A HYBRID MODEL FOR FRAUD DETECTION ON PURCHASE ORDERS

William Ferreira Moreno Oliverio

Abstract: Frauds on the purchasing area is an issue which impacts companies all around the globe. One of the possibilities to tackle this issue is through the usage of audits, however, due to the massive volume of the data available today, it is becoming impossible to check all the transactions of a company, hence only a small sample of the data is verified. This work presents a new approach through the usage of techniques of signature detection with clustering techniques for an increased probability of inclusion of fraud related documents in the sample. Due to a non-existence of a public database related to the purchase area of companies for fraud detection, this work will use real procurement data to compare the probability of selecting a fraudulent document into a data sample via random sampling versus the proposed model as well as exploring what would be the best clustering algorithm for this specific problem. The proposed model improves the current state-of-the-art since it does not require pre-classified datasets to work, is capable to operate with a very high number of data records and does not need manual intervention. Preliminary results show that the probability of including a fraudulent document on the sample via the proposed model is approximately seven times higher than random sampling.

Key-word: Fraud detection; clustering, procurement, ERP

Introduction

Enterprise Resource Planning (ERP) are systems that provide complete automation for most business processes. While the automation increases the efficiency of the company, it opens possibilities for internal fraud if the controls available on the system are not robust enough to prevent it. Frauds represent, in average, 5% of the company revenue [1,2] and combined with the fact that the procurement departments manage more than 60% of company expenditure [10], this context represents a relevant research topic.

This work focus on detecting frauds on the procurement area, which has the highest financial impact on organizations. One of the possibilities to tackle this issue is through the execution of audits; however, due to the massive volume of the data available today, it is becoming impossible to examine all the transactions of a company. Hence only a small sample of the data is verified during an audit. Due to the small number of frauds compared to the typical transactions, frequently, these fraudulent transactions are not included in the sample and hence are not verified during the audit.

This paper presents a new approach using the techniques of signature detection associated with clustering for an increased probability of inclusion of fraud related documents in the sample.

Due to non-existence of a public database related to the purchase area of companies for fraud detection, this work will use real procurement data to compare the probability of selecting a fraudulent document into a data sample via random sampling versus the proposed model as well as exploring what would be the best clustering algorithm for this specific problem.

The proposed approach improves the current state-of-the-art since it does not require pre-classified datasets to work, is capable of operating with a very high number of data records and does not need manual intervention.

Background

In this section, we are going to cover the basic concepts of fraud in the procurement area, together with their main symptoms, the concept of signature matching, and the main concepts of clustering as well.

Several authors have already studied fraudulent behaviour, which can be classified into different types. In the works of [3, 4] the main types of frauds are described and grouped into 8 categories, but they can be summarized into bid rigging, double payment, kickback fraud, non-accomplice vendor, personal purchases, redirect payment fraud, shell company and pass through. Each of these frauds has different symptoms associated. One example on the redirect payment fraud would be the changes of banking details of a vendor on the corporate system before payment.

This work focus on the occupational fraud described as: "the use of one's occupation for personal enrichment through the deliberate misuse or misapplication of the employing organization's resources or assets" [1].

The concept of signature matching was already used on papers related to several different areas, including software engineering [11], biology [5], and network security [7]. On this paper, the signature matching will be used to identify symptoms related to fraud based on the analysis of the records on the Enterprise Resource Planning (ERP) system of a company. An example of a possible signature would be to identify which purchase orders (POs) had an invoice amount higher than the request made to the supplier.

The next concept that is used on the model is clustering, which according to [6] is defined as "an automatic learning technique which aims at grouping a set of objects into clusters so that objects in the same clusters should be similar as possible, where-as objects in one cluster should be as dissimilar as possible from objects in other clusters". Besides, clustering can be classified into supervised and non-supervised approaches [9].

To the specific problem of fraud detection in procurement, a non-supervised approach will be applied due to the lack of available datasets with examples of fraudulent and non-fraudulent documents which are required by the supervised clustering algorithms. Among the several available clustering algorithms, the following were used on the proposed model: K-Means, DBSCAN, BIRCH, HDBSCAN, CURE, CLIQUE, ROCK, and Spectral Clustering.

Related works

There are several papers related to fraud detection on the procurement area and nine papers were selected for a comparison covering the following points: scalability, the capability to process data in real time, the requirement of domain knowledge by the operator, the adaptability to new types of fraud, how automated the solution is and the availability of metrics to compare the performance of the solution.

The result of the comparison can be found in figure 1.

Figure 1: Overall analysis of the selected papers

Paper	Scalable	Real time capable	Domain knowledge required	Adaptable	Automated	Metrics available
The Effective Use of Benford's Law to Assist in Detecting Fraud in Accounting Data	Yes	Partial	Yes	Yes	No	No
A business process mining application for internal transaction fraud mitigation	No	No	Yes	Yes	No	Yes
Fraud detection in ERP systems using Scenario matching	Partial	Yes	No	No	Yes	Yes
Reducing false positives in fraud detection: Combining the red flag approach with process mining	Partial	Partial	Partial	No	No	Yes
Fuzzy C-Means for Fraud Detection in Large transaction data sets	Partial	Partial	Partial	Yes	Partial	No
Screening for bid rigging – does it work?	No	No	Yes	Yes	No	Yes
Prediction of Enterprise Purchases using Markov models in Procurement Analytics Applications	Yes	Partial	No	Yes	Yes	Yes
K-Means: Fraud Detection Based on Signaling Data	Yes	No	Yes	Yes	No	Yes

After reviewing the work performed on this area, we identified that there is not a clear trend in the state of the art and each of the papers selected presented positive points and limitations as well while exploring different approaches. One of the papers was close enough on all the areas being related to the prediction of procurement orders. The paper was selected for comparison since the model for prediction could be used for fraud detection and it was related to the procurement domain.

Proposed model

In this section, we describe de approach proposal for this work. After an overview with the main motives for the proposal, the components are detailed and commented.

The model has three main modules, where the following activities take place:

a) Data extraction and cleaning: The data is extracted from the ERP environment, filtered, cleaned, and standardized.

b) Signature matching: The data previously extracted is verified on a series of checks to identify symptoms which could be an indication of fraud.

c) Clustering: The data generated on the signature matching phase is used as input to a clustering algorithm.

Data extraction and cleaning

The proposed model starts with the data extraction from the ERP system which holds the corporate data related to the purchases performed by the organization.

The data that will be extracted and used on the model are the following:

a) Purchase requisition: The original request for purchase raised by the employee

b) Purchase orders: The purchase order after the approval and any adjustments performed during the approval phase

c) Goods receipt: The records of the reception of the physical goods in stock or confirmation of services provided for each purchase order.

d) Invoice receipt: The invoice received for the supplier

e) Payments: The payments performed to the supplier

Accounting documents:

f) Customer, supplier, plant, storage location and company codes master data: The details for each of the object (eg: creation date, banking account, address, etc)

g) System logs: The details for the changes performed on all the items above (eg: changes on supplier banking accounts, changes performed on the unit price after a purchase order is approved, etc).

Since the ERP system holds data for more than 180 different countries, the methodology chosen was to extract the purchase orders which were created over a period of 1 year for 9 different countries.

The countries were chosen based on the Corruption Perception Index [8], which classify all the countries across the globe according to the perceived corruption level. Three countries were selected from among the most corrupted countries, three from the cleanest countries and three from the countries with an average score. The selected countries can be found in figure 2.

Figure 2: Selected countries for the experiment

Position	Country	Corruption Prevention Index	Classification
3	Switzerland	85	Very Clean
11	Germany	81	Very Clean
13	Australia	77	Very Clean
60	Croatia	49	Average Results
61	Romania	48	Average Results
61	Malaysia	47	Average Results
149	Bangladesh	28	Highly Corrupt
149	Kenya	28	Highly Corrupt
168	Venezuela	18	Highly Corrupt

After the data is extracted from the ERP system, it was standardized and cleaned. Some of the activities performed on the phase where the conversion of the amounts paid to a single currency using the average exchange rate for the period taken into analysis. Another activity is removing the data related to intra-company transactions, which are the purchases performed across different companies of the same group. This occurs when a company in one country purchases raw materials from another country. Besides, were also removed the data from some vendors which have purchase orders created automatically and based on long term contracts. This occurs when raw materials used on factories have the purchase orders created automatically whenever the stock of raw material is low, where the amounts, prices and conditions are defined on long term contracts.

Signature matching

On the signature matching phase, we use SQL scripts to compare the data extracted previously in order to identify symptoms which could be related to fraud. A total of 17 signatures were created. Some examples are available below:

- **Goods not delivered for a paid invoice:** When there is no document related to the good/service receipt before an invoice is payed
- **Invoice amount higher than the order:** Compares the original purchase order, with the invoice received for the PO and identify any orders where the amount paid is higher than the original PO
- **Sequential invoices number from a supplier:** Verify how sequential the invoices from a specific supply are. In the case of a supplier which is only trading with one company, their invoices should have a sequential number.

The output of the signature matching is calculated and stored on a new dimension in the dataset, the format of the data is always numeric so it can be used as input on the clustering module. Finally, the output can be either binary returning 1 if the purchase order is relevant to the signature or 0 if it is not relevant or a floating number.

The signature for goods not delivered for a paid invoice, returns a binary value of wheatear or not the invoice was paid without a goods receipt while the signature for sequential invoices number from a supplier would return a score based on the formula below, where MaxINV is the highest invoice number received in the 1 year of data analysed, MinINV is the minimum invoice number received on the same period and NumINV is the number of invoices received.

$$Score = (MaxINV - MinINV) / NumINV$$

The last item generated on this module is the calculation of the average unit values for the quantity and prices of the purchased goods. This information is summarized at three levels, which are the supplier, the material and finally the supplier combined with the material. Based on the average values, new scores are created to identify when specific purchase orders have values significantly higher than the average quantity or price purchased in the period.

Clustering and evaluation

In the clustering model, the values of the signatures are used as input for the clustering algorithm. However, some pre-processing activities will be performed before the execution of the clustering algorithm. We choose to scale the data before the processing since the results of the signatures can have values with a very high variation. Some scores have a result of 1 (binary output) while others can have a substantially high number (eg: the difference of the price of a specific good price in one PO versus the historical average price)

The data were scaled using Scikit learn [12], through the usage of MinMaxScaling which scales and translates each feature individually, so the data is in the range of the data set. The result would be between zero and one.

Due to the very high number of dimensions and based on [13], we decided to reduce the dimensionality of the dataset from 25 dimensions to 6 and 7 dimensions trough application of Principal Component Analysis (PCA) from SciKit Learn Library in order to reduce the processing time and increase the cluster accuracy.

The following clustering algorithms were selected to be used on this experiment based on the review of clustering algorithms performed by [14]: K-Means, DBSCAN, BIRCH, HDBSCAN, CURE, OPTICS, Spectral Clustering, CLIQUE, ROCK.

For each clustering algorithm, we selected at least one input parameter (eg: parameter K on K-Means algorithm and EPS on DBSCAN) in order to identify the best clustering algorithm for this specific problem and the best value for the input parameter in order to achieve the best index performance. In order to select the best clustering algorithm and parameters, the executions of the clustering algorithms will be compared using Davies-Bouldin index, Calinski-Harabasz index and Silhouette index.

Finally, the results on the clustering algorithm will be analyzed based on the values of the signatures as well as performing a manual analysis of the purchase orders directly on the ERP system.

Results

For this experiment 147,898 POs and their related information (according to the details on the data extraction chapter) were extracted from the ERP system and copied to a separate database based on Microsoft SQL Server 2014. The data then was filtered by eliminating intra-company transactions (when a company buys a good or service from another company of the same group) and repetitive procurement, reducing the number of POs to 147,898 which became the scope of this exercise.

Next step performed was to normalize all the amounts available on the data to a single currency. After the data was filtered and normalized, it was processed using a SQL Script to generate the scores for the signatures. The score for each signature was saved on a column and then exported to a text file.

After that, we executed the PCA on the dataset available on SciKitLearn library. Results are available in figure 3, which shows that keeping 6 dimensions would maintain 97.9% of original data variance and 7 dimensions would maintain 98.6%.

Based on the results from PCA, the data was executed on the clustering algorithms in three different formats: a) No PCA executed; b) PCA executed and keeping the top 6 principal components; c) PCA executed and keeping the top 7 principal components.

Figure 4 shows the run-time of each clustering algorithm can be found as well as the number of executions of each clustering algorithm.

Figure 3: Data variance by principal component

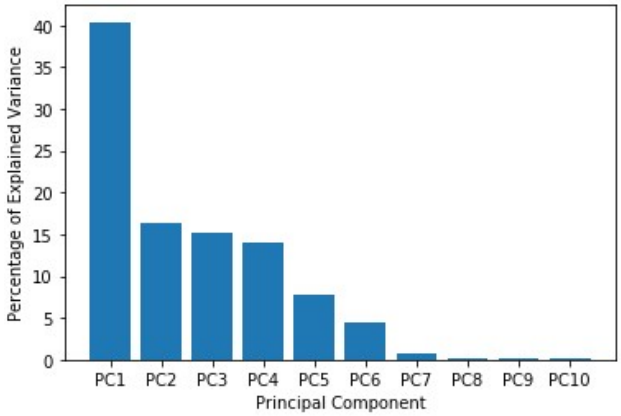

Figure 4: Performance of clustering algorithms

Average run-time in seconds						
Algortihm	Source data		PCA - 7 Dimensions		PCA - 6 Dimensions	
	Number of executions	Average run time	Number of executions	Average run time	Number of executions	Average run time
K-Means	100	245.70	100	173.41	100	236.58
DBSCAN	100	725.14	100	446.92	100	526.66
BIRCH	6	10.83	6	6.56	6	5.89
HDBSCAN	241	210.18	241	13.29	241	7.56
CURE	-	Lack of memory	-	Timeout - 12 hours	-	Timeout - 12 hours
OPTICS	-	Lack of memory	-	Timeout - 12 hours	-	Timeout - 12 hours
Spectral Clustering	-	Lack of memory	-	Lack of memory	-	Lack of memory
CLIQUE	-	Lack of memory	-	Lack of memory	-	Lack of memory
ROCK	-	Lack of memory	-	Lack of memory	-	Lack of memory

BIRCH clustering algorithm was executed only 6 times since the clustering algorithm took more than 12 hours to finish the processing of the data when the parameter Threshold was raised above 0.11. The algorithms identified with "Lack of memory" failed to start the data processing right at the start due to lack of system memory. Algorithms marked with "Timeout – 12 hours" started the data processing but didn't finish the processing after 12 hours of execution and were cancelled. Applying PCA on the dataset reduced the processing time of all the algorithms, with HDBSCAN being the clustering algorithm which had the best positive impact on applying PCA to the dataset.

The next step performed was to choose among all the executions which clustering algorithm and which parameter generated the best clusters based on internal cluster validation. Figure 5 shows the best result for each of the three indexes by clustering algorithm and utilization of PCA on the data.

Figure 5: Best index values by clustering algorithm and usage of PCA

PCA Performed / Dimensions	Algorithm	Max of CH Index	Min of DB Index	Max of Silhouette Index
YES-6	BIRCH	7,184.14134	1.18336	0.31691
YES-6	DBSCAN	31,299.97551	0.78809	0.49017
YES-6	HDBSCAN	12,179.67985	0.75381	0.49686
YES-6	K-Means	99,112.15729	0.84192	0.48378
YES-7	BIRCH	9,036.43290	1.09877	0.32323
YES-7	DBSCAN	31,299.97551	0.79820	0.49207
YES-7	HDBSCAN	56,817.80226	0.63224	0.53223
YES-7	K-Means	99,112.15729	0.84155	0.48378
No PCA applied	BIRCH	15,582.07629	0.68663	0.34748
No PCA applied	DBSCAN	15,621.01747	0.68818	0.48885
No PCA applied	HDBSCAN	58,000.35616	0.53937	0.55652
No PCA applied	K-Means	99,112.15729	0.84163	0.48378

K-Means algorithm had the best score for Calisnki Harabasz index but had worst results on both DB index and Silhouette index. Based on figure 5, the clustering algorithm chosen was HDBSCAN without application of PCA which have the best score for Silhouette index and DB index. The parameter used on HDBSCAN clustering algorithm which provided the best result was the minimum cluster size = 950 documents, which generated 8 clusters. The output of the clustering algorithm was then analyzed, and the following results were reached.

Cluster 0: This cluster has most documents which were retroactive (when the PO is created after the date on the invoice) as well as missing contact details (address or phone number) on the supplier master data in addition to a significantly high number of PO which the price on the requisition is higher than the one received on the invoice. This cluster has 9,270 documents and is considered as high risk due to the possibility of bid rigging.

Cluster 1: All the POs on this cluster have at least another PO created by the same person, on the same company, country, date and amount, which is an indication of a PO which was split into multiple smaller ones. Another important characteristic is that this cluster has the highest number of suppliers which has bank accounts which are the same as other suppliers. This cluster has 2,934 documents and is considered as a medium risk due to the case of suppliers sharing banking accounts with other entity.

Cluster 2: This cluster has the highest deviation of prices between the invoice and the purchase requisition in addition to all the cases of a sleeping vendor (when a supplier starts to trade again after a long time of inactivity). Additionally, all the documents on this cluster were not approved through the regular workflow and all the suppliers do not have any phone details available. Finally, this cluster has the highest ratings of suppliers blocked before a PO is created as well as changing the payment terms of the PO right before it is created. The cluster has 2,275 POs it is considered a high-risk cluster due to possible frauds with sleeping vendors.

Cluster 3: Similar to cluster 1, all the POs on this cluster have at least another one created with the same company, country, date and amount, which is an indication that a PO was split into smaller ones. All the POs on this cluster were not approved through the regular workflow for the approval process. This cluster contains 1,037 POs and can be considered as low risk since there is an indication of split PO which could a deviation of company policy but there is no evidence of possible fraud.

Cluster 4: Similar to cluster 2, this cluster has the highest deviation of prices between the invoice and the purchase requisition together with the fact that most of the suppliers included on this cluster create sequential invoice numbers, which is strong evidence that the supplier only trades with a single company. Another characteristic is that POs part of this cluster had the highest increase of the unit price of the good/service being purchased coupled with the fact the prices for the goods acquired are significantly higher than the average price for the same good and supplier. This cluster contains 7,267 POs and is considered a high-risk cluster due to significantly higher prices coupled with suppliers which have an indication that trades only with the company which is the object of this study and hence could be an indication of pass-through frauds, shell company fraud or bid rigging.

Cluster 5: This cluster has some indications of signatures, but the average values are significantly lower than the values found on the clusters analyzed. This cluster

contains 35,940 POs and is considered low-risk due to no significant evidence has been found.

Cluster 6: Like cluster 5, there is no indication of any signatures with a significative value, with the only exception being the fact that all the POs were not approved through the standard approval workflow of the company. This cluster has 73,538 POs and is classified as low-risk due to the absence of any further pieces of evidence.

Cluster 7: This cluster is very similar to cluster 6, without any significant values of any signatures apart from the fact that the POs were not approved through the standard approval workflow. This cluster 7,934 POs and is classified as low-risk.

Outliers: Apart from the 8 clusters created, the HDBSCAN algorithm classified 7,703 POs as outliers during the clustering execution. The POs classified as outliers were analyzed as a regular cluster and no signatures showed a significant value, which categorizes the POs classified as outliers as low-risk as well.

The average results for the signatures used in this study can be found in figure 6.

Figure 6: Average of the signature values per cluster generated

	Clusters								
Signatures	**-1**	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Invoice amount is higher than order	5.57385	9.59139	2.49477	11.90282	3.21719	11.15531	5.08216	3.24015	7.54994
Sequential invoice numbers	5.43798	0.96783	0.44948	0.66005	0.65712	0.01752	8.77515	8.98506	0.04961
Sleeping vendor	0.10954	0.00000	0.00000	0.00134	0.00000	0.00000	0.00000	0.00000	0.00000
Po performed outside standard workflow	0.53989	0.00000	0.00000	1.00000	1.00000	0.00000	0.00000	1.00000	1.00000
Retrospective po (goods receipt)	0.34313	1.75245	0.02395	0.10995	0.00000	0.18319	0.10710	0.42905	0.02652
Retrospective po (invoice receipt)	1.72309	6.42033	0.56446	2.22764	3.78004	1.09807	1.89774	0.53515	2.15882
Vendors without phone number	0.07252	1.00000	0.00000	1.00000	0.00000	0.00000	0.00000	0.00000	0.00000
Vendors without address	0.00000	0.00000	0.00000	0.00000	0.00555	0.00001	0.00000	0.00000	0.00382
Difference between supplier creation and first sale	940.26641	662.74771	1,140.91986	909.64929	1,399.19039	1,138.55339	924.42452	1,301.93214	1,314.11256
Vendor with the same bank account as another vendor	0.99752	1.19161	1.41551	0.96200	1.10536	1.23491	1.00296	0.97425	1.13300
Average difference between sales dates	82.51298	36.89338	19.82622	26.96369	22.40758	28.63565	30.73886	13.93278	26.22971
Purchase order unit increased after po creation	0.38804	0.83208	0.23732	0.07818	0.06771	1.90029	0.77561	0.03214	0.13608
Price above average for supplier and material	0.00640	0.01761	0.01459	0.05294	0.00094	0.32540	0.00012	0.14097	0.04624
Quantity above average for supplier and material	0.00061	0.02968	0.01356	0.00562	0.00000	0.05105	0.00680	0.01826	0.00486
Price above average for supplier	2.86795	1.01479	2.46260	4.24204	2.69779	2.29521	8.84686	3.55636	1.61569
Quantity above average for supplier	2.12643	1.11151	0.99598	0.24006	0.01394	0.92948	1.14071	0.11307	1.76379
Price above average for material	3.85852	0.01761	0.01457	8.17647	0.00094	0.32568	0.00402	24.72948	0.90559
Quantity above average for material	0.02962	0.06553	0.01356	0.06817	0.00000	0.05151	0.00273	0.12180	0.00927
Multiple po for same creator, supplier, amount and date	0.09427	0.00000	1.00000	0.00000	1.00000	0.00000	0.00000	0.00000	0.00000
Bank data changed for supplier	0.00038	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
Supplier blocked before a purchase	0.03359	0.00484	0.00436	0.39677	0.00924	0.01732	0.00638	0.00966	0.04134
Changes performed on payment terms before a purchase	0.06794	0.13201	0.04355	0.65467	0.15157	0.21918	0.09682	0.01391	0.16242
Purchase order blocked	0.00382	0.00129	0.00000	0.00168	0.00000	0.00229	0.00177	0.00000	0.00486
Number of POs per cluster	**5240**	**9295**	**2296**	**2974**	**1082**	**74140**	**8459**	**7766**	**36646**

Conclusion

The utilization of signature matching coupled with clustering algorithms reduced the scope of data considered to be analyzed to just 13% of the original scope, from 147,898 documents to 19,471 documents by removing from the scope the low-risk documents.

Additionally, since the documents are classified into clusters, the sample can be done by selecting documents from each cluster, ensuring that all possible types of fraud would be analyzed.

The best clustering algorithm among all the ones compared for this specific problem was HDBSCAN which generated the best results according to both Silhouette index and Davies-Bouldin index, however, K-Means had the best results according to Calinski Harabasz index.

Finally, it was identified that reducing the dimensionality of the data trough PCA improved the performance of the clustering algorithm but did not increase the results of the indexes for the clusters generated.

References

1. ASSOCIATION OF CERTIFIED FRAUD EXAMINERS (ACFE). **Report to the nations.** Austin, USA, 2016. Available on: https://www.acfe.com/rttn2016/docs/2016-report-to-the-nations.pdf Access on: March 3rd 2019
2. ASSOCIATION OF CERTIFIED FRAUD EXAMINERS (ACFE). **Report to the nations.** Austin, USA, 2018. Available on: https://s3-us-west-2.amazonaws.com/acfepublic/2018-report-to-the-nations.pdf Access on: March 3rd 2019
3. HUBER, M.; IMHOF, D. Machine Learning with screens for detecting Bid-Rigging Cartels. **Working Papers SES**. [s. l.], n. 494, p. 1-28, 2018
4. IMHOF, D.; KARAGÖK, Y.; RUTZ, S. Screening for bid rigging-does it work? **Journal of Competition Law and Economics**. [s. l.], V. 14, n. 2, p. 235–261, 2018.
5. JONASSON, J., OLOIFSSON, M., MONSTEIN, H. J. Classification, identification and subtyping of bacteria based on pyrosequencing and signature matching of 16S rdna fragments. **Apmis.** [S.1], V. 110, n. 3, p. 263-272, 2002
6. POPAT, S.; EMMANUEL, M. Review and Comparative Study of Clustering Techniques. **International Journal of Computer Science and Information Technologies.** [S.1], V. 5, n. 1, p. 805-812, 2014
7. SMITH, R. et al. Evaluating GPUs for network packet signature matching. In: 2009 IEEE International Symposium on Performance Analysis of Systems and Software, Boston, MA, USA. **Proceedings…** IEEE, 2009, p. 175-184
8. TRANSPARENCY INTERNATIONAL. **CORRUPTION PERCEPTION INDEX.** 2017, Available on: https://www.transparency.org/news/feature/corruption_perceptions_index_2017#table Access on: March 3rd 2019
9. XU, R.; WUNSCH II, D. Survey of Clustering Algorithms **IEEE Transactions on Neural Networks.** [S.1], V. 16, n. 3, p. 645-678, 2005
10. WESTERSKI, A. et al. Prediction of enterprise purchases using Markov models in procurement analytics applications. **Procedia Computer Science.** [S.1], V. 60, p. 1357-1366, 2015
11. ZAREMSKI, A., WING, J. Signature Matching: A key to reuse. In: 1st ACM SIGSOFT symposium on Foundations of software engineering, Los Angeles, California, USA. **Proceedings…** ACM, 1993, p. 182-190
12. PEDREGOSA et al. Scikit-learn: Machine Learning in Python, **JMLR** 12, pp. 2825-2830, 2011.
13. DING, C., HE, X. K -means clustering via principal component analysis, 2004 29. https://doi.org/10.1145/1015330.1015408
14. XU, D. and TIAN, Y. A Comprehensive Survey of Clustering Algorithms. **Annals of Data Science.** [S.1], V. 2, n. 2, p. 165-193, 2015