Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 190 (2021) 210–216

Procedia Computer Science

www.elsevier.com/locate/procedia

2020 Annual International Conference on Brain-Inspired Cognitive Architectures for Artificial Intelligence: Eleventh Annual Meeting of the BICA Society

Detection of non-typical users of the electronic marketplace "Freight transportation" to prevent the competitive intelligence

Jenny Domashova*, Anastasia Zasypkina

National Research Nuclear University "MEPhI", 31 Kashirskoe shosse, Moscow 115409, Russia

Abstract

This article presents results of usage of cluster analysis and machine learning to detect non-typical users of the electronic freight transportation marketplace. The most significant red flags to identify system users who collect commercial information have been identified, a competitor's profile has been drawn up. The use of various methods of clustering and classification in the modeling of this problem is studied. The procedure of formation of a training sample in the absence of labeled data has been carried out. The algorithm has been developed to detect non-typical users based on machine data analysis methods in order to prevent the real-time commercial intelligence. This algorithm is based on density-based spatial clustering of applications with noise (DBSCAN) and balanced iterative reducing and clustering using hierarchies (BIRCH) cluster analysis methods, identification of anomalous objects using the Isolation Forest method and ensemble Gradient Boosting algorithm with decision trees as the base classifiers. It can be used by electronic marketplace's owners, as well as by other companies that apply similar technologies while providing their services. As a result, software tool in Python has been developed, which allows to timely detect users collecting commercial information and hence to reduce a negative impact from possible competitive intelligence.

Keywords: competitive intelligence; machine learning algorithms; anomaly detection; cluster analysis; classification

* Corresponding author Jenny Domashova.
E-mail address: janedom@mail.ru

1877-0509
10.1016/j.procs.2021.06.026

1. Introduction

Competitive intelligence is a recognized and widely used tool in business, most often aimed at aggravating competitor's negative trends, while the rest of the marketing is searching for new niches. In this regard, many companies need to be able to defend themselves against it. In addition to statistical estimates of service costs and finding out strategy, competitive intelligence can be used to resell the services of a competing company at inflated prices. It should be noted that such activity entails not only poaching customers, but overloading the system, resulting in unnecessary costs. [1]

Competitive intelligence is the collection and processing of data from public sources, conducted in compliance with the legal and ethical standards, for the support of management decisions to improve the competitiveness of a commercial organization. The technical purpose of competitive intelligence is to obtain the maximum amount of up-to-date information about supervised object. The result of the competitive intelligence is not to copy competitor's strategy, but to build its own long-term strategy based on the received information. Thus, competitive intelligence is mostly aimed at aggravating competitor's negative trends, therefore, identification of non-typical users of the resource, possibly involved in competitive intelligence is mostly relevant.

2. Materials and methods

The purpose of the study is to identify the typical characteristics of non-typical users with the help of machine learning algorithms based on the analysis of the data of the electronic marketplace.

Selection of competitive intelligence methods is individual for each organization. Currently, the main way of collecting open-source information is to study a competing organization's web applications / platforms. [2]

The structure of the electronic marketplace "Freight transportation" (EM FT) system is three-tier and consists of the following main components: application server, communication server and Database Management System (DBMS) server. Before using the services of the EM FT, the user must be identified. The main business process in which the platform user participates is ordering a service. Its main participants are client, the platform and the service providers. The service ordering process consists of four main stages: registration / identification of the client, service pre-order, ordering and its execution. Client identification consists of the user entering credentials, confirming these data and verifying or registering this client in the external system. After registration / identification, the client can make preliminary calculations of services. To do this, he needs to select the transportation data, after which the system generates a list of services, sends requests to suppliers and offers options to the client. The final stage is the formation of an order, in which the client is directly involved only in payment.

The most complicated problem is that initially there is no information about the characteristics of the user, who can collect business information. Formally, there is a task of unsupervised learning, but the list of data for processing is not initially known. Therefore, it is necessary to compile a profile of a competitor of the EM FT platform.

It should be noted that if the payer in the order is the same organization that placed the order, then this customer is most likely a representative of the forwarding organization. Forwarders are the main clients of the platform, which means that they are very unlikely to be involved in competitive intelligence. Whereas a user who collects commercial and technical information, but is not interested in paying and placing orders, is most likely a competitor.

Thus, the competitor of the EM FT platform is characterized by the following actions:

- Preliminary calculations are often made in the same directions, with the same cargo and types of rolling stock
- The total number of pre-orders is high
- The client's organization is engaged in forwarding activities with zero probability
- The client or his organization does not pay for the placed orders
- The client or his organization rarely places orders (chooses a profitable option in the list of preliminary calculation services)
- The periods of activity of such users are continuous
- Frequent orders through the automated control system (ACS)

In order to form a feature space describing client's behavior the Entity-Relationship model (ER-model) of client-related EM FT data was formed. Basic tables are Orders, Pre-Orders, Document Operations and Registered Users.

In order to solve the task, the following step-by-step procedures are required:

1. Formation of a feature space.
2. Dimension Reduction in feature space.
3. Detection of atypical users.
4. Interpretation of atypical users based on clustering and perceptions of possible actions of competitors on the platform.
5. Classification of users.
6. Comparison of received models.
7. Description of resulting composed model.

To solve the stated problems, various methods of filling gaps in the data, cluster analysis, and classification methods were examined and applied.

The gaps were filled using a special value - the minimum value for the attribute. The preliminary cluster analysis was carried out using the K-means method, which is a simplified analogue of the EM algorithm. First, the initial approximation of the centers of the clusters is given. [3] BIRCH was used to interpret atypical objects in the sample. The BIRCH process begins with a hierarchical division of objects using a tree structure and applies other clustering algorithms to refine the clusters. [4] The first step in classifying a user is to determine whether he is anomalous via the Isolation Forest method. Isolation Forest is an unsupervised detection algorithm for anomalous objects. Conventional methods are not optimized for detecting anomalies; instead, they are optimized for finding normal instances which cause the anomaly detection result in either too many false positives or too few outliers. [5] The final classification of users into good, bad, and suspicious is done by the ensemble gradient boosting algorithm. Gradient boosting is a machine learning technique for regression and classification problems that creates a prediction model in the form of an ensemble of weak prediction models, usually decision trees. [6]

3. Results

The features describing EM FT user’s behavior are formed on the basis of ER-model.

Characteristics of groups of selected features are the following:

- Basic user’s information (user’s organization, the lifetime and activity of the account)
- The average time interval between changes in the status of the main documents
- Coefficient of variation of the main pre-order parameters selected by the user
- Number of main documents in different status
- The nature of the calculation of pre-orders by the user
- Number of additional order parameters
- The nature of orders payment

For example, the group "The nature of orders payment" shows the probability of the platform user being a forwarder, for instance, how often this customer is a direct payer and pays himself or through an advance, and so on. In total, 23 features were calculated for 6037 users (sample size) that existed at the time of the formation of the features.

At the stage of correlation analysis, 6 features were excluded. The heat map of the correlation matrix visualization is shown in the figure 1.

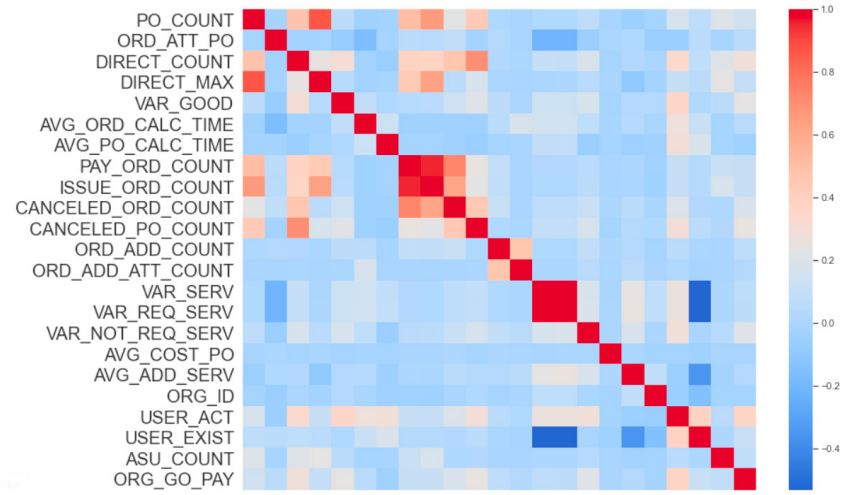

Fig. 1 - The heat map of the correlation matrix visualization

Next, a simple K-means clustering into 10 clusters was performed for temporal marking of objects and subsequent allocation of significant features. Subsequent dimensionality reduction was carried out by selecting significant features. The table 1 shows the feature sets selected by the methods SelectKBest, Recursive Feature, and ExtraTreesClassifier. [7]

Table 1. Sets of significant features.

Method	SelectKBest	RFE	ExtraTreesClassifier
Significant features	po_count, org_go_pay, pay_ord_count, user_act, avg_po_cost, direct_max	user_act, po_count, pay_ord_count, org_go_pay, asu_count, direct count	po_count, pay_ord_count, direct_max, user_act, asu_count, direct_count

The final sample of features was formed based on the results of the methods presented above and the competitor's profiles. The final selected features include: the number of calculated pre-orders, the number of paid orders, the time of account activity, the number of pre-orders calculated using ACS, forwarding activities the organization, a number of different directions in the pre-orders.

4. Discussion

To construct a training sample, you need to mark users. Let's assume that atypical users are not just noise, but anomalous objects to sample. The feasibility of this assumption can be verified by describing the outliers of the final sample.

To construct the target variable, it was assumed that atypical users are anomalous objects for sampling. The DBSCAN method was used to identify anomalous objects in the resulting sample. [8] The parameters of the algorithm: *eps = 0.005*, metric is the square of the Euclidean distance. The objects were divided into 7 clusters and outliers (anomalous objects). The outliers identified as a result of clustering contain both the "best "(because they have a high average number of paid orders) and the "worst" users (because they have a high average number of

unpaid orders), which suggests that users operating on the platform for the purpose of collecting commercial information are highly likely to be anomalous objects for the sample.

It was decided to identify atypical users using anomalous outlier detection algorithms. Due to the characteristic provided by DBSCAN models for detecting anomalous objects were trained. Using the Isolation Forest method, 897 anomalous objects were detected, 785 elliptic envelopes and 512 OneClassSVM. The minimum number of unpaid pre-orders indicates how many suspicious customers are included in the sample, whereas the number of users with a large number of paid orders shows how many "good" users there are. Figure 2 shows their comparison.

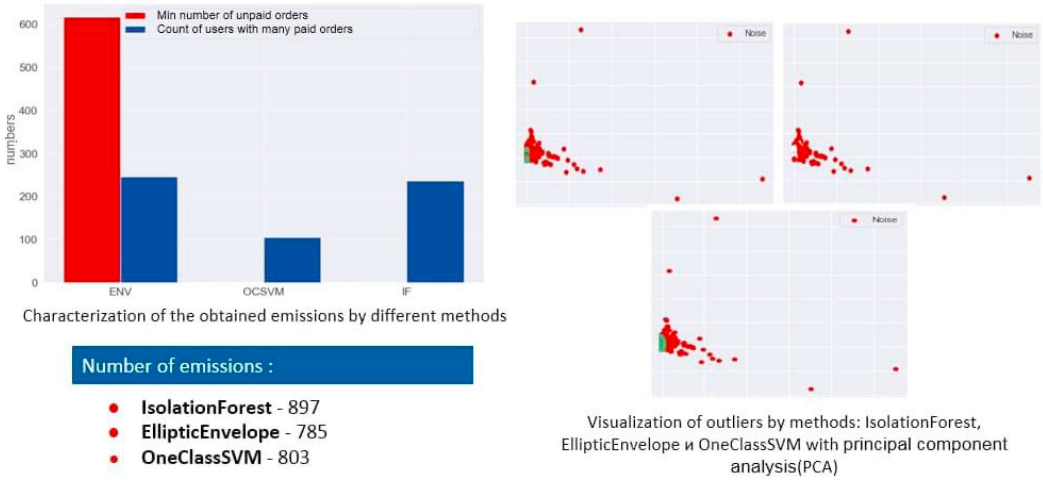

Fig. 2 – Detecting atypical users with machine learning techniques

To interpret the obtained samples with anomalous users, clustering was performed using two methods: BIRCH and K-means. The data was previously normalized. On average, the silhouette of the clusters obtained by BIRCH is higher. Using the BIRCH method, a cluster of "good" users was allocated. Such users pay for a relatively large number of pre-orders or their organization is highly likely to be engaged in forwarding activities (i.e. they are the main customers of the platform).

Business rules were developed to separate the remaining users into "suspicious" and "competitors". They are based on 4 features: the percentage of paid orders, the forwarding activity of the organization, the number of payments in the same directions and pre-orders for automated control systems (Figure 3).

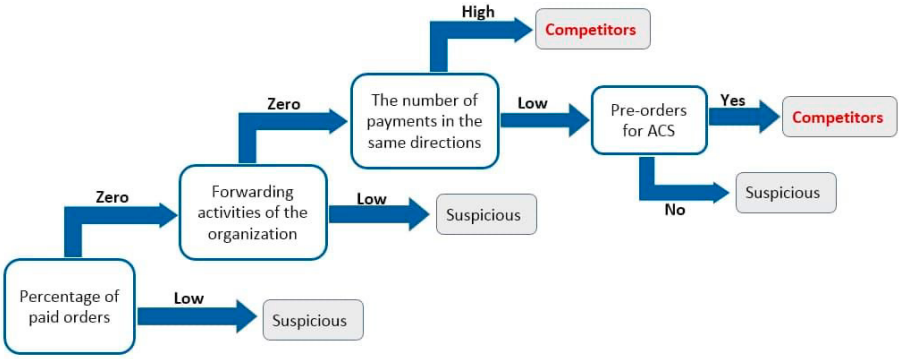

Fig. 3 - Business rules that identify suspicious users

Let's take a closer look at two features. If an organization with a high probability is a freight forwarder, then the employees of this organization are definitely good customers for the platform (freight forwarders are the system's main customers). If the user very often calculates services through the automated control system (just sends Simple Object Access Protocol (SOAP) requests to the system), then he may be suspected of monitoring the platform's prices.

Further user classification models are constructed for better interpretation. The training sample is unbalanced (~1:7 "competitors" in relation to "good"), so ensembles of classification algorithms were used: gradient boosting, adaptive boosting, and the voting classifier that combine Decision Tree Classifier and Support Vector Classification (SVC). Their main task is to use sampling data to train the model to determine the probability of an object being classified as a "competitor" by the target attribute. These classifiers output the class value (good, bad, suspicious). The parameters for each model were determined by the selecting hyperparameters among the stratified partitions of the training dataset. The results and model estimates are described in the figure 4.

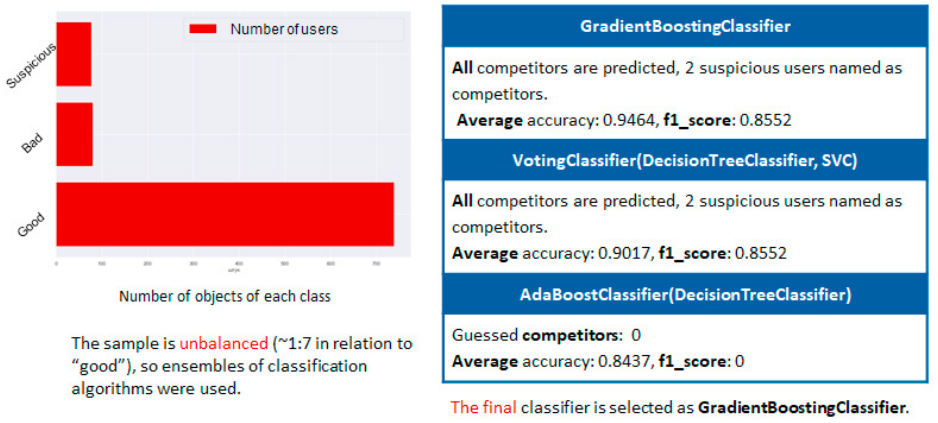

Fig. 4 - Number of objects of each class Classification of users and their interpretations, comparison of the obtained models

The evaluation of the obtained models is carried out using average accuracy and f1-score. All competitors are correctly predicted by gradient boosting methods and voting classifier, but at the same time two suspicious users are identified as a competitor, f1-score: 0.8552. The average accuracy of gradient boosting is 0.9464, and the voting classifier is 0.9017. Adaptive boosting does not recognize any competitors, respectively, the f1 parameter is exactly 0, the average accuracy in turn is 0.8437. Gradient Boosting is selected as the final classifier.

Final model consists of 3 main stages: calculation of the identified significant features, determination of the user's anomalies, and its classification. A schematic representation of the resulting model is shown in Figure 5.

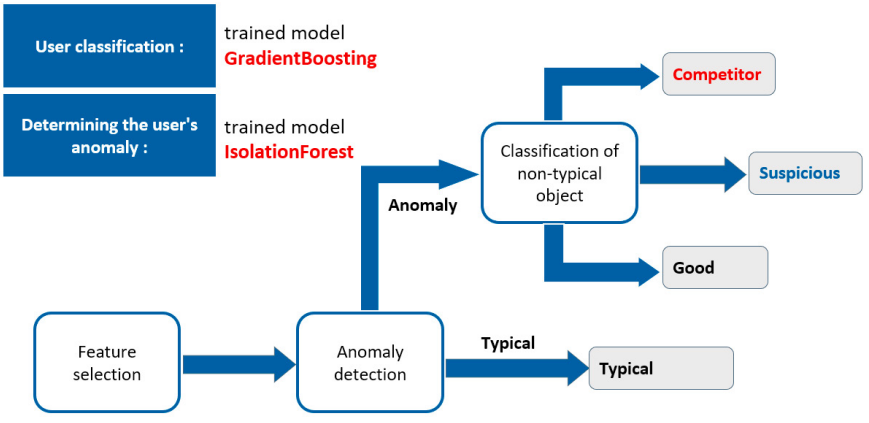

Fig. 5 – The structure of final model.

Depending on the requirements of the implementation of the software model, the structure may include the collection of data on significant signs of all users of the system and retraining of models for identifying anomalous objects and classification.

5. Summary

As a result of the work, the methods of conducting competitive intelligence on the Electronic marketplace "Freight transportation" were studied, the relevance of the work was determined, as well as atypical users of the portal are identified and described. A review of the risks associated with the collection of commercial information of the organization is carried out. The profile of the most interesting user — a competitor was created.

In addition, the analysis of data preprocessing and machine learning methods to solve the problem of identifying anomalous objects and classification was carried out. The review showed that the quality of the research results is greatly influenced by the choice of the necessary methods for solving problems, as well as the preprocessing of data for the analysis. Models for detecting anomalous objects and classification were constructed using a combination of different approaches, tested on different samples, and compared to determine the most effective model.

The initial feature space corresponding to the task was formed and calculated, the data was studied using machine analysis methods. As a result, the most informative features were identified, based on which the procedure for recognizing users of the electronic platform, including, presumably, engaged in competitive intelligence, was developed.

The article proposes a methodology that combines a system of business rules with models for identifying atypical users of an electronic platform based on cluster analysis and machine learning methods, the practical application of which will reduce the negative impact of possible competitive intelligence.

Acknowledgements

This work is supported by the National Research Nuclear University "MEPhI".

References

[1] Yushchuk E. L. Competitive intelligence: marketing risks and opportunities. - Yekaterinburg: PervoGrad, 2019 .- 264 p .;

[2] Electronic Marketplace “Freight Transportation”. Statutory documents. [Electronic resource] — URL: https://etpgp.rzd.ru/Documentation (date of the application: 03.01.2021);

[3] Everitt, Brian. Cluster analysis. — Chichester, West Sussex, U.K: Wiley, 2011. — 346 p.;

[4] Tian Zhang, Raghu Ramakrishnan, MironLinvy. BIRCH: an efficient data clustering method for large databases, International Conference on Management of Data. — Quebec: ACM-SIGMOD Montreal, 1996. —12 p.;

[5] Liu, Fei Tony & Ting, Kai & Zhou, Zhi-Hua. Isolation Forest. — Victoria, Australia: Gippsland School of Information Technology Monash University, 2009. — 10 p.;

[6] Friedman, J. H. Greedy Function Approximation: A Gradient Boosting Machine, [Electronic resource] // Education in the Statistics discipline acquaints. 1999. URL: https://statweb.stanford.edu/~jhf/ftp/trebst.pdf (date of the application: 21.12.2020);

[7] Guyon, Isabelle; Elisseeff, André (2003). "An Introduction to Variable and Feature Selection";

[8] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise // Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96) / Evangelos Simoudis, Jiawei Han, Usama M. Fayyad. — AAAI Press, 1996. — p. 226–231.

Authors

1. Jenny Domashova, Ph.D. (Econ.), Associate Professor, Institute of Financial Technology and Economic Security, NRNU "MEPhI", Moscow, Russia, janedom@mail.ru, ORCiD 0000-0003-1987-8553
2. Anastasia Zasypkina, student, Institute of Financial Technology and Economic Security, NRNU "MEPhI", Moscow, Russia, anzasypkina@mail.ru, ORCiD 0000-0002-5411-1270.