Bank i Kredyt 53(6), 2022, 587-604

BANK I KREDYT
BANK & CREDIT

Symbolic data analysis as a tool for credit fraud detection

Andrzej Dudek*, Marcin Pełka#

Submitted: 28 April 2021. Accepted: 18 August 2022.

Abstract

It can be said that the money fraud problem is as old as money itself. The development of new technologies allows criminals to develop new ways of fraud and also provides new methods to prevent them. The process of identifying if a newly authorised transaction is a case of fraudulent or genuine transaction is called fraud detection (Maes et al. 2002). Many classical methods can be used to detect money frauds. This paper proposes to apply symbolic data analysis methods, which allow describing objects in a more precise and complex way in order to handle the credit card fraud detection problem. The main hypothesis is that the decision tree for symbolic data is a better tool in credit card fraud detection than other methods.

Symbolic data analysis, unlike classical data analysis, allows describing objects in a more complex way. Symbolic data analysis makes it possible to take into account all variability and uncertainty in the data and provides suitable methods and techniques to deal with such data (see: Bock, Diday 2000; Billard, Diday 2006).

The first part is the introduction that describes the problem of credit card fraud detection and presents literature that deals with this problem. The second part presents the basic ideas of symbolic data analysis, describes all the models that will be applied in the empirical part (decision tree for symbolic data, logistic regression for symbolic data, *k*-nearest neighbour method for symbolic data and kernel discriminant analysis for symbolic data). The third part presents the results of credit card fraud detection. The data set containing 284,807 different card transactions (492 being fraud transactions) is used to build all models. The obtained results show that decision trees usually lead to slightly better results than other methods in the symbolic data case (for a single model). The last part presents the final remarks.

Keywords: credit card, fraud detection, symbolic data, machine learning, R software

JEL: G2, C02, C19, C38

* Uniwersytet Ekonomiczny we Wrocławiu; e-mail: andrzej.dudek@ue.wroc.pl; ORCID: 0000-0002-4943-8703.
# Uniwersytet Ekonomiczny we Wrocławiu; e-mail:marcin.pelka@ue.wroc.pl; ORCID: 0000-0002-2225-5229.

1. Introduction

It is obvious that frauds are as old as humanity itself. Moreover, the development of new technologies leads to new ways for criminals to commit new types of fraud (Bolton, Hand 2002). As the use of credit cards is now prevalent, credit card frauds have kept on growing. As frauds can lead to severe financial losses, those losses may become significant enough to affect not only customers (bank clients) but also banks themselves.

All actions against card fraud can be divided into fraud prevention, with attempts to nip fraudulent transactions in the bud, and fraud detection, where fraud transactions are detected over time (*post factum*).

Many different methods can prevent card frauds – the most important ones being addressed verifications systems (AVS), the card verification method (CVM), and the personal identification number (PIN). AVS involves verifying the address with the zip code of the customer while CVM and PIN involve checking the numeric code that is keyed in by the customer (Dal Pozzolo 2015).

The process of identifying if a newly authorised transaction belongs to the case of fraudulent or genuine transactions is called fraud detection (Maes et al. 2002). The cost of fraud detection should not exceed the cost of losses due to fraudulent transactions (Quah, Sriganesh 2008).

There are many different papers that address the problem of credit card fraud detection. This part of the paper will offer a literature review.

Bhatla, Prabhu and Dua (2003) show that screening 2% of transactions can reduce fraud losses accounting for 1% of the total value of transactions. Dal Pozzolo (2015) analyses and proposes new machine learning techniques for credit card fraud detection. Maes et al. (2002) analyse the use of Bayesian neural networks in credit card fraud detection. Raj and Portia (2011) and also Sonepat and Sonepat (2014) compare and study different credit card fraud detection methods (like Bayesian neural networks, the hidden Markov model, genetic algorithms, decision trees, support vector machine methods and neural networks). In conclusion, they say that if any of the methods or combinations of them is used, the probability of fraud detection increases. Carcillo et al. (2018a) present and compare active learning strategies, and they analyse their fraud detection accuracies. And when precision is considered, semi-supervised stochastic approach with random labelling and max combining (MF-SR) returns the best results with a detection improvement of 3.21%. Dal Pozzolo et al. (2014) focus their paper on the problems of dealing with unbalancedness, non-stationarity and assessment issues. In particular, they use two techniques (SMOTE and EasyEnsemble) and show that they can both improve the performance of classical undersampling. They have also used ensemble learning techniques and have demonstrated that combining information from different models improves the performance with undersampling but is not better than single models where SMOTE or EasyEnsemble is used before training.

Lebichot et al. (2019) deal with deep transfer learning approaches to credit card fraud detection and focus on transferring classification models learned on a specific category of transactions (e-commerce) to another. Best results were obtained when an adaptation of the method proposed by Ganin et al. (2016) is used.

Carcillo et al. (2018b) present a scalable real-time fraud finder (SCARFF) that integrates big data tools with a machine learning approach. Their experimental results on a massive data set of real credit card transactions show their framework to be scalable, efficient and accurate over a big stream of transactions.

Dal Pozzolo et al. (2015) analyse how undersampling affects the posterior probability of a machine learning model. The paper by Dal Pozzolo et al. (2017) tries to formalize the problem of fraud-detection and proposes a learning strategy that effectively addresses class imbalance, concept drift and verification latency. Their results reveal the crucial role of feedback in credit-fraud detection. Their experiments show that solutions which lower the influence of feedbacks in the learning process (like classifiers that mix feedbacks and delayed supervised samples or that implement instance weighting schemes) often return less precise alerts (Dal Pozzolo et al. 2017).

Shirgave et al. (2019) presented a review on credit card fraud detection made with the use of machine learning methods. They selected a supervised learning technique random forest to classify alerts as fraudulent or authorized. This classifier is trained using feedback and a delayed supervised sample. Next it aggregates each probability to detect alerts. Similar work was done by Varmedja et al. (2019), Maniraj et al. (2019) and Sailusha et al. (2020), and Lim, Lee and Sim (2021) where different machine learning methods are compared in credit card fraud detection. Carcillo et al. (2021) present a comparison of supervised and unsupervised learning methods for credit card fraud detection.

Makki et al. (2019) present an experimental study with imbalanced classification approaches to credit card fraud detection. They compared eight machine learning methods. They found out that logistic regression, the C5.0 decision tree, support vector machines and artificial neural networks were the best methods according to three considered performance measures (accuracy, sensitivity and AUPRC).

Chen and Lai (2021) present the application of a deep convolution neural network model to credit card fraud detection. The deep learning technique offers high accuracy and quick pattern identification for detecting sophisticated and unknown patterns. This model addresses the inefficiency issues of the existing models. The existing machine learning models, auto-encoder models and other deep learning models are compared with the proposed model for performance evaluation by incorporating a real-time credit card fraud dataset. Asha and Suresh (2021) presents an application of an artificial neural network in credit card fraud detection. Authors compare an artificial neural network with other models (e.g. a support vector machine, a *k*-nearest neighbour classifier) and find out that the neural network makes it possible to detect 100% of credit card frauds.

Khatri, Arora and Agrawal (2020) present a comparison of supervised machine learning methods for credit card fraud detection. Hussein et al. (2021) use an ensemble classification model based on the fuzzy-rough nearest neighbour algorithm, sequential minimal optimization, and logistic regression for credit card fraud detection. The proposed ensemble allows obtaining better results in terms of the detection rate, the false alarm rate, specificity, the positive predictive value, the f-measure, ROC curves and the AUC area.

Also papers written by Polish authors deal with problems of fraud detection and fraud prevention. Some of them focus more on the role of the statutory auditor and its role in fraud detection (see for example: Szczepankiewicz 2016; Żukowska-Kalita 2017; Rydzak 2016; Bartoszewicz, Bartoszewicz 2016; Lew 2016). Other papers present the role of statistical data analysis and the role of data analysis in general (see for example Nowak 2013; Tomanek 2014). Tomanek (2014) uses the minimum covariance determinant (MCD) estimator, where function covMCD from robustbase package of R is used. Lach (2021) presents the problem of identity theft, credit card frauds, etc. from the perspective of the judicial system. Miarzyńska (2021) compared various machine learning methods used in credit card fraud detection.

When dealing with symbolic data analysis in general, there are no works concerning credit card fraud detection. Similar problems present within the domain of credit scoring are presented by Dudek (2013) and Pełka (2018; 2019). Papers by Dudek (2013) and Pełka (2018) use kernel discriminant symbolic data analysis, symbolic decision trees, the *k*-nearest neighbour method and the multi-model approach. In both articles, the error of the model is used to assess the model. The lowest error value was obtained for random forests and then for a decision tree. In these papers, the least suitable method turned out to be the *k*-nearest neighbour method for symbolic data. The paper by Pełka (2019) presents an adaptation of decision stumps (decision trees with only one split) in credit scoring. It was found that the application of such an approach may be useful when building an ensemble (combined) model for credit scoring.

Symbolic data analysis allows taking into account the variability and uncertainty of data in the case of credit card fraud detection. For example, it is possible to take into account all the amounts withdrawn in a certain time frame or in the case of data anonymization we can use various approaches to model the data.

When analysing the possible benefits arising from the application of symbolic data to credit fraud detection, it is certainly worth noting the possibility of a full description of objects with the use of symbolic data of various types. In the case of symbolic data we are able to take into account all the variability and all the uncertainty in the data (e.g. by using symbolic interval-valued variables, symbolic histogram variables, etc.). Besides, such symbolic data analysis provides a variety of different methods that can be used for credit fraud detection.

The main aim of this paper is to apply symbolic data analysis methods (decision trees, kernel discriminant analysis, the kNN for symbolic data and logistic regression for symbolic data with various approaches to this method) in credit card fraud detection. The work has adopted the hypothesis that the application of symbolic data may be a useful way for credit card fraud detection in practice.

2. Methodology

Each symbolic object can be described by an assortment of variables, both classical (nominal, ordinal, ratio, interval) and symbolic, like symbolic interval-valued variables, symbolic histogram variables, symbolic multivariate variables (symbolic multinominal variables), and symbolic multivariate variables with weights or frequencies (symbolic multinominal variables with weigths). Besides that, a symbolic object can take into account relations between variables, as symbolic taxonomic variables do (see, for example, Bock, Diday 2000, pp. 2–3; Billard, Diday 2006, pp. 7–30; Diday, Noirhomme-Fraiture 2008, pp. 10–19).

More details on symbolic objects and symbolic variables can be found in Bock, Diday (2000, pp. 2–8), Billard, Diday (2006, pp. 7–66), Diday, Noirhomme-Fraiture (2008, pp. 3–30), Dudek (2013, pp. 42–43).

When considering symbolic data analysis methods that can be useful for credit card fraud detection, there are many different methods and approaches available. The most important methods include decision trees for symbolic data, kernel discriminant analysis for symbolic data, the *k*-nearest neighbour method for symbolic data and logistic regression for symbolic data. All of these methods will be described and used in the empirical part of the paper.

2.1. Decision trees for symbolic data[1]

In the case of symbolic data analysis and decision tree models, we have classification trees requiring a nominal dependent variable and a set of explanatory variables, which can be either classical variables of any type or symbolic interval-valued variables, or symbolic multinominal variables without weights (see Gatnar, Walesiak 2011, pp. 282–285).

The following steps are essential in order to build a symbolic classification tree (see Gatnar, Walesiak 2011, pp. 282–285):

1. Data collection and the construction of a symbolic data table.

2. For symbolic multinominal or classical variables that are nominal or ordinal, a frequency table is built. This table presents a count of the frequency with which a given category is observed in a dataset. For symbolic interval-valued variables the arithmetic mean (midpoint) is calculated for each possible combination of upper and lower bounds of the variable ranges.

3. Two stop criterions must be selected – one is the node size n^* (number of objects in a node), and the second one is the split criterion W^*. If the number of elements in a node is lower than n^*, then it's a terminal node (a leaf). If the split criterion is greater than W^*, we can use this division to split the tree.

4. Calculation of the probability that the objects will be assigned to the left node of the tree, $p_k(l)$, must be estimated.

A. For symbolic interval-valued, ratio and interval variables, we use the mean of the distributions calculated in point 2 above. These midpoints are used as cutting values c. The probability of assigning this item to the left node is expressed as follows (see Garnar, Walesiak 2011, pp. 283–284; Dudek 2013, p. 153):

– if the value c is located within the symbolic interval-valued variable, then:

$$p_k(l) = \frac{c - \underline{v}_{kj}}{\overline{v}_{kj} - \underline{v}_{kj}} \tag{1}$$

where:

$k = 1,\ldots,n$ – the number of the symbolic object,
$\underline{v}_{kj}$ – the lower bound of a symbolic variable,
$\overline{v}_{kj}$ – the upper bound of a symbolic variable,

– if the value c is located below the lower bound for the symbolic interval-valued variable, then:

$$p_k(l) = 0 \tag{2}$$

– if the value c is located above the upper bound for the symbolic interval-valued variable, then:

$$p_k(l) = 1 \tag{3}$$

B. For ordinal, nominal and multivariant variables, the cutting value c belongs to particular variable data categories (excluding the last category). The frequencies of variable values that are lower than c and those that are higher should be added for each item. Likewise, the c-value constitutes a distinct variable category for nominal variables. For a given item, the c-value is equal to the frequency of the category.

[1] This part of the paper is based on the paper by Pełka (2019).

5. The probability of assigning an item to the right node is $p_k(r) = 1 - p_k(l)$.

6. The calculation of the quality criterion for assigning the W node to each c point (see Garnar, Walesiak 2011, p. 284; Dudek 2013, p. 154):

$$W_j(t, c) = \log \prod_{k=1}^{n} \left[p_k(l) P_l(s) + p_k(r) P_r(s) \right] \tag{4}$$

where:

$j = 1, \ldots, m$ – the variable number,

t – the node number,

c – the cutting value,

$p_k(l)$ – the probability of assigning the k-th item to the left node,

$p_k(r) = 1 - p_k(l)$ – the probability of assigning the k-th item to the right node,

$P_l(s)\left(P_r(s)\right)$ – the conditional probability of observing the class to which the k-th item belongs on the left node (this is the quotient of the sum of probabilities of assigning all items to this node).

7. The choice of the highest W values for each variable.

8. The choice of the W variable that is greater than W^* and splitting the node according to the method appropriate for the given variable, on the condition that the size of node n is greater than that of n^*. Where one of these conditions has not been fulfilled, the node may not be divided further and is thus a final node.

9. Steps 6–8 should be repeated for each node until the final nodes are obtained. At a later stage of tree construction, the questions dealt with in earlier steps are not addressed.

10. Visualising and interpreting the results.

2.2. Kernel discriminant analysis for symbolic data

Kernel discriminant analysis for symbolic data uses an intensity estimator instead of a kernel density estimator, which is used in the classical data case (see Härdle, Simar 2003 p. 27). As with symbolic data, we cannot use any density estimation and since the symbolic data space is not a subspace of the Euclidean space, we cannot use the integral operator either. Instead, the intensity estimator for symbolic data is applied (Rasson, Lissoir 2000, pp. 241–242).

The intensity estimator for symbolic data is defined as follows (Rasson, Lissoir 2000, p. 242; Gatnar, Walesiak 2011, p. 281):

$$\hat{I}_r(A_i) = \frac{1}{n_r} \sum_{k=1}^{n_r} \prod_{f=1}^{F} K_{A_i, h_f}(A_{kr}) \tag{5}$$

where:

$\hat{I}_r(A_i)$ – the intensity estimator of the i-th symbolic object and the r-th cluster,

n_r – the number of objects in the r-th cluster,

$r = 1, \ldots, u$ – the cluster number,

$f = 1, \ldots, F$ – the number of a distance measure that is used in the kernel (as many different distance measures can be applied as is appropriate for the symbolic variables),

A_{kr} – the *k*-th object that belongs to the *r*-th cluster,
h_f – the bandwidth parameter (a *prior* set parameter) associated with the *f*-th distance measure,
$K_{A_i,h_f}(A_{kr})$ – the kernel for the *i*-th symbolic object and bandwidth parameter h_f and the *k*-th object that belongs to the *r*-th cluster.

For symbolic data intensity estimator is defined as follows (Rasson, Lissoir 2000, p. 242; Gatnar, Walesiak 2011, p. 281):

$$K_{A_i,h_f}(A_{kr}) = \begin{cases} 1 & \text{for} \quad d_{ik} < h_f \\ 0 & \text{for} \quad d_{ik} \geq h_f \end{cases} \tag{5}$$

where d_{ik} is the distance measure for symbolic objects.

More about distance measurement for symbolic data can be found in Gatnar and Walesiak (2011, pp. 18–23), Malerba, Esposito and Monopoli (2002), Bock and Diday (2000, pp. 139–197).

The posterior probability of assigning the *i*-th object to the *r*-th cluster is defined as follows (Rasson, Lissoir 2000, p. 244; Gatnar, Walesiak 2011, pp. 281–282):

$$q_r(A_i) = \frac{\hat{p}_r(A_i)\hat{I}_r(A_i)}{\sum_{r=1}^{u}\hat{p}_r(A_i)\hat{I}_r(A_i)} \tag{6}$$

where:

$r = 1,\ldots,u$ – the cluster number,
$q_r(A_i)$ – posterior probability that the *i*-th object will be assigned to the *r*-th cluster,
$\hat{p}_r(A_i)$ – prior probability that the *i*-th object will be assigned to the *r*-th cluster,
$\hat{I}_r(A_i)$ – the intensity estimator for the *i*-th symbolic object and the *r*-th cluster.

Prior probabilities can be calculated as one of three possible formulas (Rasson, Lissoir 2000, pp. 242–243; Gatnar, Walesiak 2011, p. 282):

a) they can be equal for each cluster, so $\hat{p}_r(A_i) = \frac{1}{u}$ they can take into account the number of objects in a cluster,
b) $\hat{p}_r(A_i) = \frac{n_r}{n}$ where n_r is the number of objects in *r*-th cluster, *n* is the total number of objects in a dataset,
c) they can be estimated as follows:

$$\hat{p}_r(t+1) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\hat{p}_r(t)\,\hat{I}_r(A_i)}{\sum_{r=1}^{u}\hat{p}_r(t)\hat{I}_r(A_i)}\right) \tag{7}$$

where:

n – the total number of objects in a dataset *n*, iteration,
$\hat{p}_r(0) = \frac{1}{u}\ \hat{I}_r(A_i)$ – the intensity estimator for the *i*-th symbolic object and the *r*-th cluster.

Rasson and Lissoir (2000, p. 241) suggest that, usually, ten iterations (with eq. 7) are sufficient to obtain prior probabilities.

Objects are assigned to the cluster with the maximum value of the posterior probability $(q_r(A_i))$.

2.3. *K*-nearest neighbour method for symbolic data

As in the symbolic data case, we have different types of symbolic variables that can describe objects, so some assumptions in the *k*-nearest neighbour method are made (see Pełka 2010):

a) symbolic objects cannot be treated as points in a hyperdimensional space,

b) a distance measure for symbolic data has to be applied,

c) the contribution of each neighbour is weighted with respect to its distance to the classified object (so-called importance of neighbours),

d) the number of the nearest neighbours can be extracted on the basis of the cross-validation of the training data,

e) as the result of the *k*-nearest neighbour method we get posterior probabilities of assigning objects to clusters.

The estimation of the posterior probabilities can be done as follows (Pełka 2010, p. 173–174):

a) if distances from the classified object A_i and all its k neighbours are equal to 0 and all neighbours are from the same cluster (r) the posterior probability is 1,

b) if distances from the classified object A_i and all its k neighbours are equal to 0, but neighbours are from different clusters, then the posterior probability is calculated as follows:

$$p\left(A_i\right)=\frac{n_r}{n_k} \tag{8}$$

where:

n_r – the number of neighbours from the r-th class,

n_k – the total number of neighbours considered,

c) if distances from the classified object and its neighbours differ from 0, the posterior probability is estimated by:

$$p\left(A_i\right)=\frac{\frac{n_r}{n_k}\cdot\Omega_r}{\sum_{r=1}^{u}\frac{n_r}{n_k}\cdot\Omega_r} \tag{9}$$

where:

$$\Omega_r=w_i\cdot\delta\left(A_i,A_k\right) \tag{10}$$

$$w_i=\frac{1}{d\left(A_i,A_k\right)} \tag{11}$$

where:

w_i – weights associated with the i-th object to be classified, they reflect "the importance of the k-th neighbour",

$\delta\left(A_i,A_k\right)=1$ if the k-th neighbour belongs to the same cluster to which we assign the i-th object,

$\delta\left(A_i,A_k\right)=0$ if the k-th neighbour does not belong to the same cluster to which we assign the i-th object,

$d\left(A_i,A_k\right)$ – the distance measure for the i-th object and its k-th neighbour.

2.4. Logistic regression for symbolic data

The logistic regression for symbolic data uses a binary dependent variable (e.g. y – credit fraud status for a credit card transaction: 1 – fraudulent transaction and 0 – non-fraudulent transaction) and symbolic interval valued explanatory variables. The general multivariate logistic regression can be expressed as follows:

$$Y_t = b_0 X_{0t} + b_1 X_{1t} + \ldots + b_m X_{mt} + e_t = \sum_{j=0}^{m} b_j X_{jt} + e_t \tag{12}$$

where:

Y – the dependent variable,
$X_0, X_1, \ldots, X_m$ – explanatory (dependent) variables,
$b_0, b_1, \ldots, b_m$ – model coefficients,
e – the model error,
$t = 1, \ldots, T$ – the observation number,
$j = 0, 1, \ldots, m$ – the variable number.

When we are using logistic regression, we assume that we are dealing with a latent y^* variable that can't be observed directly. Nevertheless, we have the following information:

$$y_i = \begin{cases} 1 \text{ if } y^* > 0 \\ 0 \text{ if } y^* \leq 0 \end{cases} \tag{13}$$

Finally, the probability that the variable y_i will be 0 or 1 is defined as follows:

$$P_i = F\left(x_i^{\mathrm{T}} b\right) = \frac{1}{1 + \exp\left(-x_i^{\mathrm{T}} b\right)} = \frac{\exp\left(x_i^{\mathrm{T}} b\right)}{1 + \exp\left(-x_i^{\mathrm{T}} b\right)} \tag{14}$$

In the case of symbolic interval-valued variables which are represented by the lower and upper bounds of the variable as $[\underline{x}_i, \overline{x}_i]$, where $\underline{x}_i$ is the lower bound of an interval-valued variable, and $\overline{x}_i$ is the upper bound of an interval-valued variable, we need to change the well-known classical logistic regression model. De Souza, Queiroz and Cysneiros (2011) describe four different approaches when dealing with symbolic interval-valued variables (de Souza, Queiroz, Cysneiros 2011, pp. 275–278):

1. The IDPC-CSP method, where the pattern classifier for interval data is based on a single posteriori probability defined by centres of the intervals $\frac{\underline{x}_i + \overline{x}_i}{2}$. Then the probability (3) is estimated for centres of intervals.

2. The IDPC-SP method, where the pattern classifier for interval data is based on a single posteriori probability defined by the lower and upper bounds of the intervals conjointly. And in equation (3) both lower and upper bounds are used as one dataset.

3. The IDPC-PP approach, where the pattern classifier for interval data is based on a pooled posteriori probability defined by the lower and upper bounds of the intervals separately. The final probability is then calculated as the mean value of both estimations.

4. The IDPC-VSP approach, where the pattern classifier for interval data is based on a single posteriori probability defined by the vertices of the hypercubes. The combination of lower and upper bounds for m symbolic interval-valued variables $\left[\underline{x}_{i1}, \overline{x}_{i1}\right], \ldots, \left[\underline{x}_{im}, \overline{x}_{im}\right]$ is known as the **M** matrix, defined as follows (de Souza, Queiroz, Cysneiros 2011, p. 276):

$$\mathbf{M} = \begin{bmatrix} \underline{x}_{i1} & \cdots & \underline{x}_{im} \\ \underline{x}_{i1} & \cdots & \overline{x}_{im} \\ \vdots & \ddots & \vdots \\ \overline{x}_{i1} & \cdots & \underline{x}_{im} \\ \overline{x}_{i1} & \cdots & \overline{x}_{im} \end{bmatrix} \tag{15}$$

The final probability in the vertices method is calculated as the mean, the maximum or minimum probability for these combinations (de Souza, Queiroz, Cysneiros 2011, p. 277).

Usually, the approaches involving centres, lower bounds or upper bounds obtain better results than the vertices method for symbolic interval-valued logistic regression (de Souza, Queiroz, Cysneiros 2011).

The empirical part of the present paper will be based on these first three approaches.

3. Authors' study on credit fraud detection

This paper uses the dataset collected by Université Libre de Bruxelles[2] and it contains 284,807 credit card transactions with only 492 frauds in it. A classical version of this data set can be obtained from the Kaggle webpage (https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud).

This classical dataset is imbalanced and contains only numerical (ratio) variables that are the result of the PCA transformation meant to make these data anonymous; only the variable amount was not transformed by the PCA. The 'Variable' class corresponds to fraudulent transactions (value 1) or non-fraudulent transactions (value 0). This dataset will be used for the symbolic decision tree for symbolic data, the kernel discriminant analysis for symbolic data, the kNN for symbolic data and logistic regression models (with centres, and upper and lower bounds used separately and together). The classical version of this data set (containing only classical numeric variables) has been used in many papers, e.g. Dal Pozzolo et al. (2015), Dal Pozzolo et al. (2014), Dal Pozzololo (2017), Carcillo et al. (2018).

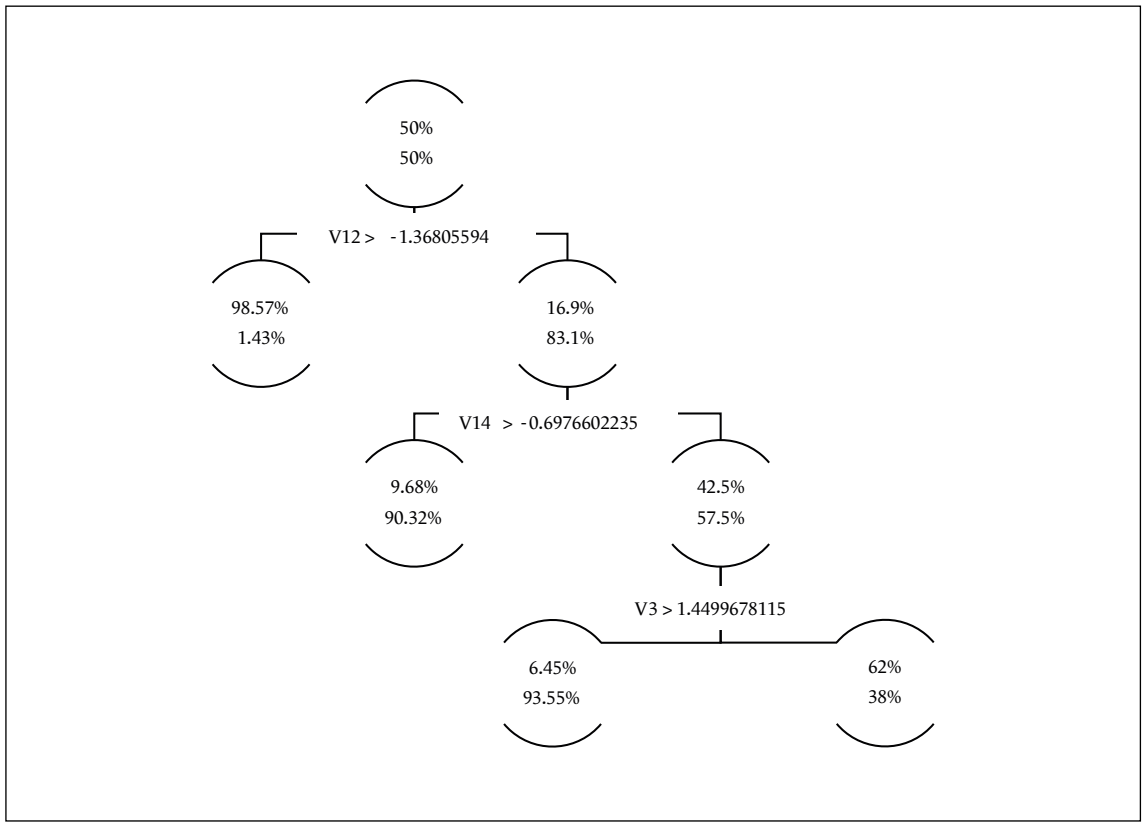
As the data set is highly imbalanced, all the 492 fraud transactions have been selected, and from the rest of the data set another 500 objects (non-fraud transactions) were drawn randomly with replacement. Such a modified data set was used as the new initial data set for all methods to be used. The sampling part was repeated 20 times, and the results were averaged. The new data set was then randomly divided into a learning part (2/3 of the data) and a testing part (1/3 of the data), while keeping the new data set structure.

[2] This dataset (the classical one) was collected and analysed during a research collaboration of Worldline and the Machine Learning Group of the Université Libre de Bruxelles (ULB) on big data mining and fraud detection. More info can be found at URL: https://mlg.ulb.ac.be/wordpress/portfolio_page/defeatfraud-assessment-and-validation-of-deep-feature-engineering-and-learning-solutions-for-fraud-detection/ (access date 8 August 2022).

The first model was the symbolic decision tree. The "best decision tree" (in terms of the prediction error for the test set) is presented in Figure 1. The error for this decision tree was equal to 0.3572139.

The averaged results for all the models are shown in Table 1.

Figure 1
Decision tree for credit fraud data



Source: own computation using the R software.

In the case of a single decision tree, the most important variable was the 12th variable, then the 14th and as the third most important came the third variable.

For example if the 12th symbolic interval-valued variable for a symbolic object contains –1.36805594 within the variable range, then the probability that the object will be assigned to the left node is calculated according to equation 1. For instance, if we assume that the interval is $\left[-2,\ 0\right]$ then this probability is equal to $p_k(l)=\frac{-1.36805594+2}{0+2}=0.31597$ and the probability that it will be assigned to the right node is $p_k(r)=1-p_k(l)=1-0.31597=0.68403$. But if the interval-valued variable's lower bound is greater than the *c* value, then this object will not be assigned to the left node. If the interval-valued variable's upper bound is below the *c* value, it will be assigned to the left node.

Due to the nature of the data set we have no knowledge of what these variables represent, it is hard to state anything more precise about them. However, very low midpoint values of variable 12 are characteristic of fraud transactions. Higher midpoint values of variable V14 are characteristic for non-fraud transactions. Midpoint values of variable V3, which are lower than 1.45, are characteristic mostly for fraud transactions, while higher values mean that we are dealing with non-fraudulent transactions.

Studying Table 1 we notice that the decision tree for symbolic data produces the best results in terms of error. The kNN for symbolic data, which was based on 10 neighbours, came second, while logistic regression, where lower and upper bounds of symbolic variables were used separately, took the third place. The worst results were obtained in the case of the logistic regression where lower and upper bounds were used jointly. Table 2 contains averaged misclassification tables for all the analyzed methods.

Table 1
Average results for the models in the new balanced dataset

Model (method)	Essential parameters	Error	Correct predictions
Symbolic decision tree	Default values of the function decisionTree.SDA from *R* software	0.359526	0.640474
Kernel discriminant analysis for symbolic data	Ichino-Yaguchi distance h (bandwidth parameter) = 1.5*	0.454695	0.545305
KNN for symbolic data	10-nearest neighbours	0.39065	0.60935
Logistic regression for symbolic data	Centres method	0.426353	0.573647
	Lower and upper bounds separate estimation	0.400293	0.599707
	Lower and upper bounds joint estimation	0.4766832	0.523317

* Arbitrary choice made upon the distances in the data set.

Source: own computation using the R software.

Asha and Suresh (2021) found out that for classical data credit card fraud detection it is the artificial neural network and the kNN for classical data that provide the best results in terms of prediction accuracy. This is similar to our second-best method – the kNN for symbolic data. Chen and Lai (2021) compared deep learning networks with support vector machines (SVMs), the logistic regression model and the random forest. They found out that logistic regression leads to lower accuracy, despite the reduction in temporal duration, and the random forest is similar to the deep learning approach in terms of accuracy but it takes more time. Accuracy and speed of SVMs are comparatively lesser than the deep learning neural network's. Khatri, Arora and Agrawal (2020) compared various machine

learning models using the classical version of the credit card fraud data set. They found out that maximum precision is obtained by the random forest, the kNN comes next in line and the decision tree and logistic regression follow. The maximum sensitivity is obtained by the Naïve Bayes classifier, then the kNN and finally the decision tree and random forest models.

Table 2
Misclassification tables for all analysed methods

Method	Misclassification table*		
Symbolic decision tree		fraud	non-fraud
	fraud	320.2	89.4
	non-fraud	270.1	320.2
Kernel discriminant analysis for symbolic data		fraud	non-fraud
	fraud	272.7	103.4
	non-fraud	123.95	272.7
kNN for symbolic data		fraud	non-fraud
	fraud	304.7	341.02
	non-fraud	113.7	304.7
Logistic regression for symbolic data – centres method		fraud	non-fraud
	fraud	286.8	257.7
	non-fraud	168.6	286.8
Logistic regression for symbolic data – lower and upper bounds method (separate estimation)		fraud	non-fraud
	fraud	299.8	242.03
	non-fraud	158.3	299.6
Logistic regression for symbolic data – lower and upper bounds method (joint estimation)		fraud	non-fraud
	fraud	261.6	267.8
	non-fraud	208.9	261.6

* Average values for 20 models.

Source: own elaboration with the application of R software.

Lim, Lee and Sim (2021) did a similar comparison of various methods of machine learning for a classical dataset. They suggest that neural networks and support vector machine models are much more capable to learn and adapt to new fraud patterns. Indeed, other supervised Machine Learning algorithms like the kNN, the Naïve Bayesian and the Decision Tree algorithms are not effective in detecting new frauds and they required a more comprehensive re-training process onto the new data. This is true also in our case, because symbolic models (e.g. decision trees) use similar ideas as classical models and the problem of re-training appears in the symbolic data as well. Makki et al. (2019) have also compared various models in credit card fraud detection and found out that the highest accuracy is obtained for the classical decision tree (C5.0), support vector machines, artificial neural networks and logistic regression. The kNN for classical data provided satisfactory results too.

Since in the case of symbolic data the literature has not formulated a proposal involving SVMs or deep learning neural networks, we can presume that a decision tree used as a part of a random forest is an adequate tool for credit card fraud detection, and so is the kNN for symbolic data as it bears resemblance to the classical kNN. The logistic regression for symbolic data does not perform as well as the classical logistic regression model. This may derive from the fact that in the case of the symbolic logistic regression model we do not use symbolic interval-valued data but other approaches (transformation methods) instead and these can lead to some information loss – see subsection 2.4 for details.

4. Final remarks

Unlike classical data analysis, where each object is represented as a vector of quantitative or qualitative measurements, symbolic data analysis makes it possible to take into account the uncertainty and variability of the data. This allows describing objects in a more complex way but on the other hand it requires new methods and algorithms capable of dealing with such data types (see: Bock, Diday 2000; Billard, Diday 2006).

The present paper shows how to apply four different methods of symbolic data analysis (symbolic decision trees, kernel discriminant analysis for symbolic data, logistic regression for symbolic data – with various estimation approaches used – and the *k*-nearest neighbours method for symbolic data) to credit fraud detection.

As the data set used in the empirical part is highly imbalanced, 492 fraud transactions from the initial data set were selected and, what is more, 500 non-fraud transactions were randomly drawn (with replacement). This procedure was repeated 20 times and the results were averaged. The best results, in terms of the error rate and the rate of correct predictions, were provided by the symbolic decision tree. Slightly worse results were obtained in the case of the kNN for symbolic data (with 10 neighbours). The worst results were produced by logistic regression for symbolic data where lower and upper bounds of interval-valued variables were used together (jointly) for model estimation.

References

Asha R.B., Suresh K.S.K. (2021), Credit card fraud detection using artificial neural network, *Global Transitions Proceedings*, 2(1), 35–41.

Bartoszewicz A., Bartoszewicz S. (2016), *Implementacja systemu zwalczania nadużyć finansowych przez Instytucję Zarządzającą na potrzeby audytu desygnacyjnego*, Studia i Prace Kolegium Zarządzania i Finansów, Szkoła Główna Handlowa, 152, 53–72.

Bock H.-H., Diday E., eds (2000), *Analysis of Symbolic Data. Explanatory Methods for Extracting Statistical Information from Complex Data*, Springer Verlag.

Billard L., Diday E. (2006), *Symbolic Data Analysis. Conceptual Statistics and Data Mining*, John Wiley & Sons.

Bhatla T.P., Prabhu V., Dua A. (2003), Understanding credit card frauds, *Cards Business Review*, 1(6), 1–15.

Bolton R.J., Hand D.J. (2002), Statistical fraud detection: a review, *Statistical Science*, 235–249.

Carcillo F., Le Borgne Y.A., Caelen O., Kessaci Y., Oblé F., Bontempi G. (2021), Combining unsupervised and supervised learning in credit card fraud detection, *Information Sciences*, 557, 317–331.

Carcillo F., Le Borgne Y.-A., Caelen O., Bontempi G. (2018a), Streaming active learning strategies for real-life credit card fraud detection: assessment and visualisation, *International Journal of Data Science and Analytics*, 5(4), 285–300.

Carcillo F., Dal Pozzolo A., Le Borgne Y.-A., Caelen O., Mazzer Y., Bontempi G. (2018b), Scarff: a scalable framework for streaming credit card fraud detection with spark, *Information Fusion*, 41, 182–194.

Chen J.I.Z., Lai K.L. (2021), Deep convolution neural network model for credit-card fraud detection and alert, *Journal of Artificial Intelligence*, 3(02), 101–112.

Dal Pozzolo A. (2015), *Adaptive machine learning for credit card fraud detection*, Ph. D. thesis, Université Libre de Bruxelles, Computer Science Department.

Dal Pozzolo A., Boracchi G., Caelen O., Alippi C., Bontempi G. (2017), Credit card fraud detection: a realistic modeling and a novel learning strategy, *IEEE Transactions on Neural Networks and Learning Systems*, 29(8), 3784–3797.

Dal Pozzolo A., Caelen O., Johnson R.A., Bonetempi G. (2015), *Calibrating probability with undersampling for unbalanced classification*, 2015 IEEE Symposium Series on Computational Intelligence.

Dal Pozzolo A., Caelen O., Le Borgne Y.A., Waterschoot S., Bontempi G. (2014), Learned lessons in credit fraud detection from a practitioner perspective, *Expert Systems with Applications*, 41(10), 4915–4928.

Dudek A. (2013), *Metody analizy danych symbolicznych w badaniach ekonomicznych*, Wyd. UE we Wrocławiu.

Ganin Y., Ustinova E., Ajakan H., Germain P., Larochelle H., Marchand M., Lempitsky V. (2016), Domain-adversarial training of neural networks, *Journal of Machine Learning Results*, 17(1), 2096–2030.

Gatnar E., Walesiak M., ed. (2011), *Analiza danych jakościowych i symbolicznych z wykorzystaniem programu R*, C.H. Beck.

Härdle W., Simar L. (2003), *Applied Multivariate Statistical Analysis*, Springer Verlag.

Hussein A.S., Khairy R.S., Najeeb S.M.M., Alrikabi H.T. (2021), Credit card fraud detection using fuzzy rough nearest neighbor and sequential minimal optimization with logistic regression, *International Journal of Interactive Mobile Technologies*, 15(5), 24–42.

Khatri S., Arora A., Agrawal A.P. (2020), *Supervised machine learning algorithms for credit card fraud detection: a comparison,* 10th International Conference on Cloud Computing, Data Science & Engineering.

Lebichot B., Le Borgne Y.-A., He-Guelton L., Oblé F., Bontempi G. (2019), Deep-learning domain adaptation techniques for credit cards fraud detection, in: L. Oneto, N. Navarin, A. Sperduti, D. Anguita (eds), *Recent Advances in Big Data and Deep Learning. Proceedings of the INNS Big Data and Deep Learning Conference INNS BDDL2019, held at Sestri Levante, Genova, Italy, 16–18 April 2019*, Springer.

Lach A. (2021), *Karnoprawna reakcja na zjawisko kradzieży tożsamości*, Wolters Kluwer Polska.

Lew A. (2016), Ryzyko istotnego zniekształcenia jako element badania przychodów i kosztów przez biegłego rewidenta, *Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu*, 440, 363–371.

Lim K.S., Lee L.H., Sim Y.W. (2021), A review of machine learning algorithms for fraud detection in credit card transaction, *International Journal of Computer Science & Network Security*, 21(9), 31–40.

Maes S., Tuyls K., Vanschoenwinkel B., Manderick B. (2002), Credit card fraud detection using Bayesian and neural networks, in: *Proceedings of the First International NAISO Congress on NEURO FUZZY TECHNOLOGIES*, 16–19 January, Havana, Cuba.

Makki S., Assaghir Z., Taher Y., Haque R., Hacid M.S., Zeineddine H. (2019), An experimental study with imbalanced classification approaches for credit card fraud detection, *IEEE Access*, 7, 93010–93022.

Malerba D., Esposito F., Monopoli M. (2002), *Comparing dissimilarity measures for probabilistic symbolic objects*, in: A. Zanasi, C.A. Brebbia, N.F.F. Ebecken, P. Melli (eds), *Data Mining III*, WIT Press.

Maniraj S.P., Saini A., Ahmed S., Sarkar S. (2019), Credit card fraud detection using machine learning and data science, *International Journal of Engineering Research*, 8(9), 110–115.

Miarzyńska Z. (2021), *Porównanie algorytmów uczenia maszynowego na przykładzie rozpoznawania oszustw w transakcjach płatniczych*, bachelor thesis, Jagiellonian University in Kraków.

Nowak A. (2013), Identyfikacja przyczyn i sprawców oszustw finansowo-księgowych – ujęcie statystyczne, *Finanse, Rynki Finansowe, Ubezpieczenia*, 61(2), 139–148.

Pełka M. (2010), K-nearest neighbour classification for symbolic data, *Acta Universitatis Lodziensis. Folia Oeconomica*, 235, 171–176.

Pełka M. (2018), Podejście wielomodelowe analizy danych symbolicznych w ocenie zdolności kredytowej osób fizycznych, *Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu*, 507, 200–207.

Pełka M. (2019), Symbolic decision stumps in individual credit scoring, *Bank i Kredyt*, 50(6), 513–528.

Quah J.T.S., Sriganesh M. (2008), Real-time credit card fraud detection using computational intelligence, *Expert Systems with Applications*, 35(4), 1721–1732.

Raj S.B.E., Portia A.A. (2011), *Analysis on credit card fraud detection methods*, 2011 International Conference on Computer, Communication and Electrical Technology (ICCCET), National College of Engineering.

Rasson J.P., Lissoir S. (2000), Symbolic kernel discriminant analysis, in: H.-H. Bock, E. Diday (eds), *Analysis of Symbolic Data. Explanatory Methods for Extracting Statistical Information from Complex Data*, Springer Verlag.

Rydzak R. (2016), *Rola biegłego rewidenta w wykrywaniu istotnych zniekształceń sprawozdania finansowego będących konsekwencją oszustwa*, Prace Naukowe, 9–19, Uniwersytet Ekonomiczny w Katowicach.

Sailusha R., Gnaneswar V., Ramesh R., Rao G.R. (2020), *Credit card fraud detection using machine learning*, 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), IEEE.

Shirgave S., Awati C., More R., Patil S. (2019), A review on credit card fraud detection using machine learning, *International Journal of Scientific & Technology Research*, 8(10), 1217–1220.

Sonepat R., Sonepat S. (2014), Analysis on credit card fraud detection methods, *International Journal of Computer Trends and Technology*, 8(1), 45–51.

de Souza R.M., Queiroz D.C., Cysneiros F.J.A. (2011), Logistic regression-based pattern classifiers for symbolic interval data, *Pattern Analysis and Applications*, 14(3), 273–282.

Szczepankiewicz E.I. (2016), *Rewizja finansowa, audyt wewnętrzny a audyt śledczy w wykrywaniu oszustw gospodarczych*, Studia i Prace Kolegium Zarządzania i Finansów, Szkoła Główna Handlowa, 152, 73–93.

Tomanek J. (2014), Analiza wielowymiarowa w wykrywaniu oszustw księgowych, *Studia Ekonomiczne*, 192, 155–169.

Varmedja D., Karanovic M., Sladojevic S., Arsenovic M., Anderla A. (2019), Credit card fraud detection – machine learning methods, in: *2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH)*.

Żukowska-Kalita J. (2017), Symptomy oszustw w sprawozdaniach finansowych i ich identyfikacja w procesie badania sprawozdania finansowego przez biegłego rewidenta, *Studia i Prace Kolegium Zarządzania i Finansów*, Szkoła Główna Handlowa, 154, 53–73.

Analiza danych symbolicznych jako narzędzie wykrywania oszustw z użyciem kart kredytowych

Streszczenie

Oszustwa finansowe nie są zjawiskiem nowym, a rozwój nowoczesnych technik i technologii, który z jednej strony poprawia standard życia, dostarcza także nowych narzędzi oszustom. Nowoczesne metody analizy danych dostarczają narzędzi, które pozwalają na szybsze i znacznie dokładniejsze identyfikowanie oszustw oraz przeciwdziałanie ich powstawaniu. Proces identyfikacji oszustw to w dużym uproszczeniu działania zmierzające do weryfikacji transakcji. Literatura przedmiotu wskazuje, że koszty takiej weryfikacji nie powinny przekraczać strat spowodowanych oszustwem (Quah, Sriganesh 2008).

W klasycznej analizie danych wypracowano wiele metod, które mogą być użyteczne w identyfikacji oszustw w transakcjach finansowych. Do metod tych zalicza się m.in. sztuczne sieci neuronowe, drzewa decyzyjne czy regresję logistyczną. W literaturze przedmiotu z zakresu wykrywania oszustw z użyciem kart kredytowych prezentowane oraz porównywane są zwykle metody klasycznej analizy danych. Zazwyczaj artykuły te prezentują zastosowanie takich metod, jak sztuczne sieci neuronowe, drzewa decyzyjne, podejście wielomodelowe czy regresja logistyczna i metoda k-najbliższych sąsiadów (zob. m.in.: Makki i in. 2019; Varmedja i in. 2019; Maniraj i in. 2019; Szczepankiewicz 2016; Miarzyńska 2021).

Niemniej jednak w literaturze przedmiotu brakuje artykułów prezentujących możliwość zastosowania różnych metod analizy danych symbolicznych do identyfikacji oszustw finansowych z użyciem kart kredytowych oraz porównujących te metody.

Artykuł stanowi uzupełnienie tej luki w literaturze przedmiotu. Proponuje się w nim zastosowanie metod analizy danych symbolicznych, pozwalających opisywać różne obiekty (na przykład transakcje kartami kredytowymi) w dokładniejszy i złożony sposób niż za pomocą danych klasycznych. Główna hipoteza brzmi, że drzewa decyzyjne dla danych symbolicznych są lepszym narzędziem wykrywania oszustw finansowych niż regresja logistyczna dla danych symbolicznych, metoda *k*-najbliższych sąsiadów dla danych symbolicznych czy jądrowa analiza dyskryminacyjna danych symbolicznych.

W analizie danych symbolicznych, w przeciwieństwie do analizy danych klasycznych (gdzie obiekty opisywane są przez pojedyncze zmienne – metryczne lub niemetryczne), można opisywać obiekty w dokładniejszy, złożony sposób. Dostarcza ona odpowiednich narzędzi do analizowania złożonych zbiorów danych, m.in. o dużej zmienności (zob. Bock, Diday 2000; Billard, Diday 2006).

Artykuł ma następującą strukturę. We wstępie scharakteryzowano problem wykrywania oszustw finansowych oraz dokonano przeglądu literatury z tego zakresu. Druga część prezentuje podstawowe idee i pojęcia z zakresu analizy danych symbolicznych oraz opisuje metody, które zostaną zastosowane w części empirycznej (drzewa decyzyjne danych symbolicznych, regresję logistyczną danych symbolicz-

nych, jądrową analizę dyskryminacyjną danych symbolicznych oraz metodę *k*-najbliższych sąsiadów dla danych symbolicznych). Trzecia część artykułu (empiryczna) prezentuje wyniki wykrywania oszustw finansowych. Zastosowano tu zbiór danych zawierający 284 807 transakcji kartami kredytowymi (tylko 492 były oszustwami). Otrzymane wyniki wskazują, że drzewa decyzyjne danych symbolicznych zwykle pozwalają osiągnąć lepsze wyniki niż inne metody analizy danych symbolicznych (dla modeli pojedynczych). Artykuł opatrzono podsumowaniem.

Słowa kluczowe: karty kredytowe, oszustwa kartami kredytowymi, dane symboliczne, uczenie maszynowe, program R