: https://www.ibm.com/support/knowledgecenter/en/SS3RA7_15.0.0/com.ibm.spss.modeler.help/c50node_general.htm
2. According to the output of IBM SPSS Modeler, the importance value is considered to be important when it is above 0.95, to be marginal when above or equal to 0.9, and to be unimportant under 0.9.
3. The results showed that the performance would improve significantly if the sample size of the positive class was doubled or tripled. However, there was only a slight improvement from the doubling to the tripling. To avoid over fitting, we chose to double the sample size of positive class in the training dataset.
4. Generally, it is proper to set pruning severity from 70% to 80%. We compared the results of the model respectively when it ranged from 65%, 70%, 80% and 85%. After comparing the measures of recall, accuracy and tree depth, we chose the pruning severity of 75%.

Xin Zhou

Dr Xin Zhou is an associate professor in the Department of Customs Administration at Shanghai Customs College. Her research interests focus on risk management of customs, data analysis, and supply chain management. She holds a PhD in Management Science from Tongji University. She is studying for Executive Master of Customs and Supply Chain Compliance at Rotterdam School of Management, Erasmus University in Netherland.