Anomaly Detection in Public Procurements using the Open Contracting Data Standard

Elisabeth Kehler[1] , Julio Paciello[2] and Juan Pane[3]

[1] Universidad Nacional de Asunción, Paraguay, maelikehler@gmail.com
[2] Universidad Nacional de Asunción, Paraguay, julio.paciello@pol.una.py
[3] Universidad Nacional de Asunción, Paraguay, jpane@pol.una.py

Abstract. Public procurement typically presents a potential source of corruption. For this reason, the detection of anomalies in public procurements can improve the quality of purchases, and consequently enable a better quality of life in the country through the correct use of public funds. In this paper, we use as a case study the public contracts of Paraguay, which are in the open data format of the Open Contracting Data Standard (OCDS), for training an unsupervised learning model for anomaly detection, based on the Isolation Forest algorithm. The resulting classification allows to obtain a measurement or scoring of contracts that can be used to identify outliers. Given a local dataset of cases of procurement processes with protests with judgments in favor of the protestant or with citizen complaints, the preliminary results show that the trained model classifies as anomalous more than 45% of the potentially anomalous dataset. A detailed validation considering the public procurements local regulations is needed, with the purpose of building a tool that allows an intelligent sampling of contracts with atypical behavior to review, applicable to Paraguay and other countries that implement the OCDS.

Keywords: Open Contracting Data Standard, Open data, Anomaly detection, Unsupervised learning, Artificial Intelligence.

1 Introduction

Transparency is an important tool to avoid corruption and is essential in a process of public procurements. The Open Contracting Data Standard (OCDS) [1] is a tool that helps to implement the transparency needed and provides the possibility of analyzing the data on a machine learning level. This work uses the publicly available data of the public procurements of Paraguay as a case study. The Public Procurements Office (DNCP) publishes since 2010 contracts in the OCDS open data format. The total number of contracts published by the DNCP since 2010 to 2019 amounts to 311,782.

Anomalies detection in public contracts is especially important in order to find, prevent and take actions of possibly misappropriated funds. These funds can then be redirected to areas with an important social impact, such as education or public health care. The regulatory analysis of the conformity of public procurement processes according to the local legislation is performed manually by a team of public officials of the DNCP, analyzing each procurement process separately. Given the volume of public contracts that are managed annually and the manual work of public officials to perform this task, it is possible to clearly notice that an exhaustive control of all

contracts is not feasible. In addition to the control carried out by the DNCP team, there are also journalistic publications on cases identified as possible frauds. However, journalistic investigations focus mostly on contracts that are potentially more striking for public opinion, which represent a small portion of the total.

Considering the public procurements as a potential source of corruption, and therefore a way for the misuse of the public funds, performing regulatory control to ensure that the processes are aligned with the local legislation represents an important task for a country. In Paraguay, to date, this analysis is done manually by specialized staff of the DNCP. They determine the classification of the data according to the local laws [8] and also considering known fraud schemes, as for example the Red Flags scheme [9]. So, the main problem is that having an ever growing amount of data, a proportional growing number of staff members to analyze the data is required. The use of the OCDS format makes it possible to apply Machine Learning techniques, as unsupervised learning, for anomaly detections of possible outliers to the expected behavior, serving the DNCP as an automated tool for implementing a smart sampling of procurement processes that can require an in-depth verification. In this work, this problem is addressed by proposing the automation of control tasks in a first instance, which allows the DNCP staff to obtain a smartly selected sample of relevant procurement processes for manual review.

The work is organized as follows, in section II the State of Art is mentioned, in section III we explain the proposed solution and in section IV we present the preliminary results and final discussion.

2 State of Art of anomaly detection techniques

Conti and Naldi in [4] present an statistical anomaly detection approach in procurement auctions using an average bid based method evaluating with the detection probability and the false alarm probability. Vaserhelyi and Issa illustrate K-Means Clustering applied to a labeled refund transactions dataset in [5]. Deng and Mei combine Self-Organizing Map (SOM) and K-Means Clustering for an unsupervised approach to detecting Fraudulent Financial Statements in [6]. Panigrahi, Kundu, Sural and Majumdar propose a fusion approach for credit card fraud detection using a rule-based filter, a Dempster-Shafer adder and a Bayesian Lerner in [7].

As we can appreciate there are multiple previous works that address the anomaly detection in certain stages of public procurements, and also in financial transactions. This work proposes to implement a tool that focuses on all stages of public procurement, during the call for bids, the award and contracts, and contract modifications specifically in the format of the OCDS. This approach also differs from the mentioned state of art by applying unsupervised isolation forests to determine the anomaly score of the data points.

Isolation Forest [2, 3] creates a forest of binary trees by randomly selecting a feature and also randomly selecting the split value at each node. The anomaly score is obtained by getting the length of the path to the data point in the isolation tree. It was

chosen for its independence to distance and density, especially because the data sets are high-dimensional, and for its good computational performance.

3 Proposed Solution

The semi-structured version of the data in the OCDS format allows for the implementation of algorithms to analyze it. The goal is to train a unsupervised learning model that separates anomalies from the data for their consequent human analysis to determine if a fraud is taking place.

As seen in Figure 1, first the data had to be cleaned, considering human errors when loading the data. Then a feature selection is done, selecting the variables based on three criteria: a) if the variable has data (is not empty), b) if the data is structured (E.g. no free text or links to text), and c) if the data is part of the local regulations for the procurement process. For an algorithm to be able to analyze the data efficiently, the data needs to be converted to numerical values, in this approach hashing and binary vectors are used to perform these tasks. The data is divided into 3 datasets, a) data in the planning and tender stages of the procurement process, b) contracts, and c) contracts modifications. Finally the data was normalized for use as an input to the algorithm.

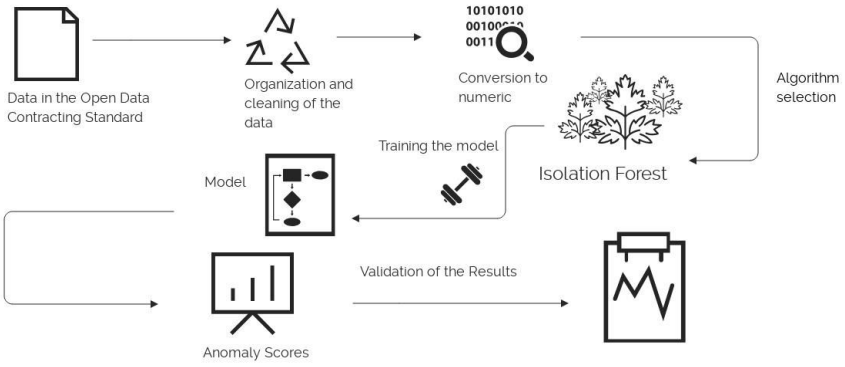

Fig. 1**.** Work flow of the proposed solution.

After training the model and getting the anomaly score for the input data, the DNCP provided to this work a dataset containing potentially anomalous procurement processes from 2010 onwards, in order to obtain a preliminary validation of the effectiveness of the classification model. The dataset includes potentially anomalous procurement processes with protests with judgments in favor of the protestant or with citizen complaints. Protests are internal disputes in the procurement process whereas complaints are external complaints with identity protection about the procurement

process. The scores of these cases obtained with the isolation forest implementation were then analysed to measure the accuracy of the trained models.

4 Preliminary Results and Final Discussion

The results consists of three trained models and the anomaly score for each of the data points. This anomaly score ranges between -1 and 1, where if it is less than 0 it is considered an anomaly and if it is more than 0 is considered as normal. The following Table 1 shows the total data points analysed per dataset, the total data points with protests with judgments in favor of the protestant and with complaints per dataset, the percentage of data points with protests and complaints detected as anomalous and the execution times. The computational platform used was an Intel Core i7, with 16 GB of RAM, running the iforest algorithm implementation of the Python scikit-learn library with a 1000 estimators and 50 samples. The implementation and input/output data can be found at https://gitlab.com/MaEliK/otherframeworks.

Table 1**.** Data points analysed and percentage of anomalous points detected

Data Set	Number of data Points	Protests in dataset	Complaints in dataset	Protests detected as anomalous	Complaints detected as anomalous	Execution Time
Planning and Tender	108470	599	297	48.75%	45.79%	9,119 s
Contracts	137783	1305	457	43.37%	42.45%	42,125 s
Contract Modifica-tions	29406	168	144	16.07%	31.94%	3,538 s

The obtained percentages show the proportion of data classified by the algorithm as anomalous data. It can be noticed that it detects almost half of the known potentially anomalous data, consistently in the planning and tender phases and the contracting phase. Finally, this work proposes an alternative to automate the regulatory control of procurement processes based on data analysis in OCDS format. An unsupervised learning based technique is proposed that could classify in seconds as anomalous more than 45% of the potentially anomalous dataset provided. Next steps are to validate the obtained results with the DNCP staff and check results according to local regulations. Also a better interpretation of the variables that influence high anomalous scores is required.

References

1. OCDS Homepage, http://standard.open-contracting.org/latest/en/, last accessed 2019/03/14.

2. Liu, Fei Tony, Ting, Kai Ming and Zhou, Zhi-Hua. “Isolation forest.” Data Mining, 2008. ICDM‘08. Eighth IEEE International Conference on.
3. Liu, Fei Tony, Ting, Kai Ming and Zhou, Zhi-Hua. “Isolation-based anomaly detection.” ACM Transactions on Knowledge Discovery from Data (TKDD) 6.1 (2012): 3.
4. Conti, Pier Luigi & Naldi, Maurizio. (2009). Detection of Anomalous Bids in Procurement Auctions. SSRN Electronic Journal. 10.2139/ssrn.1493346.
5. Issa, Hussein & Vasarhelyi, Miklos. (2011). Application of Anomaly Detection Techniques to Identify Fraudulent Refunds. SSRN Electronic Journal. 10.2139/ssrn.1910468.
6. Deng, Qingshan & Mei, Guoping. (2009). Combining Self-Organizing Map and K-Means Clustering for Detecting Fraudulent Financial Statements. 2009 IEEE International Conference on Granular Computing, GRC 2009. 126-131. 10.1109/GRC.2009.5255148.
7. Panigrahi, Suvasini & Kundu, Amlan & Sural, Shamik & Majumdar, Arun. (2009). Credit card fraud detection: A fusion approach using Dempster–Shafer theory and Bayesian learning. Information Fusion. 10. 354-363. 10.1016/j.inffus.2008.04.001.
8. Ley 2051/03 “De Contrataciones Públicas”, https://www.contrataciones.gov.py/documentos/download/marco-legal/12760, last accessed 2019/03/14.
9. Development Gateway, Open Contracting Partnership, “Red Flags for integrity: Giving the green light to open data solutions,” in press.