ECONTECHMOD. AN INTERNATIONAL QUARTERLY JOURNAL – 2019, Vol. 08, No. 2, 3 – 8

Ensemble-based Method of Fraud Detection at Self-checkouts in Retail

P. Vitynskyi, R. Tkachenko, I. Izonin

Department of Publishing Information Technologies, Lviv Polytechnic National University, S. Bandery 28a, 79008 Lviv, Ukraine; e-mail: pavlo.vitynsky@gmail.com, roman.tkachenko@gmail.com, ivanizonin@gmail.com

Received: May 17, 2019, accepted June 19, 2019

Abstract. The authors consider the problem of fraud detection at self-checkouts in retail in condition of unbalanced data set. A new ensemble-based method is proposed for its effective solution. The developed method involves two main steps: application of the preprocessing procedures and the Random Forest algorithm. The step-by-step implementation of the preprocessing stage involves the sequential execution of such procedures over the input data: scaling by maximal element in a column with row-wise scaling by Euclidean norm, weighting by correlation and applying polynomial extension. For polynomial extension Ito decomposition of the second degree is used. The simulation of the method was carried out on real data. Evaluating performance was based on the use of cost matrix. The experimental comparison of the effectiveness of the developed ensemble-based method with a number of existing (simples and ensembles) demonstrates the best performance of the developed method. Experimental studies of changing the parameters of the Random Forest both for the basic algorithm and for the developed method demonstrate a significant improvement of the investigated efficiency measures of the latter. It is the result of all steps of the preprocessing stage of the developed method use.

Keywords: classification, Ensemble-based method, Random Forest, fraud detection, retail, Ito decomposition, imbalanced dataset

INTRODUCTION

Nowadays artificial intelligence tools are widely used in various business areas: retail trade in goods and services, including e-commerce, financial companies, start-ups, etc. in order to analyze and increase the volume of sales of goods and services. Many machine learning technics are used for different intellectual systems that were created to solving real-life problems [1, 13]. One of the concrete examples is introducing self-scanning systems (for example Fig. 1 [2]) in stores that allows clients to scan their items using mobile scanners while shopping [3].

This is a great opportunity to reduce queues at the cash registers, cut down costs by reducing the number of cashiers [4] and in general it is an opportunity to give consumers more freedom [5]. However, at the same time there are a number of problems. In particular, there is the open risk for retailers for this kind of payment. Some customers will not use this possibility in a proper way and will not scan all of the items in their basket. In this case damage is provided to seller. Empirical research has shown that around 5% of all transactions are fraud (or may contain technical problems with equipment) [6].

Fig. 1. Example of the self-scanning system provided by Diebold Nixdorf [2]

Therefore, the task of the identification of such cases using artificial intelligence is very important today.

THE LITERATURE OVERVIEW

There are a few different approaches that are used for fraud detection [7]. Expert systems with a multitude of statistical and logical rules aimed at detecting suspicious transactions are traditionally widely used to detect fraudulent transactions. However, this approach has several disadvantages. Machine learning based methods help to reduce risks of incorrect detection of fraud. The use of machine learning methods in conjunction with statistical rules helps reduce the risks

associated with the limitations of expert systems — in particular, reduce the number of cases in which legitimate transactions are erroneously identified as fraudulent, and increase the number of successfully identified fraud transactions. Machine learning algorithms make it possible to detect dependencies that are not obvious to humans by quickly analyzing huge amounts of data. Both types of learning algorithms are used for this problem: learning with a teacher (supervised learning) and without a teacher (unsupervised learning). In the first case, we are talking mainly about classification algorithms, when there is a training dataset with previously known answers, and in second, there are no such answers. Classification algorithms require historical data set with the fraud label (true or false) for each instance for training the model [14]. A trained model is used for predicting the probability of fraudulent transaction. The following algorithms are commonly used for classification (neural networks, random forest, support vector machines, logistic regression, gradient boosting). The main complexity of supervised learning is that an imbalance of classes is inevitable: the number of legitimate transactions is hundreds of thousands of times higher than the number of fraudulent transactions.

In [8, 9, 10] the approach to classification is proposed, used extension of the input variable space using Ito decomposition and application different machine learning algorithms:

- Support Vector Regression [8];
- Logistic Regression [9];
- Random Forest [10];
- Probabilistic Neural Network [10].

Applications of the Probabilistic Neural Network for solving real classification tasks not in all cases provide the good accuracy [10]. As investigated in [8], Support Vector Regression with different kernel provides a precession result. The training algorithm of one from two investigated implementation of SVR [8] provide a very fast solution. However, the problem of the searching effective kernel for different task is one of the disadvantages of this method. Classifier based on Logistic regression do not provide effective results in condition of the Big Data processing [9]. As investigated in [10], Random Forest algorithm has many advantages of solving classification task. But this algorithm does not always provide a good result when working with imbalance dataset

The aim of this work is to develop robust method of fraud detections and to prove the feasibility of the created methods in real-life problem with imbalanced data sets.

PROPOSED ENSEMBLE-BASED METHOD

To improve an accuracy of classification authors have proposed using composition of feature transformations in conjunction with Random Forest classifier. Our approach is composed of two phases: data preprocessing and classification.

Data preprocessing phase include such step-by-step procedures:

- Normalization by MAX value;
- Row-wise normalization by Euclidean norm;
- Weighting by correlation;
- Polynomial extension.

The flow-chart of the proposed method is shown in Fig. 2.

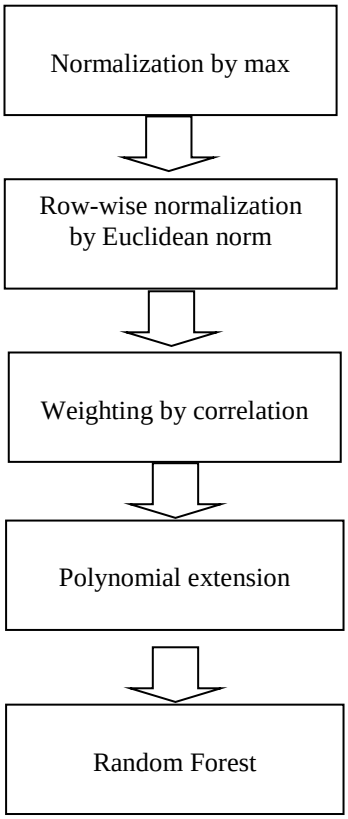

Fig. 2. Flowchart of the proposed method

Let's us consider all method's procedures in details.

Step 1. Linear scaling. Normalization is the most common part of data preprocessing in machine learning. Since the random forest is based on recursive partitioning it is invariant to monotonic transformations, therefore, attributes column-wise scaling doesn't affect it. It is performed using (1):

$$t_i = \frac{x_i}{\max(x)}. \quad (1)$$

Step 2. Row-wise scaling. Row-wise scaling ensures the ratio between the individual attributes instead of the values of the attributes. The experimental results confirm the increase of information gain in conjunction with more common methods of normalization based on individual components of the vector. Each element of the

row is divided by the Euclidean norm (2) of row.

$$|x| = \sqrt{\sum_{i=1}^{n} |t_i|^2} \,. \qquad (2)$$

Step 3. Weighting by correlation. To identify the importance of the data set attributes corresponding to the label the Pearson correlation was used. The higher the weight of an attribute, the more relevant it is considered. The dataset attributes were weighted by their correlation coefficients.

Step 4. Polynomial extension. The Ito decomposition is widely used for the development of various nonlinear approximation models [10]. The general view of Ito decomposition can be written as follows [11]:

$$Y(x_1,...,x_n) = \theta_i + \sum_{i=1}^{n} \theta_i x_i + \sum_{i=1}^{n}\sum_{j=i}^{n} \theta_{i,j} x_i x_j + ... + \\ + \sum_{i=1}^{n}\sum_{j=i}^{n}\sum_{l=j}^{n} \theta_{i,j,l} x_i x_j x_l + ... + \\ + \sum_{i=1}^{n}\sum_{j=i}^{n}\sum_{l=j}^{n} ... \sum_{z=k-1}^{n} \theta_{i,j,l,...,z} x_i x_j x_l ... x_z \qquad ,(3)$$

where:

k is the degree of the Ito decomposition;

y is output attribute;

$q_j, j = \overline{1,n}$ are regression parameters;

n is feature's number in each vector $x^{(i)}$.

Members of the Ito decomposition x_i, $x_i x_j$, $x_i x_j x_l, ..., x_i x_j x_l ....x_z$ are formed sequentially for $k = 1,2,3,\ldots$ for the given $i = \overline{1,n}$, $j = \overline{i,n}$, $l = \overline{j,n}$, ... , $z = \overline{k-1,n}$ [11].

Based on the Weierstrass theorem, it is possible to use the Kolmogorov-Gabor polynomial to increase the space of input data. This can increase the likelihood of correct imbalanced data classification [10].

Step 5. Random Forest. Random forest is a powerful classification method based on an ensemble of decision trees where each decision tree makes prediction independently and the class for which most classifiers have voted becomes a final prediction. The main parameters of the method are:

1) The number of decision trees in the forest;
2) The number of randomly selected training set features for building trees;
3) The maximum depth of the tree.

The method is non sensitive to noise in dataset and robust to overfitting.

The disadvantages of the method are complexity and consumption of large amount of computational resources.

But the composition of the method based on the use preprocessing procedures (scaling by maximal element per column with row-wise scaling by Euclidean norm, weighting by correlation and applying polynomial extension) with Random Forest algorithm should show significant accuracy improvement.

MODELING AND RESULTS

Case-study descriptions. Authors used dataset from [12] for modelling. It is a real case-study from the store. The data set contains information about the scanning process for a customer, including the number of scanned items per second and the total time spent in the store. The data set contains a column with the classification into fraud and not fraud. There are 1879 instances with 9 attributes (Table 1) in data set and only 104 of them are marked as fraud (imbalanced data set).

TABLE 1. DATA SET DESCRIPTIONS

Attribute name	Min value	Max value	Mean value
trustLevel	1	6	3.4018
totalScanTimeInSeconds	2	1831	932.15
grandTotal	0.01	99.96	50.86
lineItemVoids	0	11	5.46
scansWithoutRegistration	0	10	4.9
quantityModifications	0	5	2.52
scannedLineItemsPerSecond	0.0005	6.66	0.05
valuePerSecond	0.000007	37.87	0.2017
lineItemVoidsPerPosition	0	11	0.745

The accuracy of the model is calculated using following cost matrix (Table 3). It was built based on empirical observations [12].

TABLE II. COST MATRIX FOR EVALUATION OF THE METHODS EFFECTIVENESS

		Prediction	
Actual value		0 (no fraud)	1 (fraud)
	0 (no fraud)	€ 0	€ -25
	1 (fraud)	€ -5	€ 5

The retailer earns 5 euros profit for each correctly identified case of fraud and for every fraud attempt he loses 5 euros. The wrongly accused customer might not return to this store, which is denoted by a loss of 25 euros for the retailer. An honest customer identified correctly means neither loss nor profit for the retailer. The accuracy of a model is defined as a sum of the cost or profit of all scans. For testing purposes dataset was random split into training and validation sets (77% and 33% accordantly) (Fig. 3).

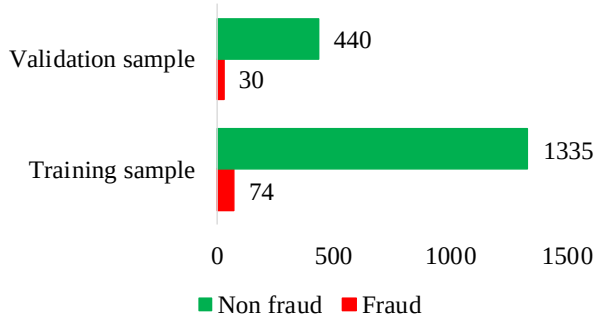

Fig. 3. Dimensionality of the fraud and non-fraud cases in training and validation samples

Results. The method's parameters used for modelling are:

- Ito decomposition of the second degree;
- Correlation coefficients for each attribute (Table 3);
- The number of trees in the Random Forest algorithm are 300;
- The maximum depth of the tree is 10;
- The seed used by the random number generator is 50;
- The variable that indicates whether to use out-of-bag samples to estimate the R2 on unseen data is true.

TABLE III. OBTAINED CORRELATION COEFFICIENTS

Attribute	Correlation coefficient
trustLevel	-0.320
totalScanTimeInSeconds	0.110
grandTotal	0.001
lineItemVoids	0.063
scansWithoutRegistration	0.074
quantityModifications	-0.001
scannedLineItemsPerSecond	-0.023
valuePerSecond	-0.029
lineItemVoidsPerPosition	-0.090

According to these parameters, based on Table 3 proposed method show the next results (Table 4).

TABLE IV. COST MATRIX OBTAINED USING PROPOSED METHOD

Actual value	Prediction		
		0 (no fraud)	1 (fraud)
	0 (no fraud)	440	0
	1 (fraud)	10	20

COMPARISON AND DISCUSSION

Results obtained by proposed methods were compared with existing ones:

- basic Random Forest algorithm;
- Stochastic Gradient classifier;
- Multilayer perceptron;
- Decision tree;
- AdaBoost.

The results of such comparison in training and test modes are describe in Tables 5 and 6.

TABLE V. RESULTS OF ALL METHOD IN TRAINING MODE

Training mode	AUC	Average precision	Profit
Random Forest	1.000	1	370
Stochastic Gradient classifier	0.500	0.05	-370
Multilayer perceptron	0.870	0.65	-45
Decision tree	0.770	0.52	-70
AdaBoost	1.000	1	370
Proposed method	**1.000**	**1**	**370**

TABLE VI. RESULTS OF ALL METHOD IN TEST MODE

Test mode	AUC	Average precision	Profit
Random Forest	0.700	0.44	-30
Stochastic Gradient classifier	0.500	0.06	-150
Multilayer perceptron	0.800	0.52	-70
Decision tree	0.640	0.19	-285
AdaBoost	0.800	0.52	-70
Proposed method	**0.830**	**0.69**	**50**

The basic parameters of the existing methods are shown in [8].

As can be seen from Table 5, *Stochastic Gradient classifier* is demonstrating the worst result according to the profit measure. *Multilayer perceptron and Decision tree* also show not very good results. The best results in the training mode are shows an ensemble method: *Random Forest, AdaBoost and Proposed one*.

However, in the application mode, taking into account the imbalance of the data set, the situation changes. All investigated methods, except those developed, show negative results of the cost matrix. This means considerable financial loss when used them for solving the task.

In the case of application of the investigated methods, *Decision tree* shows the worst result. It follows the *Stochastic Gradient classifier*. *Multilayer perceptron* and *AdaBoost* methods show the same unsatisfactory results. As far as financial losses are concerned, these methods are not the ones that should be used to solve the task.

The best result is obtained using the developed

method. The closest to it, however, with negative results, is the basic *Random Forest algorithm*.

That is why the comparison of the effectiveness of the operation of the *basic Random Forest algorithm* and the *Developed method* with different values of the Random Forest parameters is carried out in the work.

Fig. 4-7 shows the AUC curve when changing the trees number and depth of the tree for existing *Random Forest algorithm* and *Proposed method*.

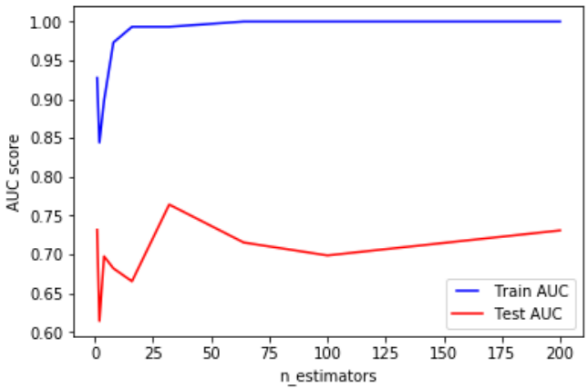

Fig. 4. Effectiveness of the *Random Forest algorithm* using different values of the trees number (from 1 to 300) in training and test modes

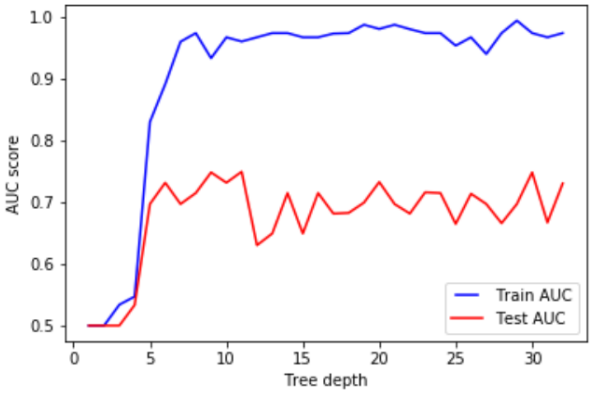

Fig. 5. Effectiveness of the *Random Forest algorithm* using different values of depth of the tree (from 1 to 30) in training and test modes

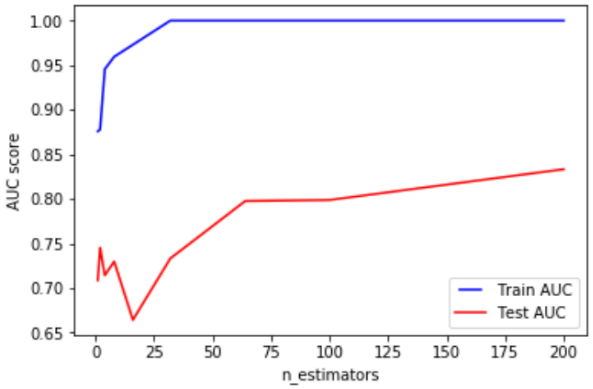

Fig. 6. Effectiveness of the *Proposed Method* using different values of the trees number (from 1 to 300) in training and test modes

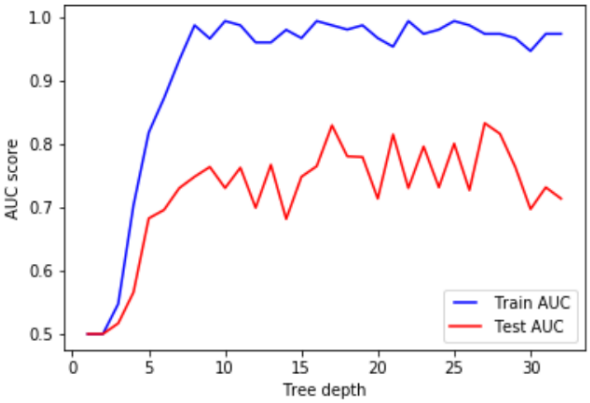

Fig. 7. Effectiveness of the *Proposed Method* using different values of depth of the tree (from 1 to 30) in training and test modes

As can be seen from Fig 4 – Fig. 7, the best results obtained using Proposed Method.

CONCLUSIONS

The effective solution of the fraud detection task in retail is one of the most important in this domain. The application of the machine learning based methods to it help to reduce risks of incorrect detection of fraud.

The authors are proposed a new ensemble-based method for fraud detection at self-checkouts in retail in condition of unbalanced data set. Algorithmic implementation of the developed methods consists of the sequential execution of set of procedures over the input data: scaling by maximal element in a column with row-wise scaling by Euclidean norm, weighting by correlation and applying polynomial extension and then – applying Random Forest algorithm.

The experimental comparison of the effectiveness of the developed ensemble-based method with a number of existing (basic Random Forest algorithm; Stochastic Gradient classifier; Multilayer perceptron; Decision tree; AdaBoost) demonstrates the best performance of the developed method.

REFERENCES

1. **Molnár E., Molnár R., Kryvinska N., Greguš M. 2014.** Web Intelligence in practice. The Society of Service Science, Journal of Service Science Research, Springer, Vol. 6, No. 1: 149-172.
2. BEETLE / iSCAN EASY SCO. URL: https://www.dieboldnixdorf.com/en-us/retail/systems/self-checkout-solutions/beetle-iscan-easy-sco (last accessed 10.06.2019)
3. **Kryvinska N. 2012.** Building Consistent Formal Specification for the Service Enterprise Agility Foundation. The Society of Service Science, Journal of Service Science Research, Springer, Vol. 4, No. 2: 235-269.
4. **Kaczor S., Kryvinska N. 2013.** It is all about Services - Fundamentals, Drivers, and Business Models. The Society of Service Science, Journal of Service Science Research, Springer, Vol. 5, No. 2, 2013: 125-154.

5. **Gregus M., Kryvinska N. 2015.** Service Orientation of Enterprises - Aspects, Dimensions, Technologies. Comenius University in Bratislava, ISBN: 9788022339780.
6. **Kryvinska N., Gregus M. 2014.** SOA and it's Business Value in Requirements, Features, Practices and Methodologies. Comenius University in Bratislava, ISBN: 9788022337649.
7. **Wang S. 2010.** A Comprehensive Survey of Data Mining-Based Accounting-Fraud Detection Research. International Conference on Intelligent Computation Technology and Automation, Changsha: 50-53.
8. **Izonin I. et. All 2018.** The Combined Use of the Wiener Polynomial and SVM for Material Classification Task in Medical Implants Production. International Journal of Intelligent Systems and Applications, Vol.10, No.9: .40-47.
9. **Tepla T.L., et all. 2018.** Alloys selection based on the supervised learning technique for design of biocompatible medical materials. Archives of Materials Science and Engineering, vol. 1, no. 93: 32–40.
10. **Tepla T., Izonin I., Duriagina Z.** 2019. Biocompatible materials selection via new supervised learning methods. LAP Lambert Academic Publishing, Riga, Latvia, 114 p.
11. **Vitynskyi P. et al. 2018.** Hybridization of the SGTM Neural-like Structure through Inputs Polynomial Extension. In: Proceedings of the 2018 IEEE Second International Conference on Data Stream Mining & Processing, 21–25 August 2018, Lviv, Ukraine, 2018: 386-391.
12. DATA MINING CUP 2019. URL: https://www.data-mining-cup.com/dmc-2019/ (last accessed 10.06.2019).
13. **Gruszczyński K. 2019.** Enhancing business process event logs with software failure data. Econtechmod. Vol 8, no 1: 27-32.
14. **Anokhin M., Koryttsev I. 2015.** Decision-making Rule Estimation with Applying similarity Metrics. Econtechmod. Vol 4, no 3: 73-78.