Combining Unsupervised and Supervised Learning in Credit Card Fraud Detection

Fabrizio Carcillo[a], Yann-Aël Le Borgne[a], Olivier Caelen[b], Yacine Kessaci[b], Frédéric Oblé[b], Gianluca Bontempi[a]

[a] *Machine Learning Group, Computer Science Department, Faculty of Sciences ULB, Université Libre de Bruxelles, Brussels, Belgium.*
(email: {fcarcill, yleborgn, gbonte}@ulb.ac.be)
[b] *R&D Worldline, Worldline, France.*
(email: { olivier.caelen,yacine.kessaci,frederic.oble} @worldline.com).

Abstract

The usage of supervised learning is pervasive in credit card fraud detection. The main assumption underlying the supervised learning techniques is that fraudulent patterns can be learned from the past. The task become challenging when taking in account customer changes in behavior and the ability of the fraudsters to invent novel fraud patterns. In this context, unsupervised learning techniques can help the Fraud Detection System to find anomalies. In this paper we present a hybrid technique which combines supervised and unsupervised techniques to address the fraud detection problem. Unsupervised outlier scores computed at different levels of granularity are compared and tested on a real annotated credit card fraud detection dataset. Experimental results show that the combination is efficient and indeed improves the accuracy of the detection.

Keywords: Fraud Detection, Ensemble Learning, Outlier Detection, Semi-supervised Learning, Contextual Outlier Detection

1. Introduction

Credit card fraud detection aims to detect whether a credit card transaction is fraudulent or not on the basis of historical data. It is a notoriously difficult problem, due to the nature of customer spending behaviors (e.g. changing habits during holiday periods) and fraudsters techniques (e.g.

adaptation to fraud detection techniques). It is now known that an effective approach to tackle this problem relies on machine learning techniques [**?**].

A typical Fraud Detection System (FDS) includes multiple layers of control which can be automated or supervised by humans [**?** **?**]. Part of the automated layer embraces machine learning algorithms which build predictive models based on annotated transactions. A large body of machine learning research for credit card fraud detection has grown in the past decade, which can be divided into supervised, unsupervised and semi-supervised techniques [**?** **?**]. In our previous works on credit card fraud detection, we investigated supervised [**?** **?**], unsupervised [**?**] and semi-supervised techniques [**?** **?**].

Supervised techniques rely on the set of past transactions for which the label (also referred to as outcome or class) of the transaction is known. In credit card fraud detection problems, the label is either genuine or fraudulent, i.e., the transaction was made by the cardholder, or by a fraudster. The label is usually known *a posteriori*, either because a customer complained or as a result of an investigation of the credit card company. Supervised techniques make use of labeled past transactions to learn a fraud prediction model, which returns, for any new transaction, its probability to be a fraud. But, not all the labels are available immediately [**?** **?**].

Unsupervised outlier detection techniques do not require knowledge of the label of transactions, and aim at characterizing the data distribution of transactions. They rely on the assumption that outliers of the transaction distribution are frauds and can be used to detect *unseen* types of frauds since they do not rely on past labeled transactions. It is worth noting that their use also extends to clustering and compression algorithms [**?**], which allow to identify separate data distributions for which different predictive models should be used (clustering), or by reducing the dimensionality of the learning problem (compression), which usually improves the performances of supervised techniques.

The two approaches are complementary: supervised techniques learn from past fraudulent behaviours, while unsupervised techniques allow to detect new types of fraud, or to cluster and compress data in order to improve supervised techniques. While complementary, the two approaches are combined in the semi-supervised techniques [**?** **?**]. Those are often used in situation where there are many unlabeled data points and few labeled ones. They aim at performing better than a supervised model which uses only the dataset of the few available labeled data points, or an unsupervised model

which does not profit from the few labels.

This paper concerns the integration of unsupervised techniques with supervised credit card fraud detection classifiers. In particular we present a number of criteria to compute outlier scores at different levels of granularity (from *high granular* card specific to *low granular* global outlier scores) and we assess their added value in terms of accuracy once integrated as features in a supervised learning strategy. The combination of unsupervised and supervised learning is not new in litterature as discussed in Section 2. In particular the approach, we refer to as global, is inspired to the *best-of-both-worlds* principle proposed by Michenkova et al. in [**?**]. What is original in this paper is the adoption of this principle in a credit card fraud detection setting and specifically the design and assessment of several outlier scores adapted to the specific nature of our problem. Section 3 introduces the standard unsupervised outlier scores used in the experimental section (Section 3.1), three original approaches to consider different levels of granularity when computing the outlier scores (Section 3.2) and the metrics used to compare the different approaches (Section 3.3). The experimental comparison is performed in Section 4, while the discussion and conclusion are presented in Section **??** and Section **??**.

2. State-of-the-art

The use of ensemble learning is very popular in the supervised learning community, e.g. boosting [**?**], bagging [**?**]. The adoption of ensemble strategies to improve the estimation of the oulier scores is common in unsupervised outlier detection too [**?** **?**]. Analogously to the boosting and the bagging philosophy there exists sequential [**?**] and parallel [**?**] ensemble versions for mixture of supervised and unsupervised outlier detection algorithms.

The integration of supervised and unsupervised techniques has already been discussed in the literature of fraud detection. In [**?**], Veeramachaneni et al. introduce the AI^2 system which consists in concatenating results from the anomaly detection approach with results from the supervised learning approach. The idea is to use concurrently an ensemble of unsupervised models and one supervised model (Random Forest). Afterwards, the strategy consists in merging both results by selecting the top $\frac{n}{2}$ results from the supervised model and the top $\frac{n}{2}$ results from the unsupervised ensemble. Note that this method requires a strategy to combine the scores deriving from different outlier detection methods and to manage the common observations

belonging to both subsets of $\frac{n}{2}$ unsupervised and $\frac{n}{2}$ supervised outputs. To tackle this issue, the authors proposed to project the different scores in the same space, e.g. to normalize the scores in the $[0, 1]$ interval.

In [**?**], Yamanishi and Takeuchi developed a system to find outliers in an unsupervised dataset through a two steps process. In the first step they use a Gaussian mixture model to score the unsupervised set and to impute labels. Successively, they train a supervised model using the dataset with the previously produced labels. Once the new dataset is available, a supervised model is applied and the points that do not result to be outliers are filtered again with the unsupervised model.

The *best-of-both-worlds* principle, is a sequential approach proposed by Michenkova et al. in [**?**]. They applied multiple unsupervised outlier detection algorithms, in order to transform an initial dataset in a collection of outlier scores. This sequential approach includes unsupervised learning in the first stage and a supervised one in the second. The outlier score vector s^o, obtained by the unsupervised model over the original dataset DS, is used to augment DS: $DS^{'} = (DS, s^o)$. They compared the results in terms of AUC-ROC and using a logistic regression model in three settings: *original dataset alone*, *outlier scores alone* and *original dataset + outlier scores*. Using two datasets, they showed that the classifier improves its accuracy when it uses outlier scores in addition to standard features. The goal of adding multiple outlier scores to standard features is to highlight different aspects of outlierness of the feature space. The great advantage of this approach is that we do not need to normalize or combine scores generated by heterogeneous methods. Furthermore, the supervised method is expected to extract automatically information from these scores.

Recently, a class of outlier detection algorithms emerged with the name of *contextual outlier detection* [**?** **?** **?**]. This class of algorithms aims to find outliers given a *context*. A *context* is a subset of the original dataset and it is usually identified by one or more *contextual attributes*. On the other hand, the *behavioral attributes* are opposed to *contextual attributes* and are used to identify the outlier score for each instance. Two instances with exactly the same *behavioral attributes*, but defined in two different *contexts*, may be respectively identified as outlier and inlier instance.

3. Our Approach

Given the nature of the fraud detection problem, and in particular, the *one-to-many* relationship between cards and transactions [**?** **?**], we propose an extension of the *best-of-both-worlds* principle introduced by Michenkova et al. in [**?**]. The extension consists in the definition of outlier scores (Section 3.1) which consider different levels of granularity (Section 3.2).

3.1. Outlier scores

An outlier score vector can be generated using different unsupervised techniques. In this section we limit the study to five outlier scores: *Z-score*, *PC-1*, *PCA-RE-1*, *IF*, *GM-1*. Out of the considered outlier scores, the simplest is the *Z-score*, which sums the squared univariate Z-scores computed on the features used for the detection. We consider two scores based on Principal Component Analysis (PCA): *PC-1* is the first component of the PCA and *PCA-RE-1* is the reconstruction error obtained as the difference between the original feature vector and the vector reconstructed using the first principal component. Variations of these two scores are denoted by changing the number in the suffix of the score name. So, *PC-2* will be the second principal component and *PCA-RE-2* will be the reconstruction error in the case of values reconstructed using the first two principal components together. *IF* is based on Isolation Forest [**?**] and it uses the length of the path between the root and the leaves of a Random Forest [**?**] as indicator of outlierness. Finally, *GM-1* is a monotone transformation of the membership degree of each point and represents the probability of a point to be generated by a Gaussian Mixture Model. The number of mixture components used to build the model is paired with the prefix GM.

3.2. Global, local and cluster granularity

The approach described by Michenkova et al. in [**?**], to augment the dataset DS, takes into consideration outlier scores computed on the whole set DS. As the cardholder behaviors are very diverse, it might be a sub-optimal solution to compute the outlier scores in a global fashion. Based on the *contextual* outlier detection, this section proposes three main approaches to define *contexts* and compute the outlier score at different levels of granularity:

1. Global granularity: all the transactions are considered as samples of a unique global distribution for which outlier scores can be computed. A transaction is considered anomalous if it deviates from the usual set of

transactions with incongruous values. This approach is the closest to the one discussed in [**?**] since no specificity of the credit-card problem (e.g. the fact that transactions belong to different customers) is taken into account.

2. Local granularity: the computation of the outlier scores is done in a card based manner and a transaction is considered anomalous only if it abnormally differs from the past transactions of the same card.
3. Cluster granularity: this is a compromise between the two previous approaches. The rationale is that both, the global and the card-based approach, have intrinsic limitations. It is not realistic to think that all the genuine cardholders behave in the same manner (large variance) but at the same time too few historical examples are retained if we go down to the card level (large bias). This approach aims to discover whether there is an optimal aggregation level where a reasonable trade-off between the bias and the variance of the two extremes approaches can be reached. The clustering is done at the card level and it is based on a set of features which describe the customer behavior, such as the amount spent over the last 24 hours.

Figure 1 is an illustrative example to better explain the three approaches. In this example, we consider the amount of a transaction as variable to build the outlier score. Our goal is to detect the most suspicious transaction, or to select the extreme transaction, considering only the amount of each transaction. From a global perspective, to detect the most suspicious transaction, we have to consider the overall average amount (65.4 in the example). The highest value (185) recorded for the cardholder CH2 is the most divergent value respect to the average amount. So the cardholder CH2 receives an alert in this case.

In the case of local approach, the most suspicious transaction is determined by the difference between each transaction amount and the average cardholder amount. In this example, cardholder CH3 gets alerted since it has the highest difference between one of its transactions (110) and its average (26.4).

Let us now cluster the cardholders in two groups based on their average spent amount. The first cluster will contain cardholders with an high average consumption (CH1 and CH2), while the second will contain cardholders with a low average consumption (CH3, CH4 and CH5). The average transaction amount is 99.8 for the high consumption group and 39.1 for the low con-

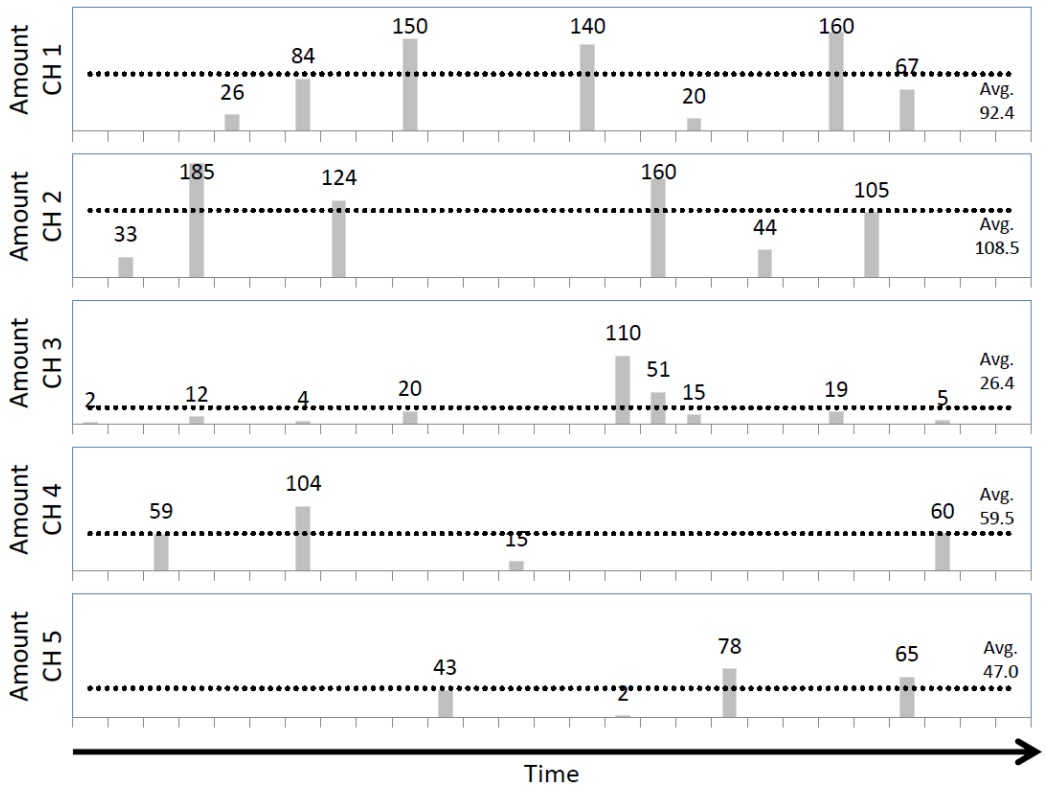


Figure 1: This illustrative example shows the behavior of five cardholders (CHi, $i = 1, \ldots, 5$) over a fixed period of time. The dotted line represents the average expenditure, while the bars represent the amount of each single transaction.

sumption group. In this case, again the cardholder CH2 gets alerted since the transaction with the amount 185 is the most divergent compared to the cluster average amount (99.8).

While it is straightforward to run the global and local approaches, the cluster approach requires the setting of some elements like the clustering algorithm and the features used in the cluster metric. We use the *k-means* algorithm [**?**] as it is simple to interpret, it runs fast with large datasets and it gives the opportunity to a priori decide an arbitrary number of clusters to be identified (letting easily control the aggregation level).

The choice of the features on which to perform clustering (*contextual attributes*) is not trivial since it can have a major impact on the final accuracy. We considered two different sets of features: the first describing the

Algorithm 1 Outlier scores at different levels of granularity

Require: gr ▷ granularity: global, local or cluster
Require: k ▷ number of cluster, if gr == cluster
Require: $cardUsage$ ▷ statistics on the card usage, if gr == cluster
Require: ot ▷ outlier models
Require: D_{tr} ▷ training set
Require: D_{te} ▷ test set

function SCORE($subD_{tr}, subD_{te}, ot$)
  $subOutD_{tr} \leftarrow subD_{tr}$
  $subOutD_{te} \leftarrow subD_{te}$
  for t **in** ot **do**
    $outlierModel \leftarrow$ fit t to $subD_{tr}$
    $trainingScore \leftarrow$ get score of $subD_{tr}$ using $outlierModel$
    $testScore \leftarrow$ get score of $subD_{te}$ using $outlierModel$
    $subOutD_{tr} \leftarrow$ append $trainingScore$ to $subOutD_{tr}$
    $subOutD_{te} \leftarrow$ append $testScroe$ to $subOutD_{te}$
  end for
  return $subOutD_{tr}$, $subOutD_{te}$
end function

if (gr == "*global*") **then** ▷ global granularity
  $(DOut_{tr}, DOut_{te}) \leftarrow$ SCORE(D_{tr}, D_{te}, ot)
end if
if (gr == "*local*") **then** ▷ local granularity
  $DOut_{tr}, DOut_{te} \leftarrow$ empty datasets
  for $card$ **in** D_{tr} **do**
    $(subD_{tr}, subD_{te}) \leftarrow$ SCORE($D_{tr}[cardID == card], D_{te}[cardID == card], ot$)
    $DOut_{tr} \leftarrow \{DOut_{tr}, subD_{tr}\}$
    $DOut_{te} \leftarrow \{DOut_{te}, subD_{te}\}$
  end for
end if
if (gr == "*cluster*") **then** ▷ cluster granularity
  $DOut_{tr}, DOut_{te} \leftarrow$ empty datasets
  $clusterLabel \leftarrow k\text{-}means(cardUsage, k)$
  $D_{tr} \leftarrow \{D_{tr}, clusterLabel\}$
  $D_{te} \leftarrow \{D_{te}, clusterLabel\}$
  for i **from** 1 **to** k **do**
    $(subD_{tr}, subD_{te}) \leftarrow$ SCORE($D_{tr}[cluster == i], D_{te}[cluster == i], ot$)
    $DOut_{tr} \leftarrow \{DOut_{tr}, subD_{tr}\}$
    $DOut_{te} \leftarrow \{DOut_{te}, subD_{te}\}$
  end for
end if

$DOut_{tr}, DOut_{te}$ ▷ augmented training and test set

cardholder's behavior and the second one summarizing personal data of the cardholder. In the former case we considered the average expenditure over the last 24 hours and the number of transactions in the last 24 hours, while in the latter case we considered the age, the nationality[1] and the gender of the cardholder. Since the cardholder behavior led to better accuracy, we only present clusters created by using cardholder behavior features in the experimental part. The principal hyper-parameter of the *k-means* algorithm is the k number of clusters to create. This hyper-parameter is also important for our case study since it defines the level of granularity of the outlier score. In our experiments, we let the hyper-parameter vary from a minimum low granularity of 10 clusters, to a maximum high granularity of 5000 clusters.

Algorithm 1 defines the process of outlier scores construction at different levels of granularity. The function SCORE (row 1) receives as input a training set, a test set and a list of outlier models to be computed. The output is the training and test set augmented with the related outlier scores. While in the case of global granularity (row 13) the entire training and test sets are passed directly to the function SCORE, in the two other cases these sets are split and then each resulting portion of transactions is passed to the function SCORE. For the local approach, the split is done at the card level (row 18) and for the cluster approach, the split is done at the level of the cluster defined at row 26.

3.3. Metrics

Several metrics have been proposed in literature to measure the quality of the detection, notably: i) Area Under the Receiver Operating Characteristic Curve (AUC-ROC) [**?** **?** **?** **?**], ii) Area Under the Precision-Recall Curve (AUC-PR) [**?** **?**], iii) F-measure [**?** **?** **?** **?**], iv) Specificity [**?** **?**], v) Recall [**?** **?** **?**], vi) Precision [**?** **?** **?** **?**].

In this work, we will limit to analyse the metrics which are considered the most relevant in fraud detection[2]: Top$n$ Precision and AUC-PR. Since the final goal of the fraud detection process is to provide the maximal number of true positives within the alerts issued for the investigators, the most pertinent metri is the Top$n$ Precision. This variant of the Precision metric refers to the ratio between the number of true positive alerts and the number of total

[1]To be more precise, we have used the risk of fraud linked to the nationality, which was learned on a similar set of data.

[2]This choice was done in agreement with our industrial partner Worldline.

alerts. We set $n = 100$ in accordance with our industrial partner since this is a plausible amount of suspicious credit cards that a group of investigators can analyse in a day. The dependency of the Top$n$ Precision on the n value has been studied in [**?** **?**] who showed that by decreasing n, the Precision increases at the cost of a reduced Recall.

Metrics that refer to the whole test set (and not simply to alerts) are the AUC-ROC and the AUC-PR. The Area Under the Curve (AUC) is a value in [0,1] which summarizes the relation of two metrics: Recall and False Positive Rate in the case of AUC-ROC; Precision and Recall in the case of AUC-PR. The AUC-ROC is equivalent to the probability that a randomly chosen fraudulent transaction has a score which is higher than the one of a randomly chosen genuine transaction [**?**]. Though the two metrics may look similar, it has been shown that it is better to use AUC-PR in the case of high class imbalance [**?**]. Furthermore, it is known that the optimization of AUC-ROC does not guarantee the optimization of AUC-PR and viceversa [**?**].

4. Experiments

To compute a consistent outlier score for the local approach, we decided to consider only those cards having at least 10 transactions in the training set. Such threshold was set to preserve statistical accuracy in the local approach since any computation of a local outlier score (even the simplest) with less transactions would be inevitably affected by large variance and deteriorate the overall accuracy. Of course such limitation does not hold for the cluster and the global approaches. However, in order to have a fair comparison between methods we kept the same dataset for all methods. Furthermore, we excluded the cards in the test set which were not present in the training set, since in this case it was not possible to pre-compute the outlier score in the training set. To make a fair comparison between the three approaches, we used the same set of cards also for the global and cluster approaches. In the case of cluster, this is beneficial, since a minimum number of transactions is needed to track the customer behavior.

In accordance with the literature, we adopted a random forest model as baseline, given its superiority in credit card fraud detection [**?** **?** **?** **?**]. We used a particular implementation of the random forest, Balanced Random Forest (BRF) [**?**], where each tree is shaped on a balanced subset of the original sample (undersampling is used to balance the two classes). A single

model is trained over the whole training set and tested on 54 days of data (*Static* approach [**?**]). The test set is one week delayed with respect to the training set, in order to emulate the verification latency.

The dataset used for the experiments refers to 334 days of transactions recorded in 2016 from February, 2 to December, 31. The dataset was provided by our industrial partner Worldline, a leader company in transactional services. It includes 76 million transactions and the percentage of frauds is 0.36%. We use transactions until September, 30 as training data while those between October, 8 and December, 31 are used as test set. The week from October, 1 to October, 7 represents the verification latency period and is not used, neither for training nor for testing.

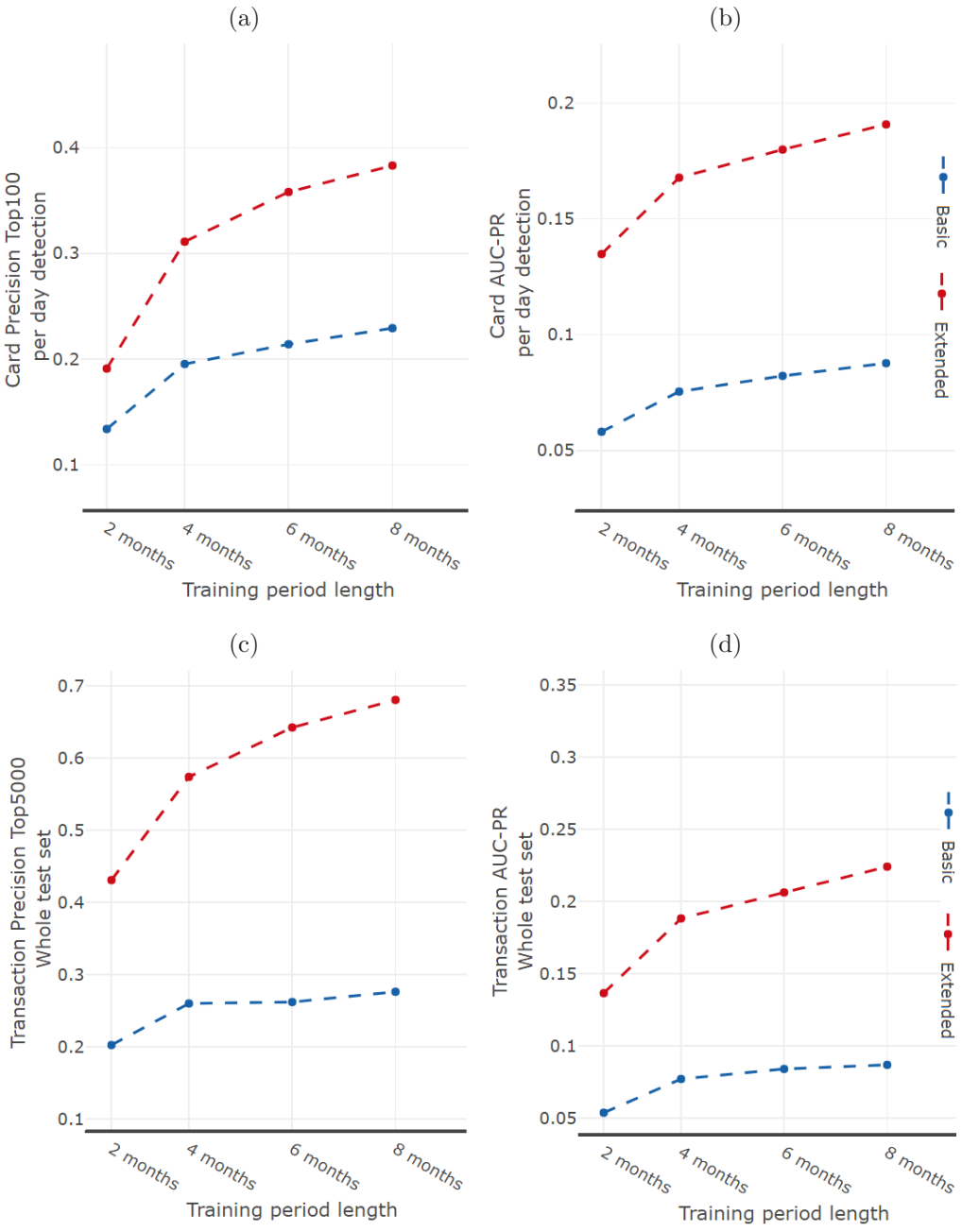
4.1. Baseline

In the this first group of experiments, we trained classifiers using datasets of 2, 4, 6 and 8 months prior to the cut-off date (October, 1), while using the original set of features alone (basic) or with aggregated features (extended). While the basic set includes the *raw* features obtained by our industrial partner Worldline, the extended set refers to the features obtained by feature engineering [**?**]. Examples of engineered features are the sum spent during the last day and the number of transactions executed by a cardholder during the last day, which we mentioned before.

We considered two metrics: the Top$n$ Precision and the AUC-PR. In Figure 2a, the Top100 Precision is plotted, while the Top5000 Precision is reported in 2c. There is a major conceptual difference between these two metrics, as the Top100 precision is computed per day and then averaged over the 54 days of the test period, while the Top5000 is computed on the whole test set without considering the daily label. The same difference occurs in the case of AUC-PR of Figure 2b and Figure 2d.

A second difference is that the detection in Figure 2a and Figure 2b aims to detect the fraudulent cards, while that of Figure 2c and Figure 2d aims to detect the transactions as independent from the cards. This second metric is less useful, since typically once a fraudulent transaction is detected the card gets blocked and no further transaction from that card can be considered. From Figure 2 it appears that, independently of the considered metric, the larger the training set, the better the accuracy [**?**]. Furthermore, the extended set of features leads to a greater accuracy compared to the basic set.

Figure 2: Average daily accuracy for a 54 days test set and for different training lengths: (a) average daily Top100 card Precision , (b) average daily AUC-PR for the fraudulent card detection, (c) the Top5000 Precision for fraudulent transactions detection and (d) the AUC-PR for fraudulent transactions detection.



4.2. Global approach

In this section, we use the best performing configuration of the baseline model, i.e. the one trained on a 8 months period and with the extended feature set (Figure 2). This configuration is compared with a model trained on the same set but with the additional features derived from the outlier scores.

In Figure 3 we can observe that, compared with the Baseline, the use of global outlier scores does not improve significantly the detection. In the worst case (i.e. "All Outliers" where we use only the outlier scores and do not consider the "Baseline" feature space) we assist to a strong deterioration of the accuracy. The outlier scores do not bring additional information even if they are used in combination with the "Baseline" features ("Baseline + All Outliers"). In this case, the additional features increase the risk of overfitting. The incapacity of global outliers in bringing useful information to the predictive model might be related to the fact that such scores are too general and then unable to capture the specificity of a given fraud behavior. It is worth reminding that the mentioned "All Outliers" and "Baseline + All Outliers" correspond respectively to the "proposed" and "proposed+" modalities presented by Michenkova et al. in [**?**].

4.3. Local approach

Figure 4 reports the results of the experiments based on local outlier scores. Whatever the metric, the use of local outlier scores is detrimental for the detection. Among all outliers, the IF local outlier score is the one performing better, while the PCA-RE performs as well as the Baseline model when considering the Precision Top5000 computed on the whole test set. Our interpretation is that while global outlier scores refer to a too large set (large bias), local outlier scores are probably computed on a too restricted set of transactions to be useful (large variance). As for the global outlier scores approach, the use of solely outlier scores (without considering the "Baseline" feature space) produces a very low accuracy.

4.4. Cluster approach

In the experiments based on cluster outlier scores, we let the number of clusters in *k-means* vary from 10 to 5000. In the first analysis all the unsupervised scores presented in Section 3.1 are used to augment the original dataset (Figure 5). Unfortunately it appears that this approach is not beneficial to the detection and the accuracy is even worse than in the global case. While

Figure 3: Accuracy obtained on a test set of 54 days, while using **global outlier scores** as additional features: (a) average daily Top100 Card Precision, (b) average daily AUC-PR for card detection, (c) Top5000 Transaction Precision on the whole test and (d) AUC-PR on the whole test.

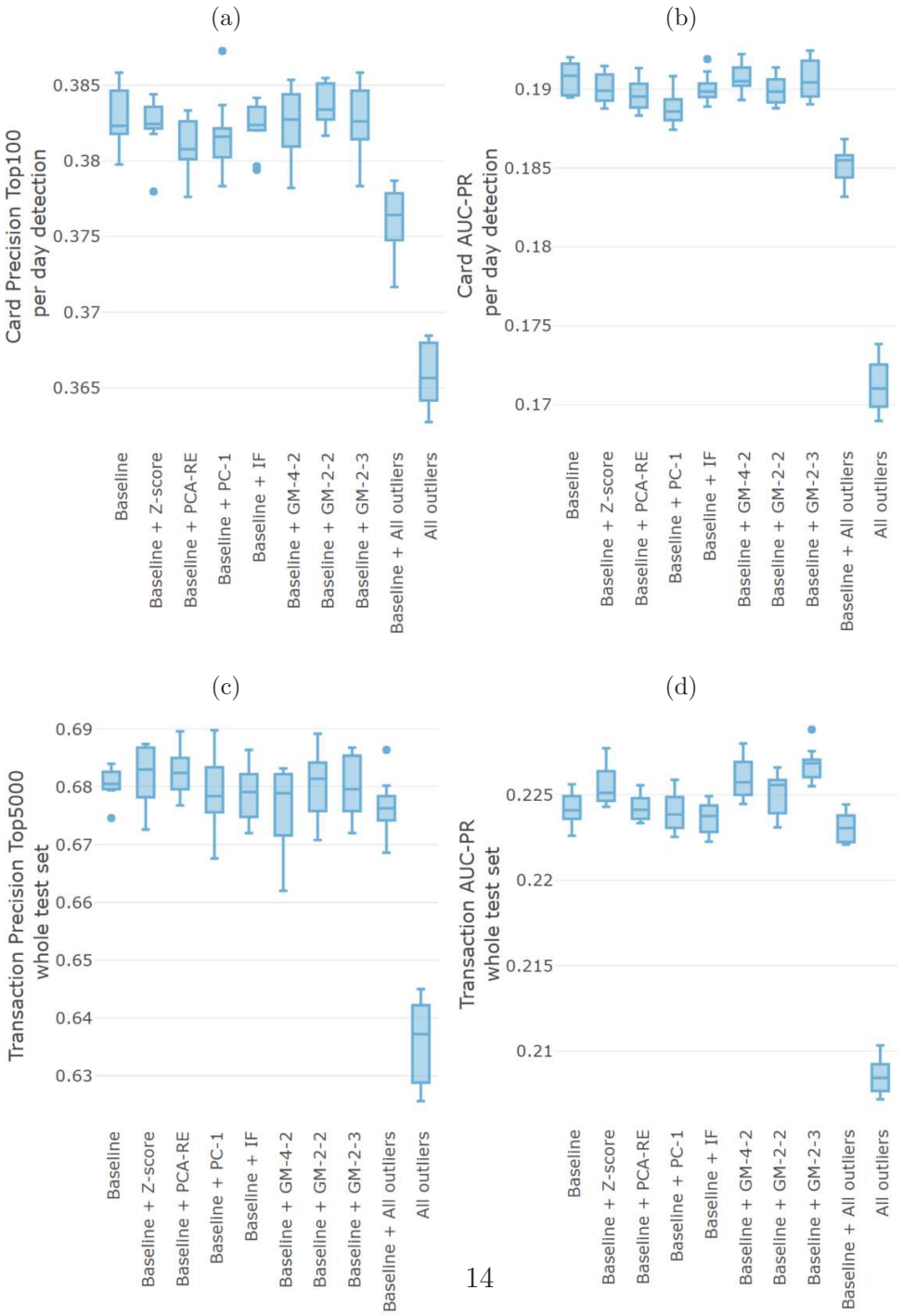


Figure 4: Accuracy obtained on a test set of 54 days, while using **local outlier scores** as additional features: (a) average daily Top100 Card Precision, (b) average daily AUC-PR for card detection, (c) Top5000 Transaction Precision on the whole test and (d) AUC-PR on the whole test.

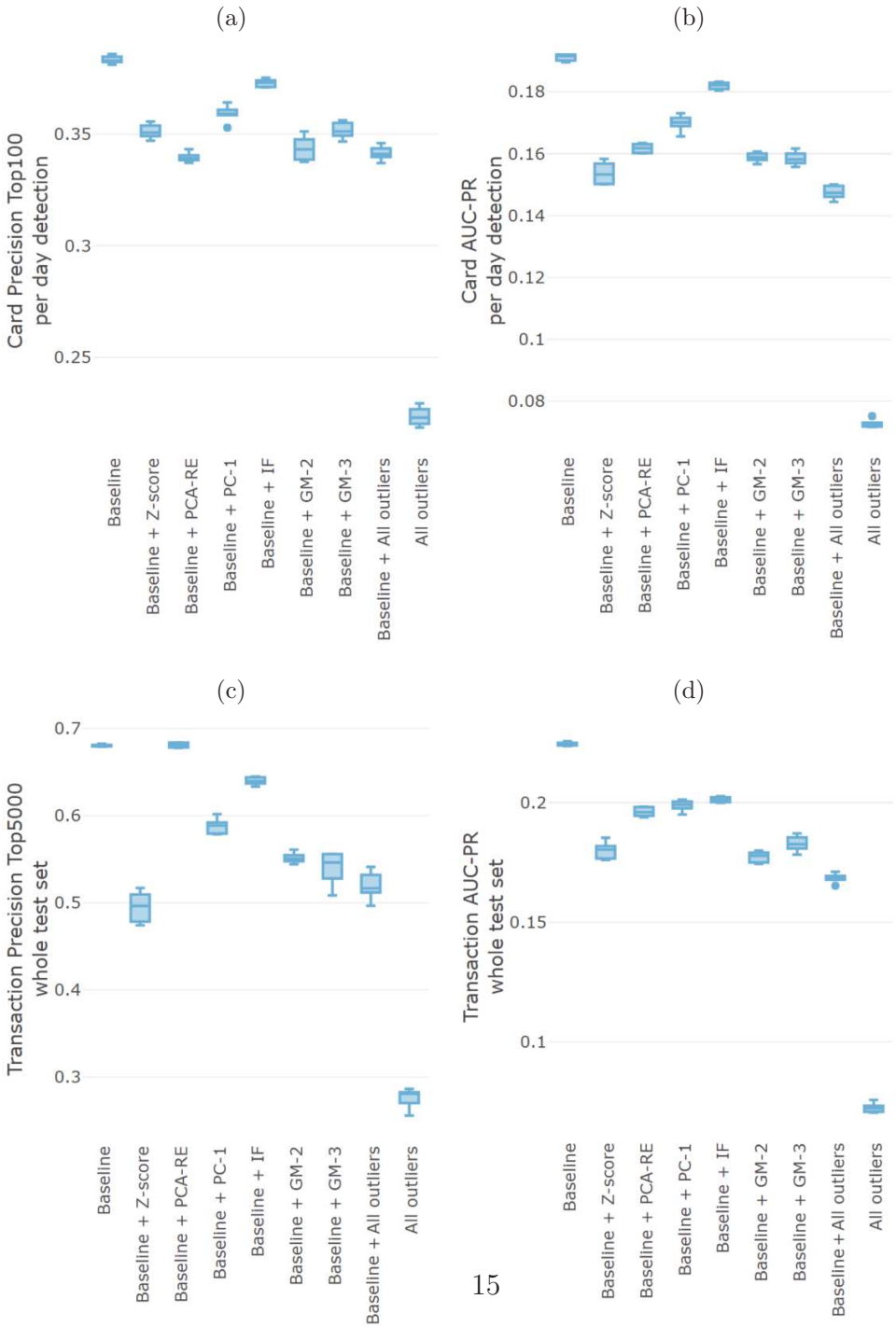
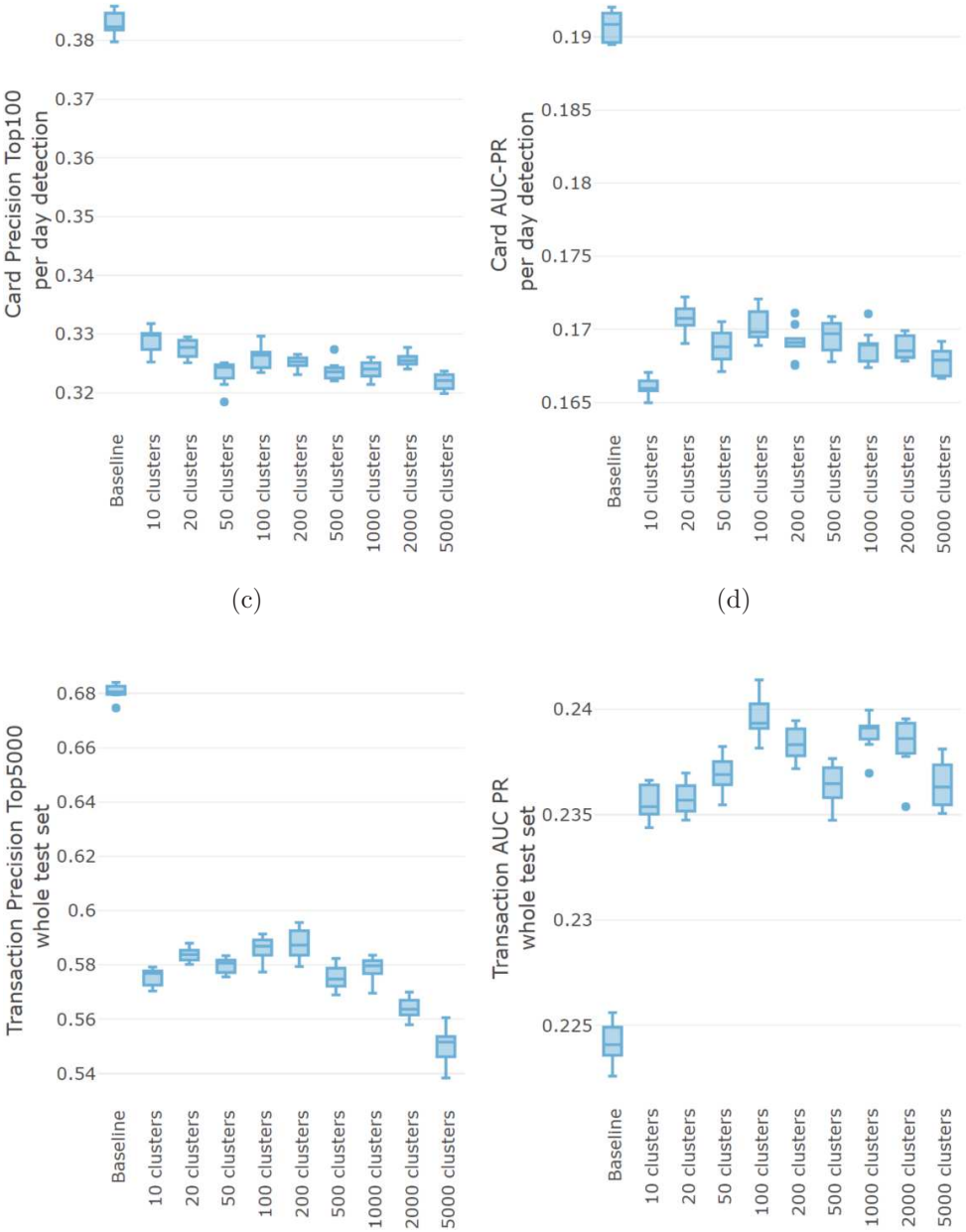


Figure 5: Accuracy obtained on a test set of 54 days, while using **cluster outlier scores** as additional features: (a) average daily Top100 Card Precision, (b) average daily AUC-PR for card detection, (c) Top5000 Transaction Precision on the whole test and (d) AUC-PR on the whole test.

(a)

(b)



(c)

(d)

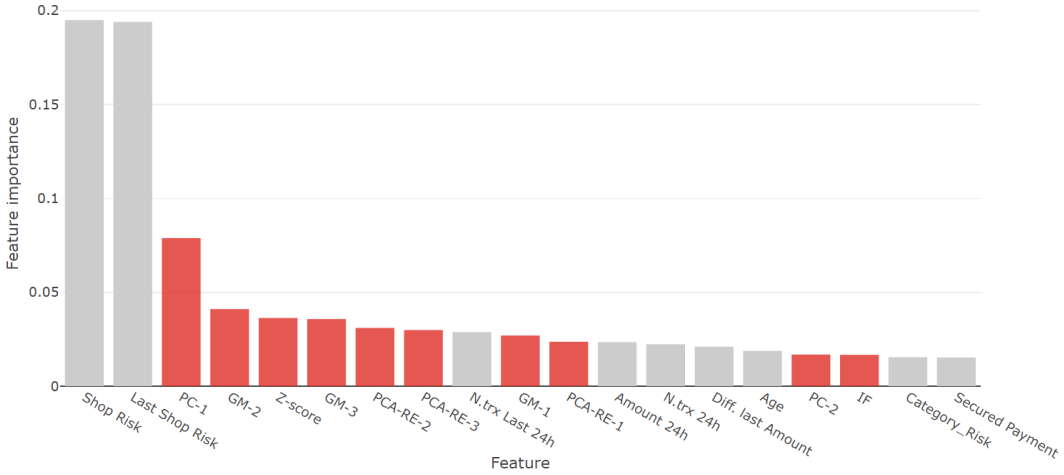


Figure 6: Features ranked by importance, obtained from the random forest classifier of Figure 5 (10 clusters case). The red bars represent the features obtained after scoring the transactions by using particular outlier scores techniques.

considering the Top100 Precision metric, we can see that the case with 10 clusters performs better than other cases even if not better than the baseline.

The second analysis concerns a study of the relevance of the outlier scores with respect to the original features. Several methods exist to assess the relevance of a feature. One of the fastest ways is to rely on the relevance returned by Random Forests. In Figure 6, we show the features used by the classifier ordered by importance (in red the outlier scores). We note that the first two features have a huge impact on the model: they refer to the risk of the shop which received the payment (*Shop Risk*) and the risk of the shop which received the previous payment (*Last Shop Risk*). Many of the outlier scores are just after these two important features, showing that outlier scores could potentially play a key role in the prediction of the risk.

For this reason the third analysis focuses only on augmenting the original dataset with a single outlier. We consider the highest ranked outlier score (*PC-1* in Figure 7) and the second highest one (*GM-2* in Figure 8). The interest of *GM-2* is witnessed also by the fact that this score appears among the ones with the highest Top100 Precision from the global perspective (see Figure 3a).

While there is no improvement for the Top100 Card Precision (Figures 7a and 8a), a significant improvement is visible in terms of Card AUC-PR and

Figure 7: Accuracy obtained on a test set of 54 days, while using **cluster based PC-1 outlier score** as additional feature: (a) average daily Top100 Card Precision, (b) average daily AUC-PR for card detection, (c) Top5000 Transaction Precision on the whole test and (d) AUC-PR on the whole test.

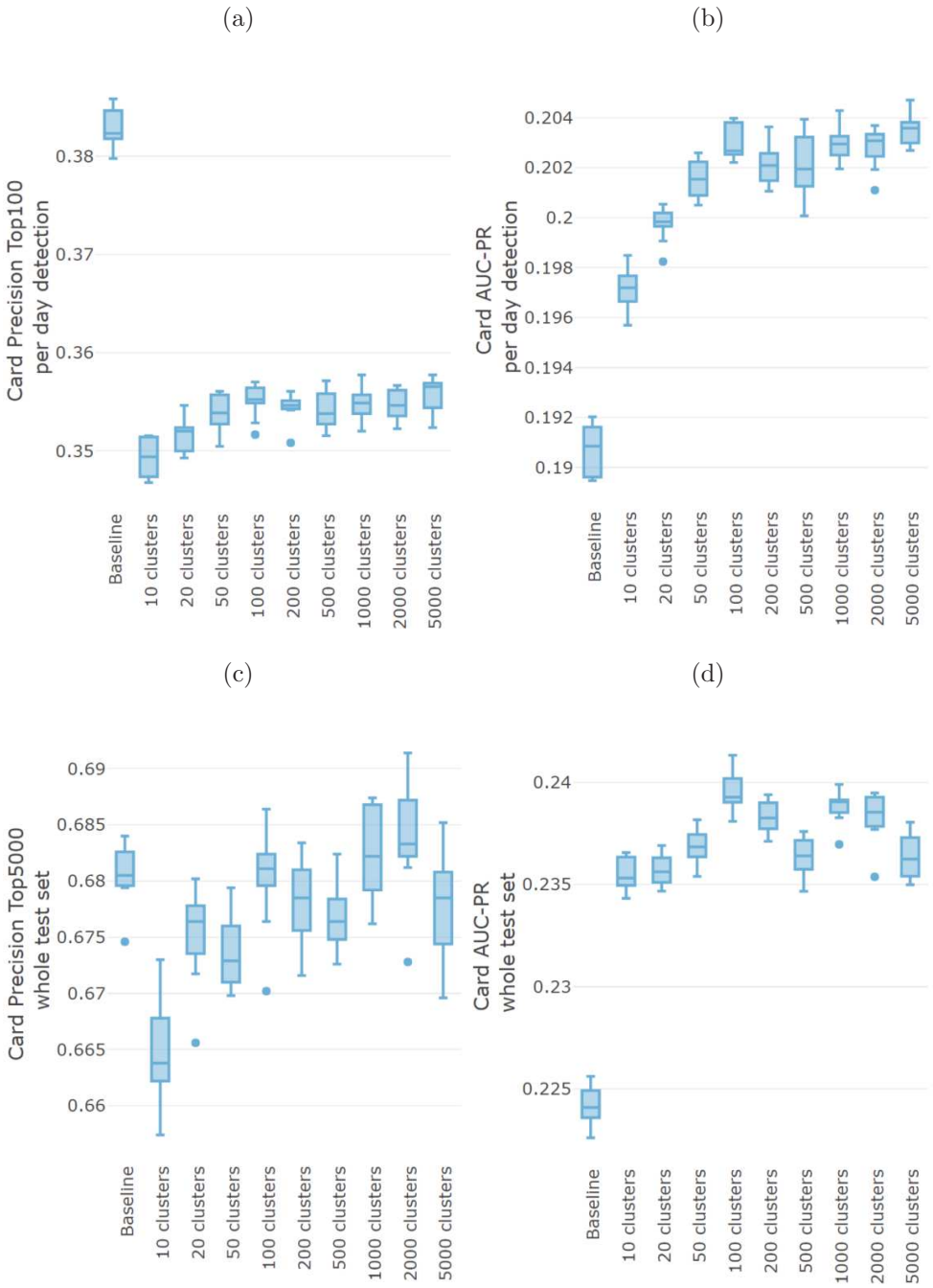
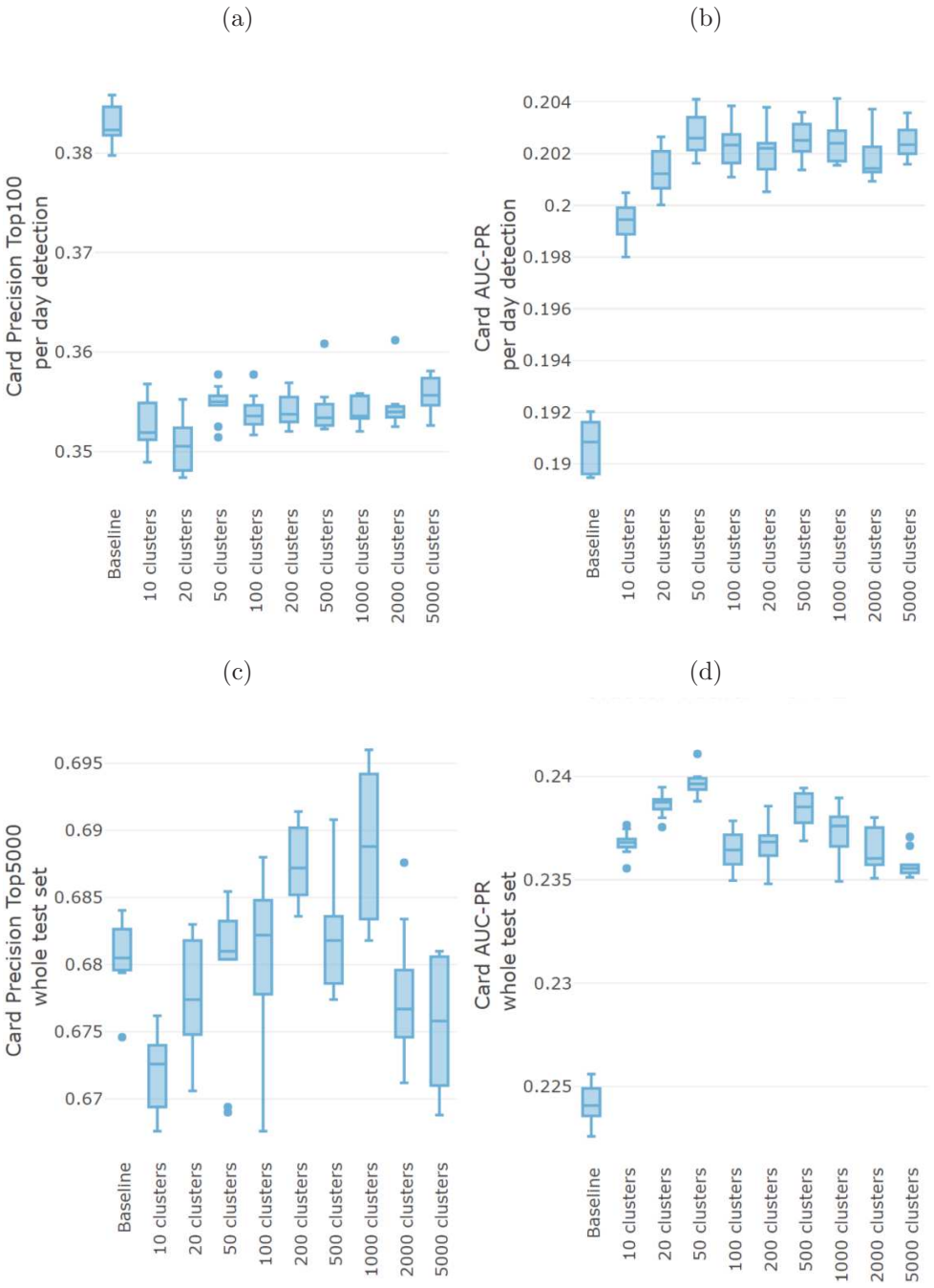


Figure 8: Accuracy obtained on a test set of 54 days, while using **cluster based GM-2 outlier score** as additional feature: (a) average daily Top100 Card Precision, (b) average daily AUC-PR for card detection, (c) Top5000 Transaction Precision on the whole test and (d) AUC-PR on the whole test.



Transaction AUC-PR. In the case of *GM-2* also the Top5000 Transaction Precision is higher.

5. Discussion

Some inconsistencies in the behavior of the Top100 Precision and the AUC-PR metrics used for assessing the cluster approach with respect to the Baseline deserve a deeper analysis. For this reason we report the entire Precision-Recall curves of the "Baseline" and the "10 clusters" local approach based on a combination of baseline features and GM-2 outlier score (in blue and in red respectively in Figure **??**).

First, we remark that the red curve is higher than the blue curve for values of Recall ranging from 0.1 to 0.3. This is coherent with the AUC-PR accuracy observed in Figures 8b and 8d, where the cluster approach outperforms the Baseline one.

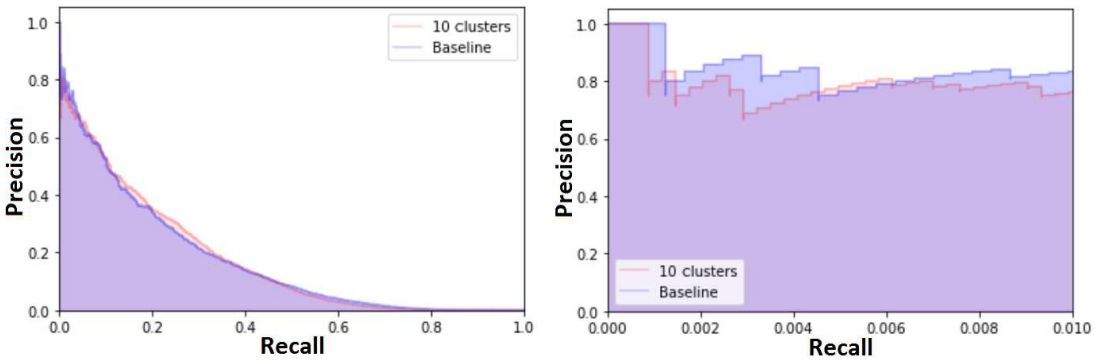
Figure **??** is a zoom of Figure **??** in the Recall interval between 0 and 0.01. We observe here that the Baseline PR curve stands often above the 10 clusters PR curve. This is consistent with the Top100 Card Precision (related to a low recall configuration) illustrated in Figure 8a, where the Baseline approach outperforms the "Baseline + GM-2 outlier score" approach.

The cluster approach showed to be the most promising method, nevertheless it also has some technical limitations. First of all, it requires the choice of some hyper-parameters. The choice of the *contextual attributes* is not trivial (Section 3.2) and the adoption of *k-means* requires the setting of the number of clusters k which has an impact on the level of granularity of the outlier score.

A further disadvantage of such method concerns the need of a minimum number of transactions for a card to be analysed. As mentioned in Section 4, in the local approach we restrained to consider those cards with more than 10 transactions in the training set. Note that the use of such filter could make some fraud patterns non observable.

6. Conclusion

This article proposes an implementation of a hybrid approach which makes use of unsupervised outlier scores to extend the feature set of a fraud detection classifier. The novelty of the contribution, beyond the application to a real massive dataset of credit card transactions, is the implementation



(a) (b)

Figure 9: Precision-Recall curve obtained on a single day of test set while using only the Baseline features and the combination of **cluster based GM-2 outlier score** with Baseline features (10 clusters): (a) Precision-Recall curve for card detection and (b) zoomed Precision-Recall curve for card detection.

and assessment of different levels of granularity for the definition of the outlier score. The considered granularity moves from the card level to the global level, considering intermediate levels of card aggregation by clustering.

The results are not convincing in terms of the global and local approach. Our interpretation is that both approaches are not at the right level of granularity for taking advantage of the unsupervised information. A more promising outcome is obtained in the case of the cluster approach (notably in terms of AUC-PR) though it appears that augmenting the data sets with too many scores could be detrimental because of overfitting and variance issues. The obtained results open the way to several considerations:

- The fact that the *best-of-both-worlds* method brings an improvement in terms of AUC-PR but not in terms of Top$n$ Precision indicates that we are probably addressing different aspects of the problem.
- Additional work in terms of different clustering algorithms and different sets of features for the clustering metric could shed additional light on the relevance of the approach.
- Though many outlier scores seem to bring information about the fraud risk (see relevance plot in Figure 6) using many of them at the same time is detrimental to the final accuracy.

- The impact of the granularity on the final accuracy suggests the interest of analysing the dataset in a stratified manner, not only in an unsupervised perspective but also in a supervised manner (e.g. by introducing some notion of locality).

Acknowledgement

The authors FC, YLB and GB acknowledge the funding of the Brufence and DefeatFraud projects, both supported by INNOVIRIS (Brussels Institute for the encouragement of scientific research and innovation).

References

[] Bahnsen, A. C., Aouada, D., and Ottersten, B. (2015). Example-dependent cost-sensitive decision trees. *Expert Systems with Applications*, 42(19):6609–6619.

[] Bahnsen, A. C., Aouada, D., Stojanovic, A., and Ottersten, B. (2016). Feature engineering strategies for credit card fraud detection. *Expert Systems with Applications*, 51:134–142.

[] Bhattacharyya, S., Jha, S., Tharakunnel, K., and Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. *Decision Support Systems*, 50(3):602–613.

[] Bühlmann, P., Yu, B., et al. (2002). Analyzing bagging. *The Annals of Statistics*, 30(4):927–961.

[] Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32. Springer

[] Carcillo, F., Dal Pozzolo, A., Le Borgne, Y.-A., Caelen, O., Mazzer, Y., and Bontempi, G. (2018a). Scarff: a scalable framework for streaming credit card fraud detection with spark. *Information fusion*, 41:182–194.

[] Carcillo, F., Le Borgne, Y.-A., Caelen, O., and Bontempi, G. (2017). An assessment of streaming active learning strategies for real-life credit card fraud detection. In *Data Science and Advanced Analytics (DSAA), 2017 IEEE International Conference on*, pages 631–639. IEEE.

[] Carcillo, F., Le Borgne, Y.-A., Caelen, O., and Bontempi, G. (2018b). Streaming active learning strategies for real-life credit card fraud detection: assessment and visualization. *International Journal of Data Science and Analytics*, pages 1–16.

[] Carneiro, N., Figueira, G., and Costa, M. (2017). A data mining based system for credit-card fraud detection in e-tail. *Decision Support Systems*, 95:91–101.

[] Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):15.

[] Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., and Bontempi, G. (2017). Credit card fraud detection: a realistic modeling and a novel learning strategy. *IEEE transactions on neural networks and learning systems*.

[] Dal Pozzolo, A., Caelen, O., and Bontempi, G. (2015). When is undersampling effective in unbalanced classification tasks? In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 200–215. Springer.

[] Dal Pozzolo, A., Caelen, O., Le Borgne, Y.-A., Waterschoot, S., and Bontempi, G. (2014). Learned lessons in credit card fraud detection from a practitioner perspective. *Expert systems with applications*, 41(10):4915–4928.

[] Davis, J. and Goadrich, M. (2006). The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240. ACM.

[] Freund, Y., Schapire, R., and Abe, N. (1999). A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 14(771-780):1612.

[] Fu, K., Cheng, D., Tu, Y., and Zhang, L. (2016). Credit card fraud detection using convolutional neural networks. In *International Conference on Neural Information Processing*, pages 483–490. Springer.

[] Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36.

[] Hartigan, J. A. and Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108.

[] Junqué de Fortuny, E., Martens, D., and Provost, F. (2013). Predictive modeling with big data: is bigger really better? *Big Data*, 1(4):215–226.

[] Jurgovsky, J., Granitzer, M., Ziegler, K., Calabretto, S., Portier, P.-E., He-Guelton, L., and Caelen, O. (2018). Sequence classification for credit-card fraud detection. *Expert Systems with Applications*.

[] Liang, J. and Parthasarathy, S. (2016). Robust contextual outlier detection: Where context meets sparsity. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 2167–2172. ACM.

[] Liu, F.T. and Ting, K.M. and Zhou, Z. (2008). Isolation forest In *Eighth IEEE International Conference on Data Mining*, pages 413–422. IEEE

[] Micenková, B., McWilliams, B., and Assent, I. (2014). Learning outlier ensembles: The best of both worldssupervised and unsupervised. In *Proceedings of the ACM SIGKDD 2014 Workshop on Outlier Detection and Description under Data Diversity (ODD2). New York, NY, USA*, pages 51–54.

[] Nguyen, H. V., Ang, H. H., and Gopalkrishnan, V. (2010). Mining outliers with ensemble of heterogeneous detectors on random subspaces. In *International Conference on Database Systems for Advanced Applications*, pages 368–383. Springer.

[] Rayana, S., Zhong, W., and Akoglu, L. (2016). Sequential ensemble learning for outlier detection: A bias-variance perspective. In *Data Mining (ICDM), 2016 IEEE 16th International Conference on*, pages 1167–1172. IEEE.

[] Sethi, N. and Gera, A. (2014). A revived survey of various credit card fraud detection techniques. *International Journal of Computer Science and Mobile Computing*, 3(4):780–791.

[] Shimpi, P. R. and Kadroli, V. (2015). Survey on credit card fraud detection techniques. *International Journal Of Engineering And Computer Science*, 4(11):15010–15015.

[] Song, X., Wu, M., Jermaine, C., and Ranka, S. (2007). Conditional anomaly detection. *IEEE Transactions on Knowledge and Data Engineering*, 19(5):631–645.

[] Van Vlasselaer, V., Bravo, C., Caelen, O., Eliassi-Rad, T., Akoglu, L., Snoeck, M., and Baesens, B. (2015a). Apate: A novel approach for automated credit card transaction fraud detection using network-based extensions. *Decision Support Systems*, 75:38–48.

[] Van Vlasselaer, V., Eliassi-Rad, T., Akoglu, L., Snoeck, M., and Baesens, B. (2015b). Afraid: fraud detection via active inference in time-evolving social networks. In *Advances in Social Networks Analysis and Mining (ASONAM), 2015 IEEE/ACM International Conference on*, pages 659–666. IEEE.

[] Veeramachaneni, K., Arnaldo, I., Korrapati, V., Bassias, C., and Li, K. (2016). Aiˆ 2: Training a big data machine to defend. In *Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing (HPSC), and IEEE International Conference on Intelligent Data and Security (IDS)*, pages 49–54. IEEE.

[] Wei, W., Li, J., Cao, L., Ou, Y., and Chen, J. (2013). Effective detection of sophisticated online banking fraud on extremely imbalanced data. *World Wide Web*, 16(4):449–475.

[] Yamanishi, K. and Takeuchi, J.-i. (2001). Discovering outlier filtering rules from unlabeled data: combining a supervised learner with an unsupervised learner. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 389–394. ACM.

[] Zareapoor, M. and Shamsolmoali, P. (2015). Application of credit card fraud detection: Based on bagging ensemble classifier. *Procedia Computer Science*, 48:679–685.

[] Zhu, X. (2005). Semi-supervised learning literature survey. Technical report, Computer Sciences TR 1530, University of Wisconsin Madison.

Zimek, A., Campello, R. J., and Sander, J. (2014). Ensembles for unsupervised outlier detection: challenges and research questions a position paper. *Acm Sigkdd Explorations Newsletter*, 15(1):11–22.