Machine Learning Approaches for Customs Fraud Detection

Rimantė Kunickaitė[a], Austėja Brazinskaitė[b], Ignas Šaltis[c] and Tomas Krilavičius[d]

[a]*Department of Applied Informatics, Vytautas Magnus University, Kaunas, Lithuania*

[b]*Faculty of Economics and Business Administration, Vilnius University, Vilnius, Lithuania*

[c]*UAB "Proit", Vilnius, Lithuania*

[d]*Department of Applied Informatics, Vytautas Magnus University, Kaunas, Lithuania*

Abstract

Customs duties are based on the origin and value of the goods and their classification (the customs tariff to be applied). Falsifying any of these factors when importing or exporting products is fraud. This includes falsely declaring the origin of the goods, declaring a lower value on the goods, misclassifying the goods and smuggling goods. In this paper we apply machine learning algorithms (Artificial Neural Network, Fuzzy Min-Max Classifier and Logistic Regression) for fraud detection in customs declarations. Performance of the models are evaluated using accuracy, sensitivity and specificity. The best results were achieved using Logistic Regression. In further research it would be useful to analyze applicability of ensemble learning methods and others fraud detection models.

Keywords

customs fraud, machine learning, classification

1. Introduction

Customs fraud is any fraudulent attempt to reduce the customs duty (ex.: tax) imposed on goods when they are imported to particular country from abroad [1]. It is common in many countries but it is hard to detect it due to huge amount of data and elaborate fraud schemes. Modern customs inspections of goods are still performed by humans, but the inspected goods are randomly selected by computer systems. Such work is relatively slow, reducing the chances of detecting fraud, and most fraudsters can slip without any consequences. Therefore it is really important to improve the work of customs and make it more efficient with detecting customs fraud. Practise shows that AI is is quite popular in this field so we believe it would really help to identify the custom fraud.

2. Literature Review

An analysis of the literature has shown that research has been mostly carried out in the areas of healthcare and financial crime. Study [2] analyzed credit card fraud. The sample for this study consisted of information on 978 cases of fraud and 22 million legal transactions. Decision tree algorithms (CART, C5.0, and CHAID) and support vector machines (SVMs) with different kernel functions were used to detect fraud cases: polynomial, sigmoidal, and linear. The results of the analysis revealed that more accurate results are obtained using decision tree algorithms.

The support vector method was also compared with logistic regression models [3]. The results showed that the regression model has more accurate results (with accuracy of 91 %) than the SVM classifier (with accuracy of 82 %).

In order to develop a methodology to detect fraudulent financial statements of companies, a data set consisting of 76 companies engaged in manufacturing was analyzed [4], from which 38 of these firms were accused of falsifying financial statements. Few methods were compared experimentally: neural networks, decision trees, and Bayesian networks. In this case, the most accurate results were obtained using Bayesian networks (with accuracy of 90.3 %), followed by the 80 % accuracy of neural networks and 73.6 % of decision trees.

All of the above methods belong to the class of supervised learning methodology and are used to create models based on historical data of fraud cases. In this paper we are also going to use these techniques.

3. Methodology

3.1. Feature Selection

Feature selection is the process of reducing the number of input variables when developing a classification model. It is recommended to reduce the number of input variables to reduce the computational cost and to improve the performance of the model. Statistical based feature selection methods involve evaluating the relationship or difference between each input variable and the target variable using statistical tests. The choice of statistical measures depends on the data type of both the input and output variables.

Feature selection between numerical input variables and categorical output variable:

1. The t-test is used to determine if the means of two samples are significantly different from each other. Suppose that $(x_1, x_2, \ldots, x_n)$ and $(y_1, y_2, \ldots, y_m)$ data samples are obtained by measuring two independent normally distributed random variables $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$, when the means μ_X, μ_Y and the variances σ_X^2, σ_Y^2 are unknown. Testing statistical hypothesis [5]:

$$\begin{cases} H_0 : \mu_X = \mu_Y, \text{the two populations are similar} \\ H_1 : \mu_X \neq \mu_Y, \text{the two populations are different.} \end{cases} \quad (1)$$

Normality condition of the two independent samples can be checked using *Shapiro-Wilk test* or the central limit theorem says no matter what distribution things have, the sampling distribution tends to be normal if the sample is large enough ($n > 30$) [5].

The independent samples t-test comes in two different forms: the standard *Student's t-test*, which assumes that the variance of the two groups are equal ($\sigma_X^2 = \sigma_Y^2$) and the *Welch's t-test* which do not assume that the variance is the same in the two groups. The F-test is applied to test the hypothesis of the equality of two independent samples variances.

2. After performing the t-test, Cohen's d effect size is calculated to measure the strength of the difference between two variables.

a) Cohen's d for the *Student t-test* [5, 6]:

$$d = \frac{\overline{x} - \overline{y}}{s_p},\tag{2}$$

where $\overline{x}$ and $\overline{y}$ represent the mean values of the two samples, s_p is an estimator of the pooled standard deviation of the two samples. It can be calculated as follow:

$$s_p = \sqrt{\frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}},\tag{3}$$

where s_x^2 and s_y^2 are the variance of the two samples, n and m are sample sizes.

b) Cohen's d for the *Welch's t-test* [6, 5]:

$$d = \frac{\overline{x} - \overline{y}}{\sqrt{(s_x^2 + s_y^2)/2}}.\tag{4}$$

T-test conventional effect sizes, proposed by Cohen, are: 0.2 (small effect), 0.5 (moderate effect) and 0.8 (large effect) [6]. Variables that have a moderate or large effect size are selected for classification model developing.

Feature selection between categorical input variables and categorical output variable:

1. The χ^2 test of independence is used to test whether there is a relationship between two categorical variables [5]. Testing statistical hypothesis:

$$\begin{cases} H_0 : \text{the variables are independent, there is no relationship between} \\ \quad \text{the two categorical variables.} \\ H_1 : \text{the variables are dependent, there is a relationship} \\ \quad \text{between the two categorical variables.} \end{cases}\tag{5}$$

The χ^2 test of independence works by comparing the observed frequencies to the expected frequencies if more than 20% of expected frequencies < 5 and at least one expected frequency < 1 the *Fisher's exact test* is used.

2. After performing the χ^2 test, Cramer's V effect size is calculated to measure the strength of the relationship between two variables [7]:

$$V = \sqrt{\frac{\chi^2}{n \cdot df}},\tag{6}$$

where n is the sample size, $df = \min(r-1, c-1)$ and r is the the number of rows and c is the number of columns in the contingency table. The interpretation of Cramér's V effect size depends on the degrees of freedom df, shown in Table 1.

Table 1
Effect Sizes for Cramer's V [6, 7]

df	Small	Moderate	Large
1	0,10	0,30	0,50
2	0,07	0,21	0,35
3	0,06	0,17	0,29
4	0,05	0,15	0,25
5	0,04	0,13	0,22

3.2. Data set balancing

In practice, unbalanced data sets often occur, which can cause major problems in classifying with machine learning algorithms. The class imbalance affect process of machine learning, algorithm completely ignores the minority class [8]. This happens because machine learning algorithms are usually constructed to improve accuracy by reducing the error. To solve this problem, class balancing is performed in two ways: Synthetic Minority Oversampling Technique (SMOTE) and a combination of Random Over-Sampling and Random Under-Sampling techniques.

Random Over-Sampling method aims to increase the number of instances in the minority class by randomly duplicate examples in the minority class. Random Under-Sampling method aims to balance class distribution by randomly delete or merge examples in the majority class. This process is done until the majority and minority class instances are balanced out [9].

SMOTE is an over-sampling method in which the minority class is over-sampled by creating new synthetic examples rather than by over-sampling with duplication. The minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any / all of the k minority class nearest neighbors [10].

3.3. Artificial Neural Network

Artificial Neural Networks are a component of artificial intelligence that is based on the functioning of a human brain and consists of units of computation called nodes or neurons. The neuron multiplies the input by the corresponding weights, then applies an activation function to the weighted sum and creates an output. To express this mathematically we can describe it by equation:

$$y(x) = f(\boldsymbol{w} \odot \boldsymbol{x} + b), \tag{7}$$

here $\boldsymbol{x}$ – input vector, $\boldsymbol{w}$ – weight vector, b – threshold value, $\odot$ – the product of the corresponding elements, f – activation function, y – output. Neuron functioning is depicted in Figure 1.
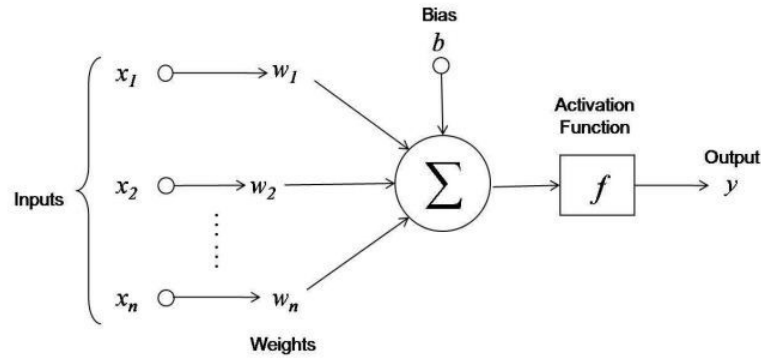
Figure 1: Functioning of Neuron [11].

3.4. Fuzzy Min-Max Classifier

The Fuzzy Min-Max Classifier that we are using in this paper is based on fuzzy logic. Suppose we have a n dimensional model of A_h and K membership functions. K membership functions are sets of fuzzy logic that define input participation in a particular class. Therefore, the input defined by the set of fuzzy logic will belong to the class to which the model gives the highest degree of membership [11, 12]. Membership degree can be defined mathematically by equation:

$$A = \{x, mA(x)\}, \tag{8}$$

here X – is a space of objects whose common element X is marked as x. Fuzzy set (class) A in space X is defined by membership function $m \sim (x)$ which connects each point in the space X with a real number in the interval $[0, 1]$, when $m \sim (x)$ the value x represents degree of membership in set A [11, 12].

The architecture of this network is similar to the architecture of a simple neural network depicted in Figure 1. The first layer is the $n - dimensional$ input layer. In a hidden layer, each node is defined in the *min* and *max* dimensional space and belongs to a particular class. The last layer consists of class nodes, which quantity is equal to the number of possible classes. It is this layer that is different from a simple neural network. Here the outputs in the layer are used as membership functions, and the class node that provides the highest estimate resolves the value of the input class [11, 12].

3.5. Logistic Regression

Logistic Regression is used when the dependent variable is categorical. From our example: to predict whether a fraud is not detected (0) or (1). Logistic regression is valid under these general assumptions: independent variables do not have to be normal, normally distributed errors are not required, and the homoscedasticity of the dependent variable is not examined. However, this model has the disadvantage of being sensitive to the problem of multicollinearity. To avoid

		Classification	
		Positive	**Negative**
Condition	+	True Positive	False Negative
	-	False Positive	True Negative

Figure 2: Confusion Matrix [13].

this, independent variables need to be chosen so that none of them are linear variables of the rest.

After selecting the required variables the probability of the event $y = 1$ is calculated:

$$\hat{P}(y = 1|\vec{x}) = \frac{e^{\hat{a}+\hat{b}_1 x_1+\hat{b}_2 x_2+...+\hat{b}_k x_k}}{1 + e^{\hat{a}+\hat{b}_1 x_1+\hat{b}_2 x_2+...+\hat{b}_k x_k}} \tag{9}$$

where $\vec{x}$ is a vector of independent variables. After this calculation a threshold can be selected and therefore class selected.

3.6. Evaluation Metrics

The following evaluation metrics were used to evaluate the performance of potential fraud detection models: accuracy, sensitivity, specificity [13]. These measures can be calculated based on the confusion matrix, which is a table with two rows and two columns that reports the number of false positives, false negatives, true positives, and true negatives (Figure 2).

The confusion matrix itself is relatively simple to understand, but the related terminology can be confusing [13]:

1. **TP** and **TN** indicate that cases of fraud and normal activity are correctly classified (predicted).
2. **FP** means that normal activity was misclassified as a fraud.
3. **FN** indicates that the fraud was misclassified as a normal activity.

Accuracy is the proportion of true results (both true positives and true negatives) and total number of cases [13]:

$$A = \frac{TP + TN}{TP + TN + FP + FN} . \tag{10}$$

Sensitivity of a classifier is the ratio between correctly identified positives and actual positives [13]:

$$SE = \frac{TP}{TP + FN} . \tag{11}$$

Specificity of a classifier is the ratio between correctly classified negatives and actual negatives [13]:

$$SP = \frac{TN}{TN + FP}.\tag{12}$$

Sensitivity shows how well positive class is predicted (in this case – declarations without irregularities). Specificity shows how well negative class is predicted (in this case – potential fraud detected from customs declarations). Accuracy describes the overall prediction accuracy of both classes.

4. Data set

The sample of the survey consists of anonymized declarations submitted to customs, which can be 15 types, the types of declarations are given in Table 2. The period for submission of declarations is 9 days. The received declaration in the system is evaluated according to the defined rules, the output of the rules system is the label *low risk of fraud*, *medium risk of fraud* or *high risk of fraud*. After inspection by a customs officer, the declaration is marked F if fraud was not detected and T if fraud was detected. Except declarations with reliable customs estimates were selected for the research, i.e. from low risk of fraud class with rating T, from medium risk of fraud class with rating T and from high risk of fraud class with rating F and T. Total number of declarations in the data set is 992. In addition to the main 306 attributes that describe the data, there are attributes-flags that describe what procedures were performed. Depending on the type of declaration, only certain attribute fields are filled in, consequently there are many omitted values in the data set. More detailed information about the data set and its attributes cannot be provided due to security requirements.

5. Experiments

5.1. Data preprocessing and feature selection

1. Due to the large number of attributes (877) and the small sample size (992 declarations), attributes with more than 50 % missing values, attributes with equal values, text attributes and categorical variables that have more than 30 categories were removed. After this preprocessing, 104 attributes remained in the data set, of which 78 are factors, 24 are numeric, identification number and label (F or T).

2. Between each numeric input variable and the target variable Cohen's d effect size for *t-test* was calculated and 11 of the 24 variables that had a moderate or large effect size were selected for classification model developing. Also between each categorical input variable and the target variable Cramer's V effect size for χ^2 test was calculated and 42 of the 78 variables that had a moderate or large effect size were selected for classification model developing.

3. In the train-test split procedure 75 % of the data set records were randomly assigned to the train data set and 25 % to the test data set.

Table 2
Types of Declarations

Type	Name
D01	Summary Declaration
D02	Entry Summary Declaration in Office of Entry
D03	Exit Summary Declaration in Office of Exit
D04	Import Declaration with Storage
D05	Import Declaration
D06	Supplementary Import Declaration
D07	Export Declaration with Storage
D08	Export Declaration
D09	Export Declaration in Office of Exit
D10	CAP Notice of Export
D11	Export Manifest
D12	Supplementary Export Declaration
D13	Transit Declaration with Storage
D14	Transit Declaration
D15	Transit Declaration with Destination

4. Data set balancing was performed in two ways: Synthetic Minority Oversampling Technique (SMOTE) and a combination of Random Over-Sampling and Random Under-Sampling techniques.

5. Categorical variables were expressed in numerical binary expression using one hot encoding technique and min-max normalization was used for numeric variables.

5.2. Results

The experiment results are presented at Table 3. It was found that the best results in terms of accuracy are obtained using Logistic Regression and Fuzzy Min-Max Classifier methods, when the training set is balanced according to the combination of over-sampling and under-sampling (accuracy is 0.92 and 0.89, respectively) and Logistic Regression with 0.92 accuracy when the data is balanced by SMOTE method. After evaluating the results according to the sensitivity index, the most appropriate methods are logistic regression and the Fuzzy Min-Max Classifier (indicators equal to 0.96 and 0.94, respectively). In terms of specificity, the highest values of the indicator were achieved using Logistic Regression and the Fuzzy Min-Max Classifier with data balanced by the SMOTE method (results equal to 0.56 and 0.53, respectively).

From the compiled classifications, the Logistic Regression model trained with a combination of balanced training data reached the highest values of all indicators. With this model, declarations without irregularities (F) are identified with 96 percent accuracy, and possible cases of fraud (T) with 56 percent accuracy. The application of fuzzy logic in the construction of the neural network improved the values of the assessment indicators compared to the neural network when no fuzzy logic was applied in the combination method with balanced data and at least improved the value of the specificity indicator only from 0.34 to 0.35 when training with SMOTE balanced methods. However, a neural network based on fuzzy logic is not the most

Table 3
Evaluation Metrics

Method	Accuracy	Sensitivity (class F)	Specificity (class T)
Logistic Regression (both)	**0.92**	**0.96**	**0.56**
Logistic Regression (SMOTE)	0.92	0.95	0.53
Artificial Neural Network (both)	0.88	0.93	0.21
Artificial Neural Network (SMOTE)	0.85	0.91	0.34
Fuzzy Min-Max Classifier (both)	0.89	0.94	0.49
Fuzzy Min-Max Classifier (SMOTE)	0.79	0.88	0.35

effective method for identifying possible cases of fraud, as models constructed by the Logistic Regression method provide more accurate classification results.

6. Conclusion

Experiments with data set of customs declarations show that:

1. The best results in terms of accuracy are obtained using Logistic Regression and Fuzzy Min-Max Classifier models, when the training set is balanced according to the the combination of oversampling and undersampling (accuracy is 0.92 and 0.89, respectively) and Logistic Regression model with 0.92 accuracy when the data is balanced by SMOTE method. According to the sensitivity, the most appropriate methods are Logistic Regression and the Fuzzy Min-Max Classifier (indicators equal to 0.96 and 0.94, respectively). In terms of specificity, the highest values were achieved using Logistic Regression and the Fuzzy Min-Max Classifier with data balanced by the SMOTE method (results equal to 0.56 and 0.53, respectively).

2. From the compiled classifications, the Logistic Regression model trained with a combination of balanced training data reached the highest values of all evaluation metrics. The application of fuzzy logic in the construction of the neural network improved the values of the assessment indicators compared to the neural network when no fuzzy logic was applied in the combination method with balanced data.

3. Identification methods of fraud, constructed as classifiers for labeled customs declarations data could be more effective if the data set were larger and contained more examples of fraud.

In further research it would be useful to analyze variety of ensemble learning methods and others fraud detection models.

References

[1] W. L. Collaborative, Customs fraud, 2020. URL: https://www.whistleblowerllc.com/what-we-do/financial-fraud/customs-fraud/.

[2] Y. Sahin, E. Duman, Detecting credit card fraud by decision trees and support vector machines, IMECS 2011 - International MultiConference of Engineers and Computer Scientists 2011 1 (2011) 442–447.

[3] S. Y. Huang, Fraud detection model by using support vector machine techniques, 2013.

[4] E. Kirkos, C. Spathis, Y. Manolopoulos, Data mining techniques for the detection of fraudulent financial statements, Expert systems with applications 32 (2007) 995–1003.

[5] V. Čekanavičius, G. Murauskas, Statistika ir jos taikymai, Vilnius: teV 1 (2000).

[6] J. Cohen, Statistical power analysis for the behavioral sciences, 2nd edn. á/l, 1988.

[7] H.-Y. Kim, Statistical notes for clinical researchers: Chi-squared test and fisher's exact test, Restorative dentistry & endodontics 42 (2017) 152–155.

[8] N. Japkowicz, S. Stephen, The class imbalance problem: A systematic study, Intelligent Data Analysis (2002) 429–449.

[9] G. E. A. P. A. Batista, R. Prati, M. C. Monard, A study of the behavior of several methods for balancing machine learning training data, SIGKDD Explorations 6 (2004) 20–29.

[10] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, Smote: Synthetic minority over-sampling technique, J. Artif. Int. Res. 16 (2002) 321–357.

[11] P. Gajjewar, Understanding fuzzy neural network using code and animation, 2018. URL: https://medium.com/@apbetahouse45/understanding-fuzzy-neural-network-with-code-and-graphs-263d1091d773.

[12] P. K. Simpson, Fuzzy min—max neural networks—part 1: Classification, IEEE Trans. on Neural Networks 3 (1992) 776–786.

[13] A. Tharwat, Classification assessment methods, Applied Computing and Informatics (2018). URL: http://www.sciencedirect.com/science/article/pii/S2210832718301546. doi:https://doi.org/10.1016/j.aci.2018.08.003.