International Journal of Engineering & Technology, 7 (4.36) (2018) 76-79

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper



Application of Artificial Intelligence Technologies for the Monitoring of Transactions in AML-Systems Using the Example of the Developed Classification Algorithm

S.G. Magomedov[1], A.S. Dobrotvorsky[1], M.P. Khrestina[1], S.A. Pavelyev[1], T.R. Yusubaliev[2]

[1]MIREA - Russian Technological University, Russia, Moscow
[2]Quality Software Solutions ltd Moscow, Russia, Moscow

Abstract

The article describes the application of artificial intelligence technologies in Anti Money Laundering (AML) systems for the purpose of transaction monitoring by the example of the developed transaction classification algorithm using machine learning methods. To improve the effectiveness of the algorithm a novel mechanism for forming a unique set of characteristics for the transactions and the participants of financial processes has been developed and a method for constructing a transaction graph has been proposed.

Keywords: *system, transactions, transaction, ccomputational, resource, processing*

1. Introduction

Improving the mechanisms for countering money-laundering is a priority for many financial institutions. Successful operation of the AML (Anti Money Laundering) system requires the solution of the following problems:
- processing large amounts of data from several sources;
- identification of potential risks;
- making prompt decisions on emerging risks in real time.

Traditional methods of solving such problems are slow, time-consuming and costly in the face of constantly evolving fraud technologies and increasing volume of unstructured and heterogeneous information. In addition IT-tools used by most companies include both outdated systems and modern applications. This leads to fragmentation and disparity of information flows and increases the complexity of their processing [1].

The key task of the AML-system is monitoring of bank transactions in order to identify potentially dangerous ones. These measures can be strengthened by using artificial intelligence technologies [1]:

1. Introduction of machine learning methods into MDM (Master Data Management) technology in AML-systems makes it possible to systematize data obtained from different, sometimes conflicting sources, to analyze the interaction of data flows, associated events and related objects. Data is supplemented by comments and optimized for the convenience of their further use. All the data related to the customer are combined into a single pool, which provides comprehensive information in the framework of the AML-system.
2. The use of artificial intelligence technologies in the method of probabilistic reconciliation of data enables in-depth analysis of information in order to exclude repetitions in the knowledge base.
3. Intelligent analytics based on templates replaces the outdated analysis engine based on a set of rules. The latter does not allow tracking of illegal transactions in situations when the relationship between transaction participants or different types of transactions is not obvious.
4. Methods of fuzzy processing allow making adjustments in the decision-making system as new information is received or existing information is changed.
5. The use of machine learning methods [2,3,4] in the transaction classification algorithms allows:
- minimizing the number of missed suspicious transactions;
- reducing the probability of erroneous classification of transactions;
- reducing the processing time of large transaction flows in real time.

This article proposes a new algorithm for transaction classification using methods of machine learning and a graph-based approach for describing participants of financial processes, their attributes and transactions.

2. Research Method

The task of the proposed classification algorithm is the detection of suspicious bank transactions.

A set of suspicious transactions is selected, which is combined with the same number of normal transactions selected randomly. The resulting set of transactions is divided into training and validation sets. The training set is used for the machine learning process (model acquisition), and the validation set is intended to test the acquired model on transactions classified as suspicious or normal. In the process of learning cross-validation is also used.

Methods of machine learning used in the algorithm include logical regression, SVM (support vector machine), random forest.

Input data of the algorithm include database of suspicious transactions and database of all transactions. Output data of the algorithm consists of the model with the highest classification accuracy among all trained models obtained.

The algorithm includes the following stages:

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

1. Reading the data.
2. Construction of the graph of bank transactions.
3. Identification of communities for vertices involved in the formation of characteristics.
4. Formation of a set of characteristics for the transactions and the participants of financial processes.
5. Analysis of characteristics.
6. Training on a small part of the data using cross-validation and building of a model.
7. Evaluation of the quality of classification.
8. Application of the obtained model for calculating the result on the whole set of transactions.

A graph is a formal representation that most comprehensively reflects information about transactions, as well as relationships between transaction participants (individuals and legal entities). Sample information for building the graph of bank transactions is shown in Table 1.

Table 1. Information about transactions

Field name	Value
DealTransact	Transaction identifier
BankOtprID	Sender bank identifier
ClientOtpr	Sender data (organization, name, address or account number)
AccIDClientOtpr	Sender account identifier (without detailed data)
AccClientOtpr	Sender account number
RestAccClientOtpr	Additional sender data
BankPolID	Recipient bank identifier
ClientPol	Recipient data (organization, name, address or account number)
AccClientPol	Recipient account number
Date	Transaction date
RealQty	Transaction amount in rubles
Comment	Comment to transaction

Graph G has the following structure

$$G = (V_1 \cup V_2, E_1 \cup E_2) \quad (1)$$

where the set V1 of vertices of the graph is a set of individuals and legal entities with their attributes conducting bank operations (transactions); the set V2 of vertices of the graph is a set of individuals added from the database of legal entities; the set E1 of graph edges is a set of bank transactions from the transaction database with the necessary attributes; the set E2 of edges of the graph is a set of relations between transaction participants.

There are the following types of graph edges:

Type I - undirected edge without weight existing if the addresses of two vertices (transaction participants) are similar;

Type II - undirected edge reflecting the relationship between the founder and the organization, the weight of the edge equals the share of the founder in the authorized capital;

Type III - undirected edge without weight reflecting the relationship between the organization and its head.

The method for constructing a graph of participants of financial processes includes the following steps:

1. Creating the set E1 of graph edges based on the database of all transactions and highlighting unique vertices (set V1).
2. Adding vertices from the database of legal entities (set V2) to the set V1.
3. Construction of Type II and Type III edges of the set E2 on the basis of the description of the relationship between the founders, managers and organizations in the database.
4. Construction of Type I edges of the set E2 on the basis of grouping of individuals by their addresses. Transaction edges can be added to the graph dynamically.

3. Results and Analysis

For graph G the following classification of characteristics [5-7] exists according to hierarchy levels in the graph:

1. Vertex level:

1.1. Degree - the number of edges of the graph incident to the vertex.

1.2. Indegree/outdegree - the number of incoming/outgoing edges incident to the vertex.

1.3. Betweenness centrality [8-10] CB(v) in a graph G = (V, E) with n vertices represents the importance of a vertex in relation to all possible paths in the graph. CB(v) of the vertex v is given by the expression (2):

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}} \quad (2)$$

where σst is the number of shortest paths from the vertex s to the vertex t, σst(v) is the number of shortest paths from s to t going through the vertex v.

1.4. The PageRank characteristic represents the importance of a vertex in relation to all other vertices in the graph [8-10].

2. The level of a pair of vertices (edge level): transaction time; transaction size.

3. The level of a vertex group within the neighborhoods of radius 1 and 2 (all vertices located respectively at a distance of no more than 1 or no more than 2 edges from a given vertex): number of vertices in the neighborhood; minimum, maximum and average degrees of a vertex in the neighborhood; minimum, maximum and average indegree/outdegree; number of edges in the neighborhood; amount of transactions in the neighborhood.

It should be noted that the direction of the edges does not matter when constructing a neighborhood.

The listed characteristics are calculated for all vertices in the graph. Then for each transaction the characteristics of the sender and the recipient are selected and assigned to the edge (transaction). As a result, each transaction corresponds to a set of 52 characteristics.

The set of characteristics is defined obviously redundant, as it is impossible to predict at the stage of theoretical studies which characteristics are significant for identifying suspicious transactions. At the same time the calculation of characteristics is the most expensive operation in terms of time and resource consumption. Therefore it is advised to optimize the composition of the analyzed characteristics in relation to their contribution to identification of suspicious transactions, which will reduce the resource intensity of the system.

For example, the process of analyzing a bank transaction involving an entity A and a person B includes the following steps:

1. Defining all paths from A to B, since the presence of such path can be a sign of a suspicious transaction.
2. Finding connected components or strongly related components (cliques), because according to statistics suspicious transactions are usually committed within a certain group of entities - components or clique sub-graphs, in which edges exist between each pair of vertices.
3. Determining the betwenness centrality for each vertex.
4. Identifying other characteristics that can be assigned to each participant in the transaction:

- degrees of vertices in the graph of paths (maximum, minimum and average degrees can be useful for analysis);
- the existence of paths, the length of paths in the graph;
- the number of the connected component to which the vertex belongs;
- the age of accounts A and B (for example, in case they specifically created an account for laundering for a short period);

- the number of transactions (corresponding to different periods), the amount of transactions (how much money is being transferred) between A and B both directly and through different paths;
- indegree and outdegree for A and B (same meaning as the previous point, though not only between A and B, but corresponding to total number of individuals involved in transactions with them, amount of transactions, etc.);
- the characteristics obtained from all vertices on the paths connecting the vertices A and B;
- the parameters by which vertices are connected on all paths;
- the confidence coefficient for each object of the transaction graph.

The time efficiency of the proposed procedure for calculating the characteristics for the transactions and the participants of financial processes can be evaluated by estimating of its computational complexity. For this purpose a weighted oriented graph is considered [9-10]:

$$G = (V, E), |V| = N, |E| = M \tag{3}$$

The maximum degree of a vertex is defined as max_degree.

The analysis of computational complexity of calculating characteristics for each edge of the graph can be carried out as follows. Each edge has two incident vertices. The complexity of calculating characteristics for one vertex corresponds to the number of vertices in a community of radius 2 around the vertex, which amounts to max_degree2 operations in the worst case. It can be assumed that the computational complexity for one vertex of the graph is O(1). The number of vertices of the graph for which the characteristics are calculated corresponds to:

$$O(\min(N, M)) \tag{4}$$

Where min is the minimum function for two numbers, since if M is lower than N the characteristics can be calculated only for the significant vertices instead of all vertices.

If the calculations are carried out in parallel then, given the number of processes equals p, the computational complexity can be estimated as:

$$O(\min(N, M) / p) \tag{5}$$

Characteristics of the created graph are shown in Table 2.

Table 2. Characteristics of the created graph

Number of vertices	15 034 710
Number of edges	781 440
Number of vertices in the sub-graph of suspicious transactions	349
Number of edges in the sub-graph of suspicious transactions	715
Total number of objects (transactions) – size of the data set selected for machine learning	1430 (50% suspicious and 50 % normal transactions)
Size of the training set: class 0 objects (normal) class 1 objects (suspicious)	1019 524 411
Size of the validation set: class 0 objects (normal) class 1 objects (suspicious)	411 191 220

Based on the validation results the criteria of effectiveness of the developed algorithm have been calculated (Table 3).

Table 3. Criteria of effectiveness of the developed classification algorithm

Criteria name	Value
Accuracy – classification accuracy	90.53%
AUROC (Area Under ROC curve) - the area bounded by the receiver operating characteristic (ROC) curve and the false positive rate axis	0.971
Sensitivity - percentage of suspicious transactions that can be detected	0.868
Specificity - percentage of correctly classified normal transactions in relation to the total number of normal transactions	0.927
Precision - percentage of actually suspicious transactions among all transactions classified as suspicious	0.932
NPV (Negative Predictive Value) - percentage of correctly classified normal transactions in relation to the total number of transactions classified as normal	0.859
FNR (False Negative Rate) - model error for wrongly classified suspicious transactions	0.132
FPR (False Positive Rate) - model error for wrongly classified normal transactions	0.073
F1 score - harmonic mean of precision and sensitivity metrics	0.899

4. Conclusion

The conducted research leads to the following conclusions:

1. Using the technologies of artificial intelligence in AML-systems brings transaction monitoring to a qualitatively new level due to:
- possibility of processing large amounts of heterogeneous data;
- deep analysis of transactions which minimizes risks and allows identifying new types of fraud in real time;
- application of machine learning methods in classification algorithms.

2. The developed algorithm for classification of transactions based on machine learning methods has the following qualities:

2.1. Demonstrates high classification efficiency (Table 3).

2.2. Uses graph structure to describe information about transactions and transaction participants which allows:
- presenting most comprehensively the relationship between transactions and individual participants;
- presenting the complex set of characteristics of transactions and transaction participants using different types of edges;
- taking into account the relational nature of fraud during construction and analysis of the transaction graph, for example:

1) scenario "Possible fraud" - if an object commits fraud, it is likely that related objects can also commit fraud;

2) scenario "Organized fraud" - fraud committed in close cooperation with the relevant group;
- providing resistance to hacker attacks, since the structure of the transaction graph makes it difficult to locally change any part of it.

2.3. Uses a unique set of characteristics which has the following advantages:
- linear dependence of the computational complexity of calculation of this set of characteristics on the number of vertices and edges, which is a priority when processing large amounts of data. For example, if the number of vertices is doubled, the time for calculating the characteristics for the entire graph will increase by no more than two times with fixed amount of computational resources;
- calculation of this set of characteristics has the property of locality, which means there is no need to process the entire graph for its calculation, which is especially important for real time processing: it is possible to calculate the characteristics only for the vertices

being added, and the time for such calculation will not depend on the size of the entire graph, that is the size of the database of accumulated transactions.

Acknowledgment

The research is being conducted with the financial support of the Ministry of Education and Science of the Russian Federation (Contract №14.574.21.0142) Unique ID for Applied Scientific Research (project) RFMEFI57417X0142. The data presented, the statements made, and the views expressed are solely the responsibility of the authors.

References

[1] Using Artificial Intelligence and Machine Learning to Help Financial Institutions Increase Compliance with Know Your Customer (KYC) Regulations. 2017. P. 2. URL: www.h2o.ai/wp-content/uploads/2017/06/Know-Your-Customer_v3_pages.pdf (24.09.2018)

[2] Pozzolo A. D., Caelen O., Borgne Y.-A. L. et al. Learned lessons in credit card fraud detection from a practitioner perspective. Expert Systems with Applications. 2014. Vol. 41. P. 4915–4928

[3] Savage D., Wang Q., Zhang X. et al. Detection of Money Laundering Groups: Supervised Learning on Small Networks. 2017. URL: https://aaai.org/ocs/index.php/WS/AAAIW17/paper/view/15101 (24.09.2018)

[4] Voit A., Stankus A., Magomedov Sh., Ivanova I. Big data processing for full-text search and visualization with elasticsearch International Journal of Advanced Computer Science and Applications. 2017. T. 8. № 12. C. 76-83. DOI: 10.14569/IJACSA.2017.081211

[5] Savage D., Wang Q., Chou P. et al. Detection of money laundering groups using supervised learning in networks. 2016. URL: https://arxiv.org/pdf/1608.00708.pdf (24.09.2018)

[6] C. Suresh, K. T. Reddy, and N. Sweta, \A Hybrid Approach for Detecting Suspicious Accounts in Money Laundering Using Data Mining Techniques," Information Technology and Computer Science, pp. 37-43, 2016.

[7] Akoglu L., Tong H., Koutra D. Graph Based Anomaly Detection and Description: A Survey. Data Min. Knowl. Discov. 2015. Vol. 29, # 3. P. 626–688. URL: https://arxiv.org/pdf/1404.4679.pdf (24.09.2018)

[8] Magomedov Sh. Organization of secured data transfer in computers using sign-value notation. ITM Web of Conferences. 2017. T. 10. DOI: 10.1051/itmconf/20171004004

[9] Brandes U. A faster algorithm for betweenness centrality. Journal of mathematical sociology. 2001. Vol. 25, # 2. P. 163–177. URL: http://www.algo.uni-konstanz.de/publications/b-fabc-01.pdf (24.09.2018)

[10] Page L., Brin S., Motwani R. et al. The PageRank citation ranking: Bringing order to the web. Tech. Rep. Stanford InfoLab. 1999.