Triclustering based outlier-shape score for time series in a fraud detection platform

Pierre Lejeail[1,2], Vincent Lemaire[1], Antoine Cornuéjols[2], Adam Ouorou[1]

[1] Orange Labs, 2 Avenue Pierre Marzin, 22300 Lannion, France
[2] AgroParisTech, 16 rue Claude Bernard, 75005 Paris, France

Abstract. This paper presents a triclustering based outlier-shape score for time series in the context of a fraud detection platform for wholesale traffic for a telecommunications carrier. We propose to use triclustering as an exploration module for outlier shape detection using whole time series. Three main steps compose this approach: (1) projection of data in a new space of time series related features (e.g. derivative), (2) estimation of the density of known normal data using a triclustering method (3) computation of an outlierness score quantifying the distance to the estimator from step (2). We conduct an evaluation of the methodology by focusing on its ability to separate data from different classes. Our preliminary results to assess this approach are very encouraging.

1 Introduction - Context

Telecommunications companies in different countries use a variety of international telecoms routes to send traffic to each other. In a "wholesale market", telecom carriers can obtain traffic to make up a shortfall, or send traffic on other routes, by trading with other carriers in the wholesale or carrier-to-carrier market. Minutes exchanges allow carriers to buy and sell call terminations (dispatch of a call to its destination). Prices in the wholesale market can change on a daily or weekly basis. A carrier will look for least cost routing function to optimize its trading on the wholesale market.

The quality of routes on the wholesale market can also vary, as the traffic may be going on an illegal route.

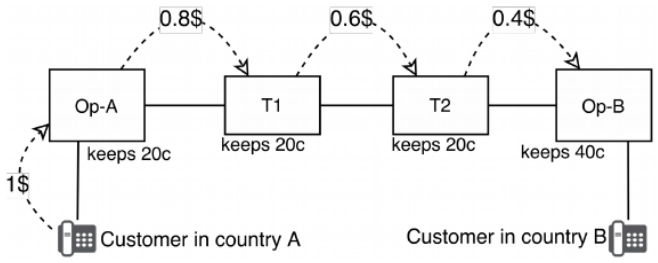

Fig. 1. A value chain providing a call termination between two users

A value chain (as the example in Fig. 1) exists between operators that provide the connection between two customers. But this value chain could be broken if a fraudster finds a way to generate communication without paying. For the last decades, fraud has been a growing concern in the telecommunication industry. In a 2017 survey [1], the CFCA[3] estimates the annual global fraud loss of $30 billion (USD). Therefore, detecting and preventing, when possible, is essential in this domain. Regarding wholesale market, a list of fraud is known [2] by the operators and fraud detection platforms already exist.

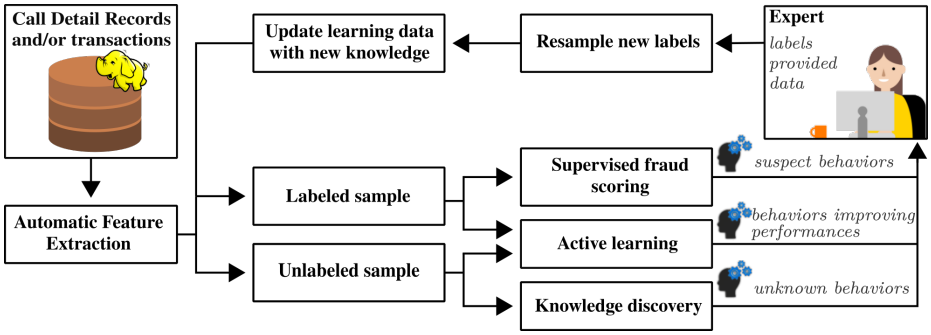

Fig. 2. The Orange platform for fraud detection

One of these platforms, shown in in Fig. 2, has been realized by Orange (as a wholesale operator). This platform contains modules which **exploit** the information given by the expert (a scoring module, for example, which is a classifier), others **explore** the data to interact with the expert(s) by finding new patterns, including (but not only) malevolent ones. This goal is achieved by knowledge discovery and active learning techniques. Those exploration modules are responsible for adapting to the constant evolution of fraudster's behaviors under (in the case of this platform) the constraint of the limited time that experts can afford to spend in this exploration. This platform share similarities with the one presented by Veeramachaneni et al. in [3]. Both platforms combine a supervised model for predictions with unsupervised models for exploration of unknown pattern, and takes into account the user feedbacks in the learning phase.

In this paper, we are interested in introducing a preliminary work on the knowledge discovery module and we propose a new methodology for the exploration of unlabeled data in an unbalanced data sets. To do this we suggest a new approach to find anomalous behaviors where individuals are represented as time series with a fixed length. By anomalous[4] behavior, we mean a behavior different from the behaviors **known** as **normal**.

[3] Communications Fraud Control Association

[4] In this paper, we also use the term 'outlier' as a synonym.

2 Outlier shape detection

2.1 Time series in wholesale fraud detection

Wholesale call record details (call duration, timestamps of the call, destination,...) are grouped by calling phone numbers. Thus, each phone number is represented by multivariate time series. Depending on the number of calls made, time series can have various numbers of points. Observations are kept on a fixed duration, W, (e.g. 6 hours), meaning that all time series have the same "length" (W). The fraud detection platform must be used in a batch protocol and on past telecommunications (past calls) since the billing is not performed online[5] between the operators concerned by a given call. The aim is to find abnormal behavior time series using comparison to a time series database used before to train a model. To inform the other operators that the calls related to those time series are potential frauds, the expert is responsible to assign or not a fraudulent character label to those time series. As we are interested by the global shape of the time series, our method tries to find the abnormality using the entire time series (on the whole length W) rather than searching for outlier points/subsequences in the time series. Therefore, we propose an outlier shape detection algorithm for the whole time series rather than an outlier points detection algorithm.

2.2 Related works

Outlier detection has been extremely studied in the last decades, see for instance the Aggarwal's book [4]. More specifically, for temporal data, there are surveys like the one of Gupta and al. in [5]. Figure 3, from [5], shows a perspective on the state of the art with respect to the data type aspect.

The methodology described in this paper regards time series and thus the red square of Figure 3. As explained in Section 2.1, the wholesale setting implies to work with outlier shape detection algorithms for the whole time series. The related state of the art is therefore in the blue square of the figure: "Direct Detection Time Series Outliers" and the "Unsupervised Discriminative Approaches"[6]. The latter uses three key elements: a time series on a give representation, a clustering method and a similarity measure.

[5] Fraud disputes between operators are resolved at billing time[2].

[6] To save space, we do not give details on this part of the state of the art (see [5] for more details), but in a potential extended version paper our method will be compared with competitive approaches. The reader may also note that our method uses density estimation and therefore the state of the art on density estimation could be of interest as well.

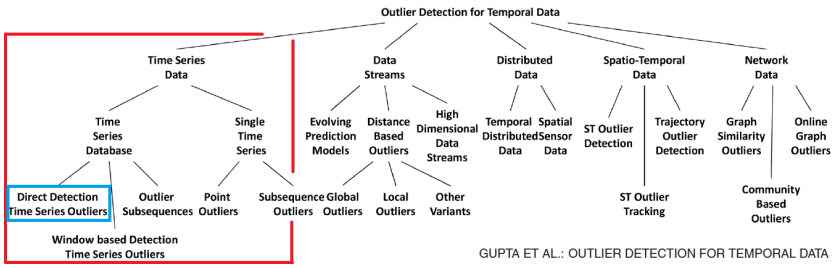

Fig. 3. A perspective on the state of the art for temporal data

3 Suggested approach of outlierness score

Our intent is to investigate a new suggestion concerning two of the three key elements of the "Direct Detection Time Series Outliers" family methods: the representation and the clustering method. We propose in a preliminary work a methodology based on the estimation of the density of normal events using a triclustering method, on a panel of time series representations and then to compare distances between this density and new individuals. The farther an individual is from the 'normal behavior' distribution, the more likely it is anomalous. The main steps of this methodology are:

1. Transform each time series in several representations;
2. Estimate the density of this set of individuals using a triclustering method;
3. Choose the best representation of time series;
4. Compute an indicator of distance to find unknown behaviors to this density.

3.1 Time series representations

In the context of classification, the study of Gay and al. [6] indicates that the choice of representation could have an important impact on the results. In our case it is impossible to know in advance which representation will be the best for density estimation vs. outliers detection. Therefore, we conduct, in Section 4, experiments using a "pool" of representations. The representation used in this first approach are the same as Gay and al. in [6]: *the original representation, the Fourier transform coefficients, the first derivative, the second derivative, the cumulated integral, the double cumulated integral and the autocorrelation function*. Each of those representations captures different information of time series. Other representations will be used in future works.

3.2 Estimation of time series density

We suggest to use a triclustering approach on the following dimensions: customers (C), timestamps (T) and a given variable (X) (for example the call duration)[7].

[7] The multivariate aspect is not studied in this preliminary work.

The input data set D is considered to be a collection of N time series, denoted S_i with $i \in 1, \ldots, N$. The key idea of triclustering with time series is that a curve can be seen as a set of points is that a curve i can be represented as a set of triplets, i.e. $S_i = \{(C_i, T_j, X_{ij}) \forall j \in 1, \ldots, m_i\}$. The total number of data points is denoted by $m = \sum_{i=1}^{N} m_i$. No assumptions are made on the time series; they could have different number of points and different timestamp values. A triclustering model aims at jointly discretizing the numerical variables T, X and at making a partition of time series by grouping the values of the categorical variable C. The objective of such models is to estimate $P(X, T, C)$, the joint probability of C, T and X. The output of the model gives groups of time series which are characterized by a bivariate discretization which estimates $P(X, T|C)$.

Amongst the state of the art of the triclustering methods we decided to use the MODL one [7]. MODL estimates the constant piecewise joint density of C, T and X with a multidimensional data grid. The optimal grid M is found by a bottom-up greedy optimization of a Bayesian criterion which makes a trade-off between accuracy and robustness of the model:

$$cost\,(M) = -log\,(p\,(M)\,p\,(D|M))$$

From the information theory point of view, this criterion can be interpreted as the encoding length of the grid plus the length of the dataset D, knowing M, according to the Minimum Description Length principle developed by Rissanen in [8]. This method is advantageous for unsupervised learning since it is free of user parameters.

3.3 Indication about the representation used

The MODL method gives a criterion named "level" (a compression rate value between 0, for a null compression gain, and 1, for a perfect compression gain) which indicates the quality of the density estimation made. One triclustering will be made per representation. Ideally, the triclustering maximizing the quality of the density estimator (value of the level) will be the best, i.e. it would also most likely provide the best separation between known normal time series and outlier time series.

3.4 Outlier score

The found partitioning gives the relationship between clusters of behaviors and groups of variables and time intervals. Thus, the outlier behavior of a customer can be linked to a specific interval of time and a set of variables. For a given new time series (not present when training the model) the purpose will be to compute how far this time series appears with respect to the model trained.

To do this we use the similarity criterion which has been defined to assess the proximity between a new time series and the clusters of a MODL coclustering model in [9]. This criterion evaluates the decrease of the level (Section 3.3), when a time series is merged within a particular cluster. This decrease can be linked to

the impact on the intra-cluster inertia of adding a new individual to its nearest cluster. Therefore, the less typical an individual from its nearest cluster, the higher the score.

4 Experimental setting and results

This section presents the data sets, the validation protocol and the results of the proposed methodology. The implementation of our method is based on the tool Khiops [10] and on R scripts for the automation of the methodology.

Our fundamental assumption is that a time series assigned to an unknown class is more likely to be anomalous than if it has been recognized as a member of a known 'normal class'. In the following experiments, we test this hypothesis by using supervised data sets where we emulate the normal/abnormal aspect of the behaviors by choosing a class as 'normal behaviors' and another as 'abnormal behaviors'. Our goal is to answer three questions :

1. Is a triclustering useful to find outlier-shape?
2. Can we find a good time series representation to detect outliers and help the triclustering?
3. Is the quality of density estimation of the triclustering correlated with the obtained results?

4.1 Protocol

As often in cyber criminality contexts, it is difficult to share data and results for confidential reason. Therefore, we experimented our method on 9 data sets from the UCR [11], offering a wide variety in terms of number and length of time series.

Each chosen dataset is composed of two classes of time series. As our model learns on a training dataset only composed of expected normal data, the train-test setting is not the one predefined in [11]. We use the following protocol to build our train-test setting:

1. Choose the majority class as normal data;
2. Choose the minority class as anomalous data;
3. Draw 70% of normal data as train data;
4. Keep remaining data as test data.

The evaluation of the score provided by the method for each representation is made with the AUC on test data. This indicator measures how well the score sort the individuals of the different classes, i.e. a good AUC that individuals from the minority class tend to have the highest scores.

4.2 Results and discussion

Performances in term of AUC (rounded to the second decimal) are reported in Table 1 for each representation (TS: original time series, DV: first derivated, ddv: second derivated, IV: cumulated integral, IIV: double cumulated integral, PWS: Fourier transform, ACF: auto-correlations). The best result for each dataset is written in bold. We detail below the results in view to the three questions questions set in Section 4.

Dataset	Train	Test	L	ROC AUC							
				TS	DV	DDV	IV	IIV	PWS	ACF	Best
Wafer	4481	2134	152	0.98	1.00	**1.00**	0.99	0.99	0.98	0.99	1.00
ECG5000	2043	1201	140	**0.99**	0.98	0.78	0.93	0.91	0.90	0.99	0.99
Computers	175	325	720	0.32	0.51	0.44	0.60	0.54	0.50	**0.64**	0.64
MoteStrain	479	793	84	**0.96**	0.95	0.84	0.85	0.66	0.61	0.82	0.96
PhalangesOutlinesCorrect	1188	567	80	0.67	0.64	0.65	0.62	0.55	**0.71**	0.63	0.71
ItalyPowerDemand	384	712	24	0.90	0.68	0.76	0.60	0.43	0.64	**0.92**	0.92
SonyAIBORobotSurface	384	377	70	0.54	0.64	0.61	**0.87**	0.79	0.58	0.66	0.87
Phalanx	235	641	80	0.40	0.30	0.31	0.41	**0.55**	0.52	0.44	0.55
WormsTwoClass	105	153	900	0.77	0.67	0.56	**0.78**	0.72	0.35	0.72	0.78

Table 1. Description of time series data sets and comparison of different reprentations. (L=Length). The column "Best gives" the best results per line for the AUC.

Is a triclustering useful to find outlier-shape? The answer seems positive: At least one AUC is better than random for each dataset (except phalanx with an AUC for double cumulated integral of 0.55).

Can we find a good time series representation to detect outliers and help the triclustering? The answer is clearly yes. The last column of the Table 1 indicates that we may find in the (small) pool of representation a representation which performs well. But no representation emerges as better that **all** others as the results of Gay and al. in [6].

Is the quality of density estimation of the triclustering correlated with the obtained results? For reasons of space limitation, we cannot give the Table of the **level** values (see Section 3.2) for all representations and data sets, but contrary to intuition, we do not find a positive correlation between the level value and the AUC. Thus the answer is negative. It appears that the method produces clusters of time series that are too specific in some cases. For example, the original representation of computers (training of 175 time series) is partitioned in 174 clusters. Despite the fact that this behavior does not necessarily imply bad results (e.g. the same problem is observed in WormsTwoClass), it is not desirable and may induce a greater variability and worst results than what could be expected.

Discussion : The experiments provide results that are consistent with the assumption that one should observe a larger score in general for classes which were not used in training. Those results are encouraging enough to pursue our efforts in this direction. In fact, by choosing the right threshold (which is a challenge by itself) for predictions one could obtain results nearly as good as in

classification (e.g. comparing our results for Wafer, SonyAIBORobotSurface and ItalyPowerDemand with Gay and al.[6], our AUC implies an accuracy, with a threshold well-chosen, nearly as good as with supervised models). Several challenges were revealed by our results. First, there is, for the moment, no way to find a good representation of the data without using a test data set which will not be available in many production applications. This is an important issue since different representations of the same dataset can lead to significant differences between AUC (e.g. ItalyPowerDemand AUC varies from 0.92 to 0.43).

5 Conclusion

This paper has presented a fraud detection platform for telecom wholesale traffic and introduced a preliminary work on an exploration module. The methodology used is based on outlier detection techniques to detect anomalous unlabeled results. The main key features of this module are: (1) use of a non-parametric method, (2) based on the ability to learn an estimator of the data density and (3) consuming a minimal amount of the expert time. The results are encouraging. They are close to those of supervised models provided by Gay et al. Works still needs to be done, notably to find a way to select/build a good representation for anomaly detection without knowing labels.

References

1. CFCA: 2017 Global Fraud Loss Survey. Survey Results, Communications Fraud Control Association (2018)
2. I3 Forum: I3f Fraud Classification. White paper 3, I3Forum (May 2014)
3. Veeramachaneni, K., Arnaldo, I., Bassias, C., Li, K., Cuesta-Infante, A.: AIˆ2: Training a Big Data Machine to Defend. In: AIˆ2: Training a Big Data Machine to Defend. (April 2016) 49–54
4. Aggarwal, C.C.: Outlier Detection. Springer (2015)
5. Gupta, M., Gao, J., Aggarwal, C.C., Han, J.: Outlier detection for temporal data: A survey. IEEE Trans. Knowl. Data Eng. **26**(9) (2014) 2250–2267
6. Gay, D., Guigourès, R., Boullé, M., Clérot, F.: Feature Extraction over Multiple Representations for Time Series Classification. In: New Frontiers in Mining Complex Patterns, Cham, Springer International Publishing (2014) 18–34
7. Guigourès, R.: Utilisation des modèles de co-clustering pour l'analyse exploratoire des données. PhD thesis, Université Paris I, Panthéon-Sorbonne (2013)
8. Rissanen, J.: Modeling by shortest data description. Automatica **14**(5) (1978) 465 – 471
9. Guigourès, R., Boullé, M., Rossi, F.: Discovering patterns in time-varying graphs: a triclustering approach. Advances in Data Analysis and Classification (Oct 2015)
10. Boullé, M.: Khiops: outil d'apprentissage supervisé automatique pour la fouille de grandes bases de données multi-tables, available at `www.khiops.com`. In: 16éme Journées Francophones Extraction et Gestion des Connaissances, EGC. (2016) 505–510
11. Chen, Y., Keogh, E., Hu, B., Begum, N., Bagnall, A., Mueen, A., Batista, G.: The ucr time series classification archive (July 2015) `www.cs.ucr.edu/~eamonn/time_series_data/`.