International Journal of Data Science and Analytics
https://doi.org/10.1007/s41060-018-0116-z

APPLICATIONS

Streaming active learning strategies for real-life credit card fraud detection: assessment and visualization

Fabrizio Carcillo[1] · Yann-Aël Le Borgne[1] · Olivier Caelen[2] · Gianluca Bontempi[1]

Received: 13 November 2017 / Accepted: 22 March 2018

Abstract
Credit card fraud detection is a very challenging problem because of the specific nature of transaction data and the labeling process. The transaction data are peculiar because they are obtained in a streaming fashion, and they are strongly imbalanced and prone to non-stationarity. The labeling is the outcome of an active learning process, as every day human investigators contact only a small number of cardholders (associated with the riskiest transactions) and obtain the class (fraud or genuine) of the related transactions. An adequate selection of the set of cardholders is therefore crucial for an efficient fraud detection process. In this paper, we present a number of active learning strategies and we investigate their fraud detection accuracies. We compare different criteria (supervised, semi-supervised and unsupervised) to query unlabeled transactions. Finally, we highlight the existence of an exploitation/exploration trade-off for active learning in the context of fraud detection, which has so far been overlooked in the literature.

Keywords Active learning · Fraud detection · Selection bias · Semi-supervised learning

1 Introduction

The use of machine learning for credit card fraud detection requires to address a number of challenges. Some of them are related to the data distribution, notably the class imbalance of the training set (many more genuine transactions than fraudulent ones), the non-stationarity of the phenomenon (due to changes in the behavior of customers as well as in fraudsters), the large dimensionality and the overlapping classes (while fraudsters try to emulate cardholders behavior, genuine behaviors of cardholders might look strange or anomalous).

The labeling process is constrained, as every day human investigators may contact only a small number of cardholders associated with the riskiest transactions and obtain the class (fraud or genuine) of the related transactions. The high cost of human labor, for assessing the transaction labels, leads to the labeling bottleneck [2]. In this context, an automatic fraud detection system (FDS) should support the activity of the investigators by letting them focus on the transactions with the highest fraud probability. From the perspective of the transactional service company, this is crucial in order to reduce the costs of the investigation activity and to retain the customer confidence. From a machine learning perspective, it is important to keep an adequate balance between *exploitation* and *exploration*, i.e., between the short-term needs of providing good alerts to investigators, and the long-term goal of maintaining a high accuracy of the system (e.g., in the presence of concept drift).

The issue of labeling the most informative data by minimizing the cost has been extensively addressed by *active learning* which can be considered as a specific instance of *semi-supervised learning* [8,42], the domain studying how unlabeled and labeled data can both contribute to a better

This paper is an extension version of the DSAA'2017 Application Track paper titled: "An Assessment of Streaming Active Learning Strategies for Real-Life Credit Card Fraud Detection" [6].

✉ Fabrizio Carcillo
fabrizio.carcillo@ulb.ac.be

Yann-Aël Le Borgne
yleborgn@ulb.ac.be

Olivier Caelen
olivier.caelen@worldline.com

Gianluca Bontempi
gbonte@ulb.ac.be

[1] Machine Learning Group, Computer Science Department, Faculty of Sciences, Université Libre de Bruxelles (ULB), Brussels, Belgium

[2] R&D, Worldline, Brussels, Belgium

Published online: 04 April 2018

learning process. The active learning (AL) literature has proposed a number of techniques to select, within a large set of unlabeled instances, the most informative to label. AL can be typically described by the iteration of three steps [42]:

1. *Selection* given a query budget of size k, a model is used to choose which k data points, once labeled, could better describe the data generating process;
2. *Querying* selected points are submitted to an *oracle* (in our case an investigator) for labeling;
3. *Training* the labeled data points are used to train and update the model;

The *Selection* and *Querying* steps differentiate active learning from conventional (passive) learning, which is limited to the *Training* step. Note also that, in order to bootstrap the procedure, a random initialization or an unsupervised model is used [42].

A first classification of AL methods can be done according to the model used for the *Selection* step. In this paper, we will distinguish between methods using only supervised information and methods integrating other sources of information, notably unsupervised or semi-supervised. Additionally, according to the nature of the dataset, we can distinguish between two main AL approaches: *pool based* and *streaming* AL. In pool-based AL, the algorithm performs queries in the same set of unlabeled points, while in stream-based AL, the set of unlabeled data points is periodically updated. The accuracy of a pool-based AL classifier is expected to grow in time, since more and more labeled data points, from the original dataset, are used for the *Training*. This is not always true in the case of the streaming approach, since data received in different periods may differ significantly (e.g., concept drift).

Though AL is widely addressed in the literature [11,25], few articles mention the aspects of credit card fraud detection (Sect. 2.2), notably the class imbalance [17,56] (in our case study, approximately only 0.2% of transactions[1] are fraudulent), the restricted labeling budget [37,49] and the specific nature of the assessment [41]. It is worth to remark that in credit card fraud detection, though the observations are at the level of transactions, the ultimate goal is to detect fraudulent cards. There exists a one-to-many relation between cards and transactions.

A peculiarity of fraud detection is that the labeling and the assessment phase are coincident and consequently strongly dependent. For the transactional service provider, it is important that the investigations are as successful as possible (i.e., low false- positive rate). This means that the accuracy of a FDS is measured in terms of precision over the top k alerted credit cards [5,13]. This is not always the case in other AL tasks where the labels of the k queried points are not directly related to the accuracy of the training process. In a FDS, it is not only important to minimize the labeling cost but also that this labeling allows to discover as many frauds as possible. The nature and the intensity of the exploration step have a strong impact on the final accuracy of detection, and accordingly, the set of state-of-the-art AL strategies which are effective in practice is much more limited than expected. In other terms, no real-life FDS can afford a totally random labeling process since this would necessarily imply an unacceptable short-term random performance. This exploitation/exploration trade-off inherent to fraud detection, to the best of our knowledge has not been addressed in the research literature.

The contributions of this article are: (i) a taxonomy of streaming AL strategies (and a number of their variants) for credit card fraud detection, (ii) an extensive comparison of these techniques for the detection of both fraudulent transactions and cards, (iii) an experimental assessment on the basis of a massive set of 12 million transactions in terms of real-life criteria (defined by our industrial partner, Worldline, a leader company in transactional services) and (iv) a two-dimensional visualization of the effects of active learning on the distribution of the training set. The outcome is an original analysis of the exploitation/exploration trade-off in the context of a real-world FDS. In particular, we expect that visualization represents a valuable insight on the evolution of the training set during active learning. In fact, though the rationale of active learning is explicit, it is not always evident to understand how active learning modifies the distribution of the training set, notably in a task, like credit card fraud detection, characterized by large noise, nonlinearity and non-separability. Also, we expect that potential bias of AL sampling strategies is easier to assess in a visual setting.

The work is organized as follows. Section 2 presents the related state of the art. Section 3 provides the general outline of our fraud detection system. Section 4 introduces the visualization technique used throughout the manuscript. Section 5 discusses a number of AL strategies (as well as possible variants) for dealing with streaming credit card transactions. Finally, Sect. 6 presents an extensive experimental session based on a real stream of transactions.

2 Related work

In this section, we will present some important state-of-the-art works in fraud detection. The review of conventional passive learning for fraud detection is presented in the

[1] Though some papers on fraud detection present datasets with still lower rates (0.01% in [15], 0.005% in [2], 0.02% in [51] and 0.004% in [36]), our dataset is inline with other recent works on fraud detection ([47], [22] and [39] have a class imbalance rate of 0.8, 0.5 and 0.4%, respectively).

Sect. 2.1, while active learning techniques and their application to fraud detection are introduced in Sect. 2.2.

2.1 Passive learning for fraud detection

Credit card fraud detection belongs to the largest domain of outlier detection [1,7,32] also called anomaly or novelty detection. The main approaches to anomaly detection are:

1. *Classification based* these supervised techniques make the assumption that a classifier of anomalies can be learnt from a training set and that the test set distribution is not significantly different from the training one. Several supervised machine learning algorithms have been discussed in the literature [40,43,53]. In most cases, an optimal detection can be obtained by combining multiple supervised machine learning techniques. For instance, Wei et al. [51] introduce ContrastMiner, a fraud detection framework which combine the use of contrast pattern mining, neural network and decision forest to have a high precision in the detection. In our previous research, we have used and assessed several binary classifiers for fraud detection [12,13]. Also one-class classifiers, like one-class SVM and Isolation Forest, belong to this category. One-class SVM [38] fits a boundary around the known set of normal points and then classifies as outlier everything which stands over the boundary. The isolation forest [27] uses the length of the path between the root and the leaves in the trees of a random forest as outlier score.
2. *Nearest neighbor based* the rationale is that the distribution of the neighbors (or local density) of a point characterizes the genuine or abnormal nature of a point. In particular, normal instances stand in dense areas, while outliers are located far from dense areas. Local outlier factor (LOF) is a well-known density/neighbor-based technique proposed by Breunig et al. in [4]. Variants of LOF have been proposed in [23,34,46,55].
3. *Clustering based* once the training set is properly clustered, outliers are expected to be located much farther away from the clusters center than normal ones.
4. *Statistically based* they require the estimation of the multivariate distribution of data and return a score of outlierness which is inversely proportional to the density of a point. Gaussian mixture models (GMMs) [16,19,26,54] are commonly used in this context because of their flexible semi-parametric nature. Hidden Markov models have been used in [45] for credit card outlier detection. Bolton et al. [3] presented two statistical approaches which rely on monitoring the anomalous behavior of a cardholder compared to a group of peers (peer group analysis) or to cardholder behavior (break point analysis).
5. *Information theoretic* they rely on the fact that anomalies create irregularities in the dataset and can be detected with information theoretic measures.
6. *Spectral anomaly detection* these techniques transform the original dataset in a way that facilitate the separation of normal instances and outliers. A commonly used transformation technique is the principal component analysis (PCA) [28,33,50]. Shuy et al. [44] proposed to combine the effect of the major component and the minor component. While the major component is used to find extreme values, the minor component highlights those observations which do not conform to the normal correlation structure.
7. *Outlier detection combined with feature selection* unsupervised outlier detection has been recently combined with feature selection. Pang et al. [30] proposed a filter-based feature selection framework for unsupervised outlier selection. This approach improved the outlier detection rates while substantially reducing the number of features by 46% (on average). The same authors proposed a novel wrapper-based outlier detection framework [29] to iteratively optimize the feature subset selection and the outlier score. The results showed that the framework improved not only the AUC-ROC, but also the precision over the top-n ranked instances.

2.2 Active learning

Unlike passive learning, active learning modifies the size and nature of the training set by choosing from a unlabeled set a subset of points whose label is expected to improve the classifier. The AL setting is particularly promising in fraud detection because of the cost and the delay related to the labeling of instances. However, the adaptation of active learning to the specific characteristics of fraud detection data has only been partially addressed in the research literature. Fan et al. [18] carried out an empirical analysis on a fraud detection dataset to assess AL approach in the presence of concept drift. They have focused on the adaptation ability of the AL strategy, but did not address the detection accuracy. Pichara et al. [31] tested a large-scale anomaly detection approach in a synthetic dataset emulating the fraud process. Their AL schema has been able to detect the whole subset of frauds using a number of queries smaller than a Bayesian network detection approach. Multiple tests were repeated using different data noise levels, and their AL approach consistently outperformed the other techniques.

However, the use a synthetic dataset reduces the impact of these results. It is very difficult to create a reliable and synthetic credit card dataset, since transactions (frauds and genuine) are very diverse and evolve in an unpredictable way. Van Vlasselaer et al. [48] applied active inference, a network-based algorithm, to fraud discovery in social secu-

rity real data. They found that committee-based strategies, based on uncertainty, result in a slightly better classification performance than expert-based strategies. Nevertheless, expert-based strategies are often preferred in order to obtain unbiased training sets from queries.

The relationship between active learning and streaming data, notably the sampling bias issue, is discussed in [14]. The authors showed that in stream-based active learning, the estimated input–output dependency changes over time and depends on the instances previously queried. Since those instances are typically selected next to the class decision boundary of the classifier, this may lead to a biased representation of the underlying data distribution. AL and concept drift are also addressed in [57]. Here, the authors stressed how concept drift may be missed in regions far from where AL queries normally take place (e.g., boundary regions between classes). The authors showed that techniques based on classical uncertainty sampling favor close concept drift adaptation, while techniques based on random sampling are more effective in dealing with remote concept drift. The best performing techniques can strongly depend on the characteristics of the data and the size of the query budget.

The issue of sampling bias is specifically discussed in [20]. Jacobusse and Veenman present multiple solutions to tackle the sampling bias on highly imbalanced datasets and screening applications. In such conditions, a small group of targets need to be detected among the nontargets vast majority. A random selection would lead to a very poor detection. So, usually an expert knowledge-driven selection is preferred (medical screening application, law enforcement, screening of job applicants, etc.). The authors emphasize that this selection is prone to suffer of a strong bias toward the previous knowledge of the expert, causing the classifier to be trained on a non-representative dataset. A similar problem is faced in fraud detection, since only a small subset of transactions can be labeled in the short term and the selection of this subset corresponds to the riskiest credit card transactions.

The integration of AL and semi-supervised learning is discussed in Xie and Xiong [52]. They introduced a stochastic semi-supervised learning (SSSL) process to infer labels in case of large imbalanced datasets with small proportion of labeled points. The approach relies on the consideration that since the number of unlabeled points is huge and the minority class is rare, the probability of making a wrong majority assignment is very low. Consequently, they proposed the assignment of the majority class to random selection of points and adopted it with success in the context of a data competition.

Finally, an original approach that may be used to deal with the one-to-many relationships between cards and transactions is discussed in [41]. They present an AL approach for multiple instance problems where instances are organized into *bags*. Typical examples of multiple instance problems are found in text classification and content-based image retrieval. In these types of problems, a bag is said to be positive if it includes at least one instance which is positive, while the bag is negative if no positive instances are observed in it.

3 The fraud detection system classifier

Let us consider a fraud detection system (FDS) whose goal is to detect automatically frauds in a stream of transactions. Let $x \in \mathbb{R}^n$ be the vector coding the transaction (e.g., including features like the transaction amount and the terminal) and $y \in \{+, -\}$ the associated label, where $+$ denotes a fraud and $-$ a genuine transaction. A detection strategy needs a measure of risk (score) associated with any transaction. In a machine learning approach, this score is typically provided by the estimation of the a posteriori probability $\mathcal{P}_{\mathcal{C}}(+|x)$ returned by a classifier $\mathcal{C}$. We consider a streaming setting where unlabeled transactions arrive one at a time or in small batches.

The FDS goal is to raise every day a fixed and small number of k alerts. In our industrial case study, $n = 32$ and k is set to 100 on the basis of cost and work organization considerations.

The issuing of those alerts has two consequences: the trigger of an investigation and the consequent labeling of the associated transactions. The outcome of the investigation determines both the success rate of the FDS and the new set of labeled transactions.

We will present in Sect. 6 two levels of experimental validations: the first concerns the detection of fraudulent transactions, while the second focuses on fraudulent cards. In the first experiment, the classifier $\mathcal{C}$ is implemented by a conventional random forest, while in the second, we use a more complex approach (ensemble of classifiers) dictated by the more challenging nature of the detection tasks. This approach has been presented in [5,13] and consists of the weighted average of two classifiers

$$\mathcal{P}_{\mathcal{C}}(+|x) = w^A \mathcal{P}_{\mathcal{D}_t}(+|x) + (1 - w^A)\mathcal{P}_{\mathcal{F}_t}(+|x) \quad (1)$$

where $\mathcal{D}_t$ and $\mathcal{F}_t$ stand for *Delayed classifier* and *Feedback classifier*, respectively, and $w^A \in [0, 1]$ is the weight controlling the contribution of the two classifiers. $\mathcal{D}_t$ is implemented as an ensemble of balanced random trees [10,35] trained on old transactions for which we can reasonably consider the class as known. $\mathcal{F}_t$ is trained on recently alerted transactions, for which a *Feedback* was returned by investigators. It is therefore alimented by the active learning component of the fraud detection system. This *Feedback* component is very important to address concept drift.

This architecture is the result of an extensive model selection and assessment procedure which has been discussed in our previous work [5,13]. The aim of this paper is to discuss the

Table 1 Summary of active learning and semi-supervised strategies described in the paper

Id	Strategy	Type
HRQ	Highest risk querying	Baseline/BL (Sect. 5.1)
R	Random querying	Exploratory active learning/EAL (Sect. 5.2)
P	Unsupervised (PCA) querying	
U	Uncertainty querying	
M	Mix of random and uncertainty querying	
SR	SSSL on random points	Stochastic semi-supervised learning/SSSL (Sect. 5.3)
SU	SSSL on uncertain points	
SM	SSSL on random/uncertain points	
SE	SSSL on points most likely to be genuine	
SR-U	SSSL on uncertain points + random sampling	SSSL + EAL (Sect. 5.3)
SR-R	SSSL on random points + random sampling	
SR-M	SSSL on random/uncertain points + random sampling	
SRN[p]	SR with reduced x% of negative feedback	Modified SSSL (Sect. 5.3)
ROS	Random oversample	Oversample (Sect. 5.4)
SMOTE	SMOTE	
QFU	Querying by frequent uncertainty	Multiple instance learning (Sect. 5.5)
MF-...	Max combining function	
SM-...	Softmax combining function	
LF-...	Logarithmic combining function	

impact of different AL strategies, so we will not take into consideration alternative classifier architectures.

4 Dataset visualization

Credit card fraud detection deals with high-volume (millions of transactions) and large dimensionality ($n = 32$ in our example) datasets. So, it is difficult to have a visual insight of the data distribution and the nature (e.g., non-separability) of the classification task. As a consequence, most papers rely only on predictive accuracy (e.g., ROC curves) to assess the difficulty of the task. Given the dynamic and sampling nature of AL , it is however relevant to visualize the location of the selected query points with respect to the original data distribution in order to better illustrate the differences between alternative strategies. In the following section, we will complement the presentation of the different AL strategies with a visual representation of our experimental dataset (details are in Sect. 6) in the space of the two first principal components (denoted by PC1 and PC2). Figure 1 shows the class conditional distributions of the two classes (fraud in red and genuine in blue) in the PC1/PC2 space.

This visualization provides interesting insights about the distribution of frauds and genuine transactions: (i) the two classes appear to be only partially overlapping in the space, (ii) the density of the fraudulent set appears to have a higher variance than the genuine set, and (iii) as far as the first principal component is concerned, frauds have a distribution skewed to the left.

However, the class conditional nature plot should not mislead us to the conclusion that the problem is easy to solve. In fact, given the high imbalance of the classes, a large number of genuine transactions still occur in the left part of the plot (mostly red).

In Section 5, we will use Fig. 1 as a template for illustrating the distribution of the queries issued by the different AL strategies taken into consideration.

5 Active learning strategies

The rationale of AL is to select (on the basis of current information) unlabeled training samples which, once labeled, can improve the accuracy. However, there are two main unknowns concerning the effectiveness of AL in credit card fraud detection. The first concerns the strong imbalance of the class distribution: as the selection of adequate queries is the most important step of an AL procedure. This step should take into account that in such a large imbalanced problem, selecting majority class points will inevitably have a negligible impact on accuracy. The second concerns the definition of

Algorithm 1 Active Learning process

Require: k ▷ total number of alerts
Require: q ▷ exploration budget
Require: m ▷ SSSL budget
Require: D ▷ initial training set
1: **for** any new day **do**
2: $\mathcal{C} \leftarrow$ learning(D)
3: $inTrx \leftarrow$ unlabeled set
4: $scores \leftarrow \{\mathcal{P}_{\mathcal{C}}(x), x \in inTrx\}$
5: $sel \leftarrow$ selection of the $k-q$ points with highest risk $scores$ ▷ HRQ
6: **if** $(q > 0)$ **then** ▷ EAL
7: $Esel \leftarrow q$ explorative points
8: $sel \leftarrow \{sel, Esel\}$
9: **end if**
10: $queries \leftarrow$ investigator labeling of sel
11: **if** $(m > 0)$ **then** ▷ SSSL
12: $SSSLset \leftarrow m$ points based on a SSSL criterion
13: $SSSLset \leftarrow$ set label $y(SSSLset) = 0$
14: $queries \leftarrow \{queries, SSSLset\}$
15: **end if**
16: $D \leftarrow \{D, queries\}$
17: **end for**

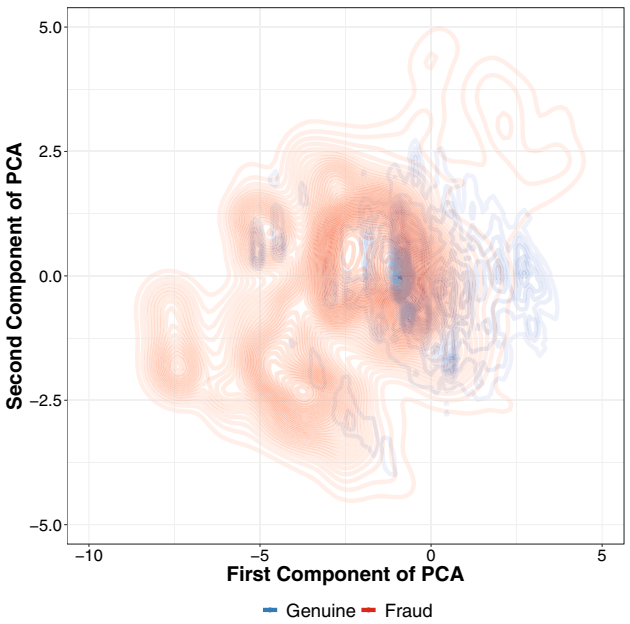

Fig. 1 Class conditional distributions (level curves of frauds in red and of genuine in blue) of transactions in the PC1/PC2 space. The dataset of transactions is collected on 15 consecutive days

accuracy: measures of detection accuracy are strictly related to the capacity of discovering frauds, i.e., querying minority class samples. This means that an AL strategy for fraud detection requires some specific tuning for being successful. To illustrate the impact of AL on FDS, we will start by considering a baseline strategy which simply queries the highest risk transactions on the basis of the current classification model. This strategy will be denoted as the highest risk querying (HRQ). Thereafter, we will introduce and assess a number

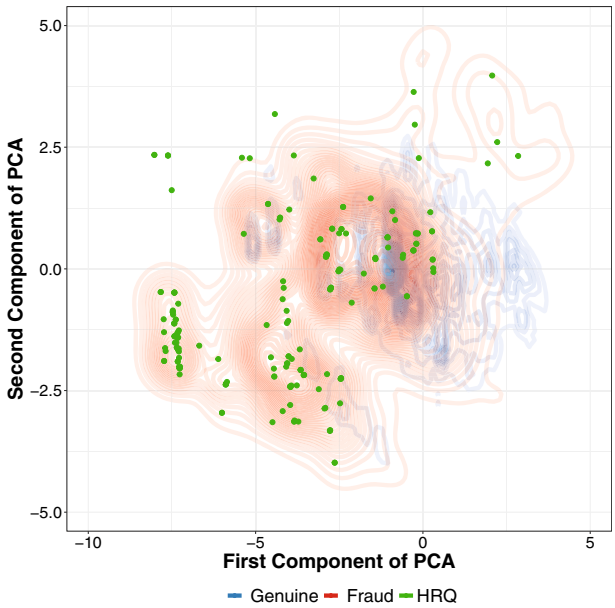

Fig. 2 Class conditional distributions in the PC1/PC2 space and the set of 200 transactions queried by the HRQ approach

of modifications of HRQ according to several principles. In order to make the comparison easier, we will define each AL strategy as an instance of a generic AL strategy detailed in Algorithm 1. The algorithm requires the specification of three parameters: the budget k of queries (i.e., maximum number of transactions that can be investigated per day), the number of q queries defined for exploration purposes and the number m of unlabeled transactions that can be set as genuine without investigation (see 5.3). When $q > 0$, the choice of the queries demands a selection criterion which plays a major role in the final accuracy of AL. Note that the criteria used by the methods discussed in the following sections can be regrouped in three classes: supervised (i.e., relying on labeled data), unsupervised and semi-supervised. The entire list of discussed AL strategies is presented in Table 1.

5.1 Highest risk querying

The idea of highest risk querying (HRQ) is simple: given a classifier $\mathcal{C}$ and a budget of queries, HRQ returns the unlabeled transactions with the highest estimated *a posteriori* probability $\mathcal{P}_{\mathcal{C}}(+|x_i)$. HRQ is the most intuitive active learning strategy for our problem if we consider that the final FDS accuracy depends on the amount of minority class querying. Note that in terms of the pseudocode in Algorithm 1, HRQ is obtained by setting $q = 0$ and $m = 0$.

HRQ is expected to have a positive impact on accuracy by discovering new instances from the minority class and improving consequently the balance of the training set. It has also some drawbacks: since its querying strategy relies

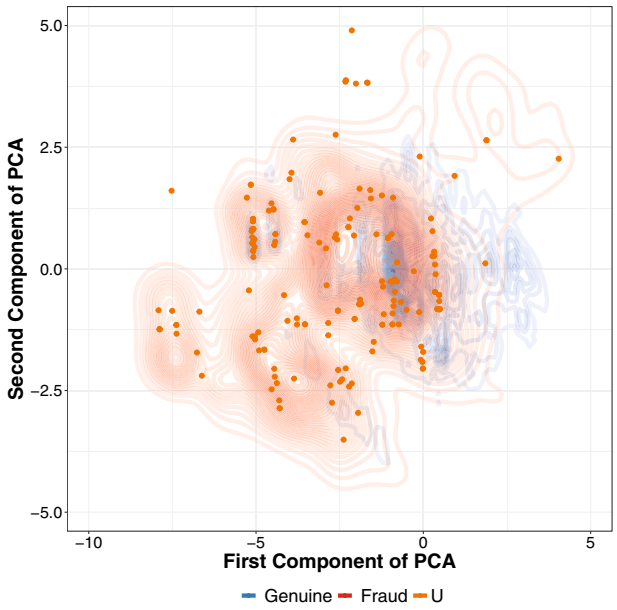

Fig. 3 Class conditional distributions in the PC1/PC2 space and the set of 200 transactions queried by the Uncertainty Sampling approach

on the classifier accuracy, this selection step could be inaccurate especially at the very beginning.
Figure 2 represents a set of 200 query points returned by HRQ. As expected, the classifier $\mathcal{C}$ selects points located in areas where the class conditional density of frauds is high (red areas). The training of a new classifier using exclusively this subset of points may suffer of selection bias, since the sample is not representative of the distribution of the genuine class.

5.2 Exploratory active learning

Exploratory active learning (EAL) strategies modify HRQ by trading exploitation for exploration. The idea is to convert a subset of the labeling budget in explorative queries. The size of the exploration budget is represented by $0 < q \leq k$ in Algorithm 1.
We may consider a number of exploration techniques for selecting the q exploratory transactions. The simplest one is random querying (denoted by EAL-R) which consists in choosing randomly the q query points. This selection is unsupervised and sub-optimal since it may query points for which the classifier is already highly confident about the class. A less naive unsupervised strategy (denoted by EAL-P) consists in using an unsupervised algorithm (e.g., PCA) to select the q queries.
An alternative with a supervised selection criterion is represented by uncertainty querying (EAL-U) which returns unlabeled data points for which the current classifier has low confidence [24]. The selection criterion is therefore supervised: given a binary classifier $\mathcal{C}$, the uncertainty querying strategy gives priority to the transactions x_i for which $\mathcal{P}_{\mathcal{C}}(+|x_i) \approx 0.5$. This value may be affected by class imbalance, and an higher threshold may be chosen in the fraud detection setting. Nevertheless, it is a good practice to balance the dataset before training the classifier. In the experimental section, we will implicitly refer to a classifier $\mathcal{C}$ which is trained on a balanced dataset.
Figure 3 presents the visualization of a set of 200 transactions selected by the uncertainty sampling approach. The transactions are expected to be selected in areas where the classifier $\mathcal{C}$ is the most uncertain. As a consequence, selected transactions will lie in areas which present a high density for both classes.
Žliobaitė et al. [57] proposed the mix of the two techniques, uncertainty querying and randomization, to tackle remote concept drift (Sect. 2). The technique (denoted by EAL-M) consists in querying by uncertainty most of the points and in querying random points from time to time.

5.3 Stochastic semi-supervised learning

The stochastic semi-supervised learning (SSSL) strategy has been introduced by Xie and Xiong [52] to infer labels in case of highly imbalanced datasets with a large number of unlabeled points. The strategy relies on a simple consideration: since the ratio between the number of frauds and the total number of transactions is very small, the probability of randomly selecting a fraud is very low.
The resulting AL learning schema is made of four steps:

1. *Selection* the current model is used to annotate all unlabeled transactions with an estimated risk;
2. *Querying* the highest risk transactions are labeled by the investigators;
3. *Majority assumption* a number of transactions are labeled as genuine by majority assumption; in this paper, we explore a number of criteria to attribute the majority class: pure random attribution (SR), uncertainty (SU), mix of randomness and uncertainty (SM) and low predicted risk (SE).
4. *Training* the labeled data points, obtained by the previous steps, are used to train/update a supervised model.

Though the selection strategy is supervised as in EAL, SSSL differs in terms of the usage of the current model $\mathcal{C}$: the predicted risk is not only used to alert and trigger the investigation but also to label (without investigation) a number of low-risk transactions.
In order to illustrate the reliability of the majority assumption, we report in Fig. 4 the distribution of the scores $\mathcal{P}_{\mathcal{C}}(+|x_i)$ over 15 days. In particular, the histograms (a), (b) and (c) refer to the score distribution for all transactions, genuine

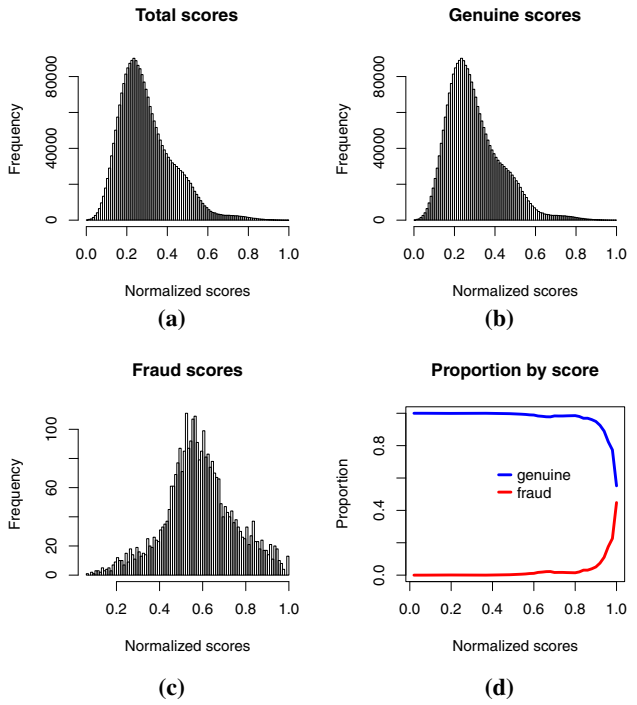

Fig. 4 Distribution of the scores obtained by $\mathcal{P}_{\mathcal{A}_t}(+|x_i)$ in a range of 15 days involving 2.4 millions of cards for: all the transactions (**a**), only genuine transactions (**b**) and only fraudulent transactions (**c**). In (**d**), the proportion of genuine and fraudulent transactions is plotted while changing the score obtained by $\mathcal{P}_{\mathcal{A}_t}(+|x_i)$

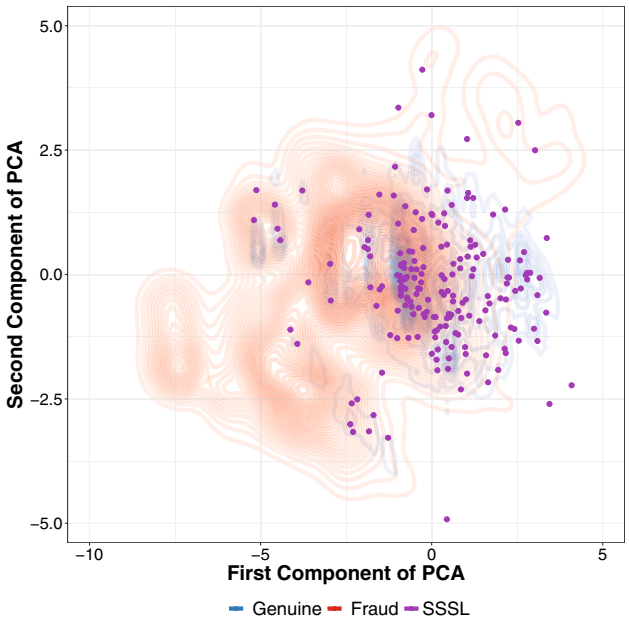

Fig. 5 Class conditional distributions in PC1/PC2 space and the set of 200 transactions selected using the stochastic semi-supervised learning (SR) approach

and fraudulent transactions, respectively. The plot (d) represents the proportion of fraudulent and genuine transactions for a given score in the range [0, 1]. Note that, though the a priori proportion of fraudulent cards in the dataset is 0.13%, it becomes 23.33% for scores higher than 0.95 and 61.90% for scores beyond 0.99. Note also that, in the area of maximal uncertainty for $\mathcal{C}$ (e.g., between 0.49 and 0.51), we find only 0.35% of frauds.

On the basis of those considerations, it is possible to define a number of stochastic semi-supervised strategies:

- SR: no exploration budget ($q = 0$) and attribution of the majority class to $m > 0$ random transactions;
- SU: no exploration budget ($q = 0$) and attribution of the majority class to the $m > 0$ most uncertain points;
- SM: no exploration budget ($q = 0$) and attribution of the majority class to the $0.7 \times m$ most uncertain points and to $0.3 \times m$ random points;
- SE: no exploration budget ($q = 0$) and attribution of the majority class to the $m > 0$ lowest risk points;

Additional variants can be created by simply allowing an exploration budget ($q > 0$). The SR-U, SR-R and SR-M strategies are hybrid strategies which combine an exploration strategy (e.g., U in SR-U) and a SSSL strategy (e.g., SR in SR-U).

Figure 5 represents the set of 200 transactions selected by an SR approach. The transactions are selected randomly and independently of the classifier $\mathcal{C}$. It is important to note that this sampling strategy basically follows the class conditional density of the genuine class. This is a consequence of the high imbalance of the dataset, since the distribution of frauds is negligible compared to the distribution of genuine transactions.

Jacobusse and Veenman [20] presented additional variants of the SSSL. They partition the original dataset $\mathcal{D}$ under study in three subsets:

- $\mathcal{D}^+$: the subset of known positive samples, which corresponds in our scenario to the subset of frauds in the feedback set received by the investigators;
- $\mathcal{D}^-$: the subset of known negative samples, which corresponds to the subset of genuine transactions in the feedback set received by the investigators;
- $\mathcal{D}^0$: the subset of unlabeled samples, which corresponds to the set of unlabeled transactions.

While SSSL proposes to add part of the samples in $\mathcal{D}^0$ and all the labeled samples in $\mathcal{D}^+$ and $\mathcal{D}^-$, the variant proposed by Jacobusse and Veenman suggests to remove $\mathcal{D}^-$ or part of it. The idea is to reduce the selection bias which can arise when the process of selection is driven in an expert knowledge way. In our context, the expert knowledge selection is replaced by

the classifier $\mathcal{C}$. In the experimental section, we will refer to this approach as SRN[p], with p being the percentage of negative samples retained from the feedback set.

5.4 Oversampling

It is worth noting that a side effect of the adoption of SSSL (Sect. 5.3) is to add a number of majority class samples to the training set. This goal is typically achieved by oversampling techniques, with the main difference that here the target class is the majority class and not the minority one. In order to assess how SSSL situates with respect to conventional oversampling, we also consider a comparison with the two main oversampling techniques: random oversample (ROS) [21] and SMOTE [9]. ROS consists in duplicating some random instances from the class to be oversampled until a given sample size is reached. SMOTE creates artificial instances from the target class in the following manner: once the k nearest neighbors from the same class have been identified, new artificial transactions are generated moving along the line segment joining the original instance and its k neighbors.

5.5 Multiple instance learning

This section deals with another specificity of the credit card fraud detection problem: the observations take place at the level of transactions but what is relevant for the company is the detection at the card level, since the investigation is performed at the card level and not at the transaction level. From an AL perspective, since multiple transactions map to the same card, we could select query points by taking advantage of such one-to-many relationship.

5.5.1 Querying by frequent uncertainty

The rationale of querying by frequent uncertainty (QFU) boils down to query those cards which are mapped to the largest number of uncertain transactions. We associate with each card c a counter representing how many of its associated transactions $x_i \in c$ are uncertain, i.e., have a score $\mathcal{P}_{\mathcal{C}_t}(+|x_i) \in [0.5 - v, 0.5 + v]$ where v determines the size of the uncertainty range. The counters are updated in real time, and the AL selection returns the k cards with the highest counters.

5.5.2 Combining function

A more advanced strategy to deal with card detection is inspired by [41]. A *combining function* can be used to aggregate all the posterior probabilities $p_i^c = \mathcal{P}_{\mathcal{C}}(+|x_i)$ of the transactions $x_i \in c$ and derive the posterior probability $\mathcal{P}_{\mathcal{C}}(+|c)$.

Table 2 Scoring of transactions

Rank	Card Id	Trx Id	p_i^c
1	A	A7	0.90
2	B	B3	0.88
3	B	B5	0.87
4	A	A2	0.83
…	…	…	…

Table 3 Scoring of cards on the basis of transactions from Table 2 with three combining functions

Card	max (p_i^c)	softmax (p_i^c)	log. (p_i^c)
A	**0.90**	0.87	34.21
B	0.88	**0.88**	**34.55**
…	…	…	…

Table 4 Additional transaction

Rank	Card Id	Trx Id	p_i^c
20,000	B	B6	0.20

The simplest combining function is the *max* function (denoted MF), which returns

$$\mathcal{P}_{\mathcal{C}}^{MF}(+|c) = \max_{x_i \in c} p_i^c \tag{2}$$

Alternatively, authors in [41] propose the *softmax combining function*:

$$\mathcal{P}_{\mathcal{C}}^{SM}(+|c) = \frac{\sum_{x_i} p_i^c e^{\alpha p_i^c}}{\sum_{x_i} e^{\alpha p_i^c}} \tag{3}$$

where α is a constant that determines the extent to which *softmax* approximates a *max* function.
In order to (i) increase the sensitivity of the card risk to high-risk transactions and to (ii) reduce its sensitivity to low-risk transactions, we propose a *logarithmic combining function* returning the score

$$\sum_{x_i \in c} -\frac{1}{\log s_i^c} \tag{4}$$

where $s_i^c = \begin{cases} p_i^c - \epsilon & \text{if } p_i^c > 0.5 \\ \epsilon & \text{otherwise} \end{cases}$ and ϵ is a very small number.
Table 3 illustrates the scores associated with the transactions of Table 2 for the three combining functions presented above. It appears that, unlike the *max* function, the other two functions are able to take into account the impact of multiple risky

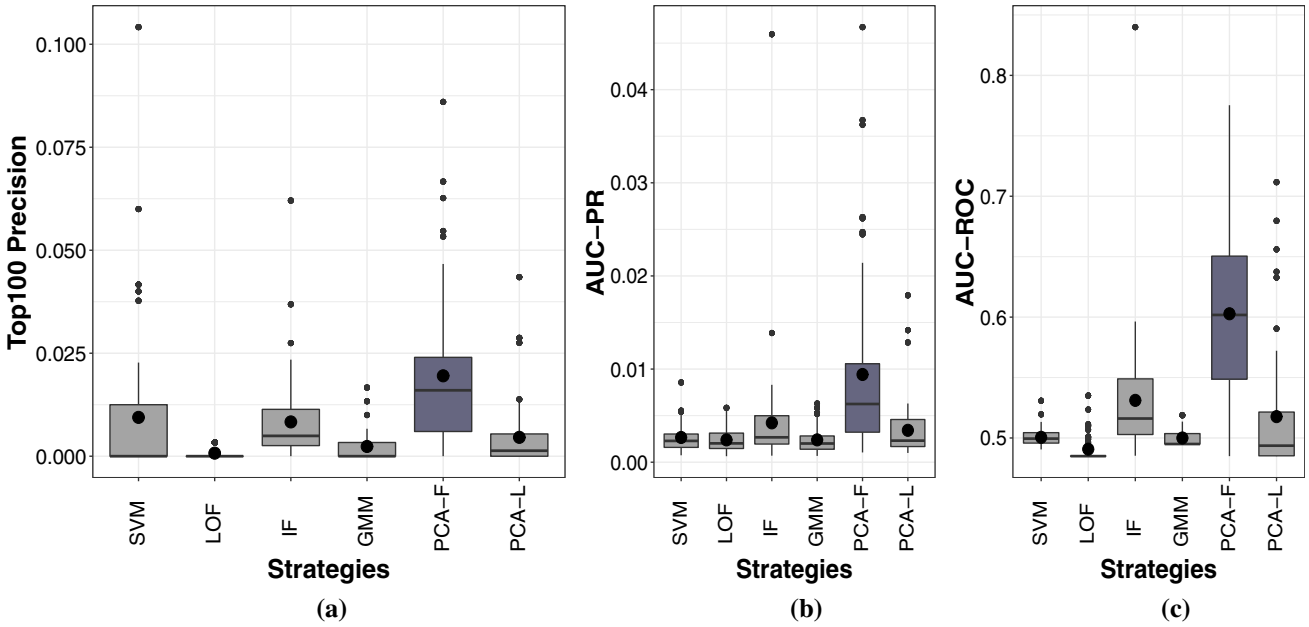

Fig. 6 Transaction-based case study with unsupervised passive learning. Boxplots summarize the accuracy measures obtained over 60 days and 20 trials. Black points indicate the mean value for each box. Comparison in terms of: Top100 precision (**a**), area under the precision–recall curve (**b**) and area under the receiver operator characteristic curve (**c**). Dark boxes indicate the best strategy (paired Wilcoxon test)

Table 5 Scoring of cards on the basis of transaction from Tables 2 and 4 with three combining functions

Card	max (p_i^c)	softmax (p_i^c)	log. (p_i^c)
A	**0.90**	**0.87**	34.21
B	0.88	0.74	**34.55**
…	…	…	…

transactions on the overall risk of a card. In other terms, two high-risk transactions weight more than a simple one with a marginal higher risk. However, the softmax and the logarithmic functions differ in the importance they give to low-risk transactions. Suppose we add a low-risk transaction (Table 4) for card "B" to the set of transactions of Table 2. Table 5 shows that the sensitivity of the card risk to such additional transaction is much larger in the *softmax* than in the logarithmic case. The counterintuitive consequence is that according to the *softmax* function the card "B" becomes now less risky than the card "A".

6 Experiments

This section relies on a large imbalanced dataset of 12 million credit card transactions provided by our industrial partner Worldline.

In this realistic case study, only a very small number ($k = 100$) of cards per day can be queried, amounting to roughly 0.2% of labeled points. The dataset has 32 features, and it covers 60 days, each day including roughly 200 K transactions.

Two sets of experiments are performed: the first measures the detection accuracy at the level of the transactions, while the second measures the detection accuracy at the card level. In the first study, for the sake of simplicity, the classification model $\mathcal{C}$ is a conventional random forest model $\mathcal{RF}$, while a more realistic model $\mathcal{A}$ (discussed in [13] and in Sect. 3) is used for the cards[2]. Since the randomization process in $\mathcal{RF}$ and $\mathcal{A}$ may induce variability in the accuracy assessment, we present the results of twenty repetitions of the streaming.

All the AL strategies are compared in identical situations and initialized with the same random and balanced set (initial training set D presented in algorithm 1). The results are presented as boxplots summarizing the fraud detection performance over the 60 days. In particular, we have considered the following accuracy measures: Top100 precision, area under the precision–recall curve (AUC-PR) and area under the receiver operator characteristic curve (AUC-ROC). In all the plots, the dark boxes are used to denote the most accurate AL strategy as well as the ones which do not differ significantly from it (paired Wilcoxon signed rank test with 5% significance level).

The precision over the Top100 alerts is expected to be larger for $\mathcal{RF}$ than $\mathcal{A}$ since multiple positive alerts for the same card

[2] The use of two different learning strategies is justified by the need to assess the robustness of the AL strategies with respect to different learning methods and different detection tasks (transaction based and card based).

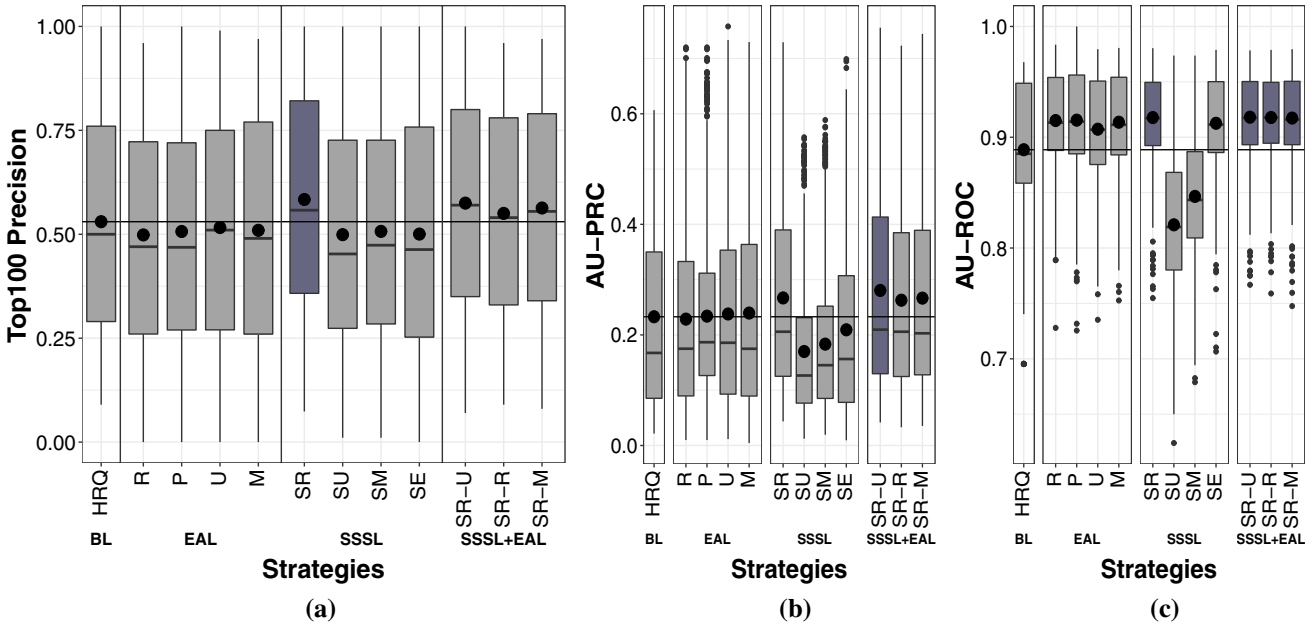

Fig. 7 Transaction-based case study. Boxplots summarize the accuracy measures obtained over 60 days and 20 trials. Black points indicate the mean value for each box, and the horizontal line indicates the mean for the baseline HRQ. The extended names for the strategies listed on the horizontal axes are found in Table 1. Comparison in terms of: Top100 precision (**a**), area under the precision–recall curve (**b**) and area under the receiver operator characteristic curve (**c**). Dark boxes indicate the best strategy and those which are not statistically different (paired Wilcoxon test)

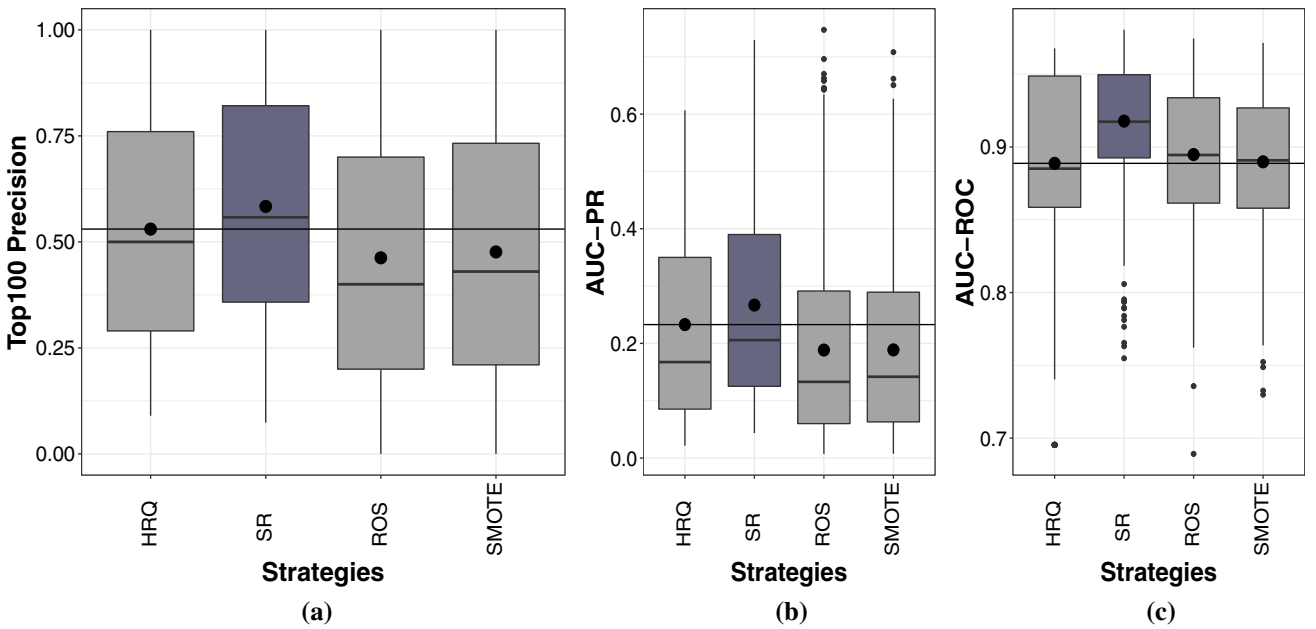

Fig. 8 Transaction-based case study. Comparison in terms of: Top100 precision (**a**), area under the precision–recall curve (**b**) and area under the receiver operator characteristic curve (**c**). Dark boxes indicate the best strategy and those which are not statistically different (paired Wilcoxon test)

will be accounted as several true positives in the transaction case but as a single success in the card case. We made all the code available on Github[3].

[3] https://github.com/fabriziocarcillo/StreamingActiveLearningStrategies.

6.1 Transaction-based fraud detection

In order to better situate the results of AL techniques, we start by reporting the accuracy of several passive (Sect. 2.1) unsupervised methods. In Fig. 6, we show the accuracy of six

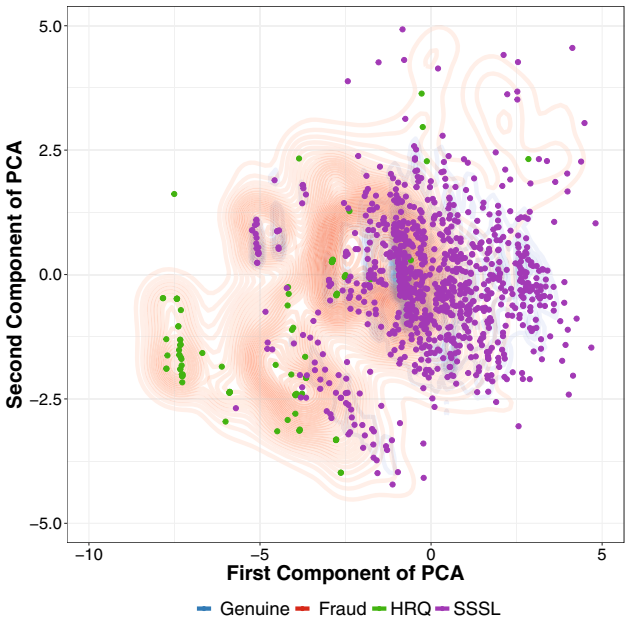

Fig. 9 Class conditional distributions in the PC1/PC2 space and the transactions selected by the SR strategy, the best one in our experimental comparison. A set of 100 transactions is selected using the highest risk querying approach (green dots), and a random set of 1000 transactions is selected and labeled as the genuine class (SSSL, purple dots)

state-of-the-art unsupervised outlier detection techniques: one-class support vector machine (SVM), local outlier factor (LOF), isolation forest (IF), Gaussian mixture models (GMMs), first component of the PCA (PCA-F) and last component of the PCA (PCA-L). Though PCA-F works better in terms of Top100 precision, AUC-PR and AUC-ROC, overall the accuracy of passive unsupervised learning is very low.

In Figure 7, we report the fraud detection accuracy of the AL techniques discussed in Sect. 5. A horizontal line is added in order to make the comparison with the baseline strategy HRQ easier. The experiments are performed with $k = 100$, $q = 5$ and $m = 1000$. These hyper-parameters have been set by trial and error and are compatible with the kind of exploration effort that our industrial partner could ask to its investigators.

The first remark is that AL (Fig. 7) significantly outperforms passive learning unsupervised techniques (Fig. 6). Second, it appears that exploratory AL alone is not able to outperform the standard HRQ strategy. In particular, unsupervised explorations are of no use: the PCA based exploration EAL-P as well as random exploration EAR are significantly worse than the baseline HRQ. The highest accuracy is obtained by SSSL (SR) or by combining SSSL with uncertainty sampling (SR-U). The SR strategy leads to an improvement in precision of 5.84%, while SR-U leads to an improvement of 5.15%. Similar improvements are observed for the AUC-PR (Fig. 7b), while a wider range of techniques perform better in terms of AUC-ROC curve (Fig. 7c). The most efficient combination in our setting is therefore obtained by a combination of stochastic semi-supervised approach with the standard HRQ strategy for active learning (Fig. 9).

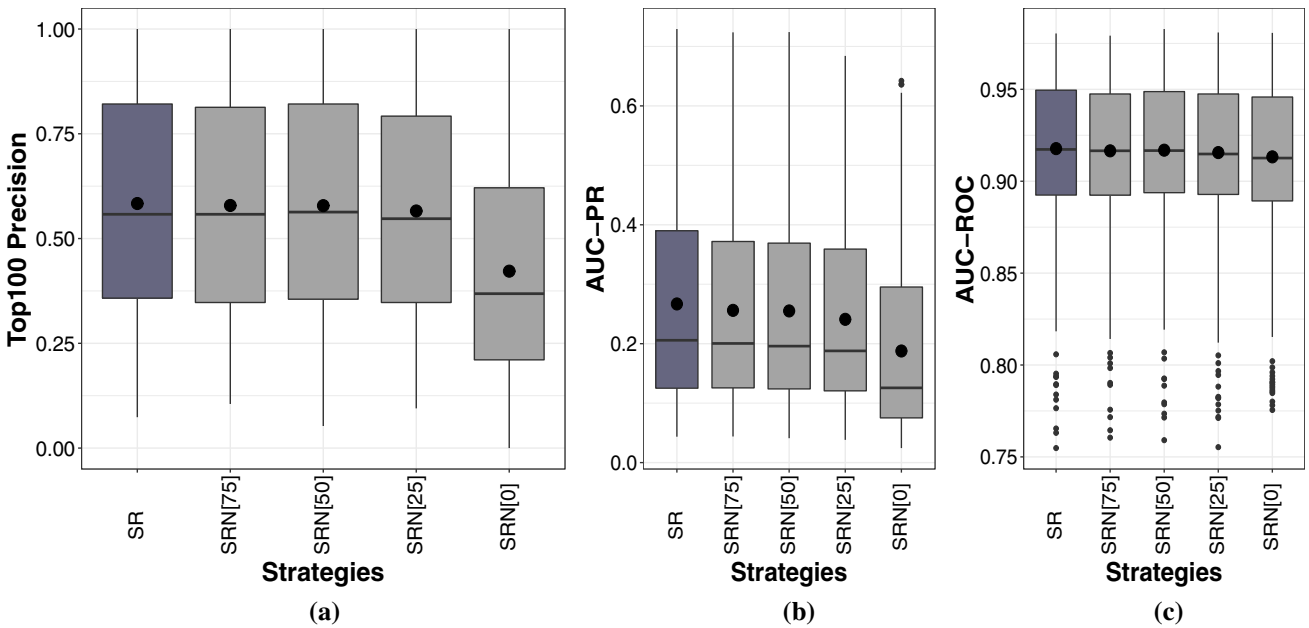

Fig. 10 Transaction-based case study. Comparison of sampling strategies SRN[p] for different p values: percentage of retained negative samples from the feedback set. Note that SR corresponds to SRN[p] with $p = 100$. Comparison in terms of: Top100 precision (**a**), area under the precision–recall curve (**b**) and area under the receiver operator characteristic curve (**c**). Dark boxes indicate the best strategy and those which are not statistically different (paired Wilcoxon test)

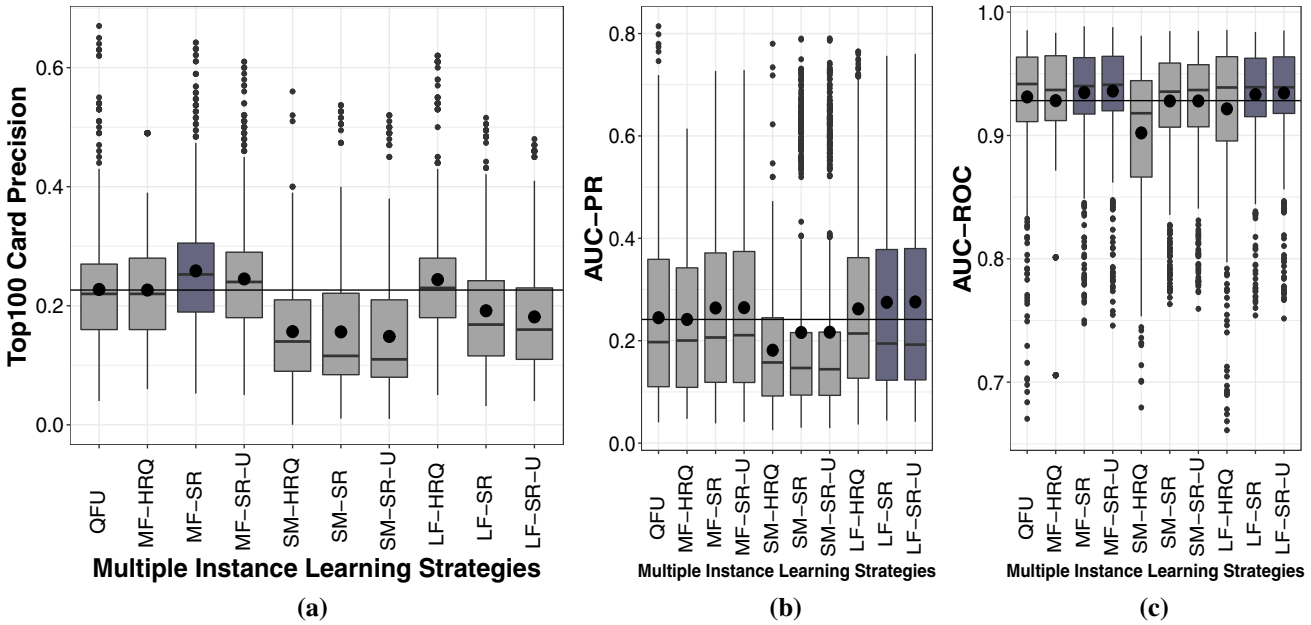

Fig. 11 Card-based case study. Comparison in terms of: Top100 precision (**a**), area under the precision–recall curve (**b**) and area under the receiver operator characteristic curve (**c**). Dark boxes indicate the best strategy and those which are not statistically different (paired Wilcoxon test)

We have also compared the semi-supervised technique SR with the standard random oversample and SMOTE oversampling techniques. As shown in Fig. 8, SR appears to be better in terms of Top100 Precision, AUC-PR and AUC-ROC. ROS and SMOTE outperform HRQ only in terms of AUC-ROC.

In Fig. 10, a comparison between SR and SRN[p] with different values of p is reported (where p is the percentage of genuine transactions retained from the feedback set). SR is statistically better than the SRN[p] strategies in terms of precision, AUC-PR and AUC-ROC and for all $p \in \{75, 50, 25, 0\}$. Our findings are similar to the results of Jacobusse and Veenman [20]. They have used a synthetic dataset with an imbalance ratio of 1%, and they have observed a higher AUC-ROC for the SRN[p] strategy in case of very small p.

In Fig. 12, we report the impact on precision of the proportion of daily transactions selected by SR. The SR experiments reported in Fig. 7 refer to a selection of 1000 points (0.50% of the daily transactions), while the highest precision (statistically significant) is attained for 2% of the daily transactions. The figure shows then the existence of a trade-off in precision between a configuration corresponding to HRQ (i.e., no additional selected points, on the left side) and a configuration with many added points (right side). Our interpretation is that both selection bias (left-side configuration), and excessive imbalance (right-side configuration) may be detrimental to precision. An accurate tuning of such trade-off is expected to be beneficial for the final performance.

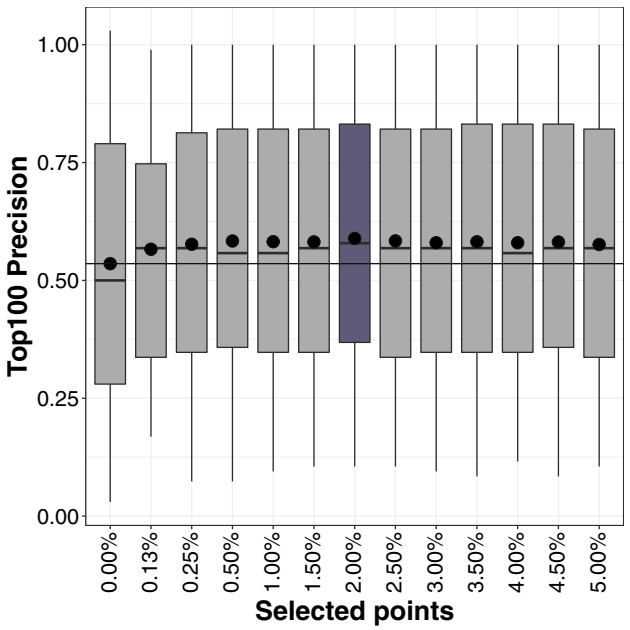

Fig. 12 Transaction-based case study. Top100 precision when varying the number of points selected in SR. The dark box indicates the best strategy (paired Wilcoxon test)

6.2 Card-based fraud detection

In this second case study, we retain the most promising techniques form the transaction assessment (namely SR and SR-U), and we compare them with the multiple instance learning strategies described in Sect. 5.5.1 ($v = 0.05$)

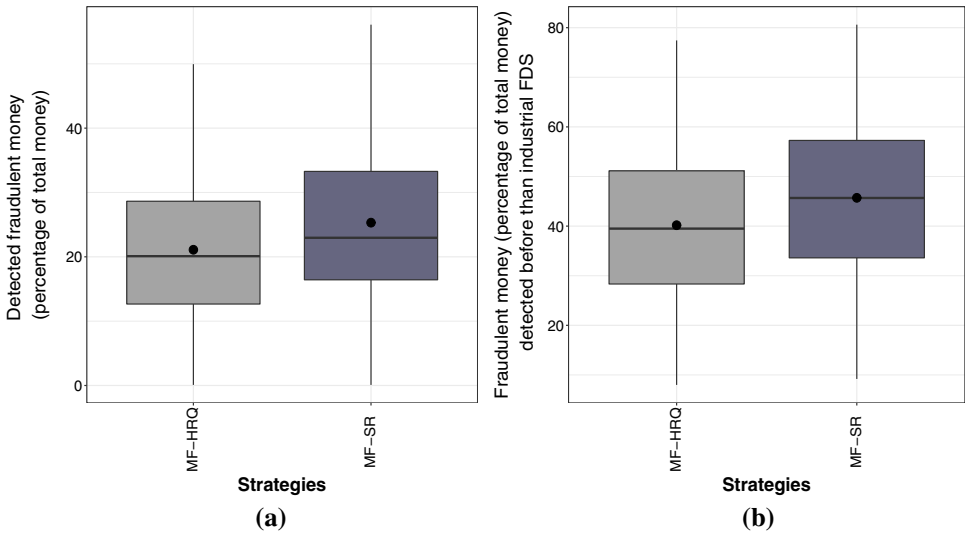

Fig. 13 Card-based case study. Daily amount of detected fraudulent money (percentage of daily transacted money) (**a**). Daily amount of fraudulent money detected before than the industrial FDS (**b**). Dark boxes indicate the significantly better strategy (paired Wilcoxon test)

and 5.5.2 ($\epsilon = 1e-3$). Figure 11 summarizes the results for the card detection study, using Top100 card precision, AUC-PR and AUC-ROC as accuracy metrics.

When the Top100 precision is considered as performance metric, stochastic semi-supervised with random labeling and *max* combining function (MF-SR) returns the best result with a detection improved by 3.21%. The strategy also performs well in terms of AUC-ROC, but is outperformed (by a small margin) by the logarithmic combining function when considering AUC-PR as the performance metric.

Similarly to the results obtained at the transaction level, the best strategies are those combining the baseline highest risk querying with stochastic semi-supervised with random labeling (SR) or uncertainty labeling (SU). The addition of an exploratory part (QFU, or SR-U strategies) did not allow to improve the detection accuracy. The performances are even slightly decreased in terms of Top100 Precision for SR-U strategies.

The results show that the combining function plays an important role. While the *max* and *logarithmic* performed best overall, the *softmax* clearly hampered the fraud detection accuracy.

In Fig. 13, we show the impact on money savings related to the adoption of the MF-HRQ and MF-SR techniques, respectively. Figure 13a reports the daily amount of fraudulent money detected by two approaches (as a percentage of daily transacted money). Figure 13b reports the daily amount of fraudulent money which would have been saved if the AL technique would have been used instead of the FDS implemented by our industrial partner (as a percentage of daily fraudulent transacted money). However, we are not able to measure the amount of fraudulent transactions detected later than our partner since from the historical data we cannot reconstruct if the fraud was indeed detected by the industrial FDS (or labeled after a claim of the cardholder).

To conclude, stochastic semi-supervised labeling (SR and SU) combined with HRQ remained the best strategies, confirming the results obtained at the transaction level. Regarding the combining functions, the *logarithmic* function provided the best performance in terms of AUC-ROC, but the *max* function significantly outperformed in terms of Top100 precision. Overall, the *max* combining function was observed to provide the most stable improvement throughout the range of explored performance metrics.

7 Conclusion and future work

This paper investigates the impact of active learning strategies on the fraud detection accuracy in credit cards. Using a real-world dataset of several millions of transactions over sixty days, we provided an extensive analysis and comparison of different strategies, involving standard active learning, exploratory active learning, semi-supervised learning and combining functions[4]. In particular, we assessed the importance of different selection criteria (supervised, unsupervised or semi-supervised). Moreover, we provided a two-dimensional visualization of the complexity (e.g., non-separability) of the fraud detection problem as well as a visualization of the distribution of the query points issued by the different strategies.

Our results show that unsupervised outlier detection has very low accuracy in the context of credit card fraud detection. This is especially true for a context where only a very small number of cards can be investigated every day and high precision is required.

[4] We made the Streaming Active Learning Strategies repository available in http://github.com/fabriziocarcillo/.

Furthermore, we find that the baseline active learning for fraud detection, the highest risk querying, can be noticeably improved by combining it with stochastic semi-supervised learning, thereby allowing to increase the fraud detection accuracy by up to five percent. Exploratory active learning techniques (including supervised or unsupervised selection criteria) have not improved the fraud detection task. This can be attributed to the highly imbalanced nature of the data and the small exploration budget that can be reasonably allocated in a fraud detection system.

Last, our results on combining functions for bags of transactions showed that the baseline strategy, implemented with the *max* strategy, was the most stable across different accuracy metrics, but that alternative functions could be worth considering.

Future work will focus on two main directions, the stratification of the dataset according to the properties of transactions or cardholders, the design and assessment of hybrid strategies to combine unsupervised (notably outlier detection) and supervised methods.

Acknowledgements The authors FC, YLB and GB acknowledge the funding of the Brufence project (Scalable machine learning for automating defense system) supported by INNOVIRIS (Brussels Institute for the encouragement of scientific research and innovation).

Funding Computational resources have been provided by the Consortium des quipements de Calcul Intensif (CCI), funded by the Fonds de la Recherche Scientifique de Belgique (F.R.S.-FNRS) under Grant No. 2.5020.11.

Compliance with ethical standards

Conflicts of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Aggarwal, C.C.: Outlier analysis. In: Data Mining, pp. 237–263. Springer, New York (2015)
2. Bhattacharyya, S., Jha, S., Tharakunnel, K., Westland, J.C.: Data mining for credit card fraud: a comparative study. Decis. Support Syst. **50**(3), 602–613 (2011)
3. Bolton, R.J., Hand, D.J., et al.: Unsupervised profiling methods for fraud detection. In: Credit Scoring and Credit Control, vol. VII, pp. 235–255 (2001)
4. Breunig, M.M., Kriegel, H.P., Ng, R.T., Sander, J.: Lof: identifying density-based local outliers. In: ACM Sigmod Record, vol. 29, pp. 93–104. ACM (2000)
5. Carcillo, F., Dal Pozzolo, A., Le Borgne, Y.A., Caelen, O., Mazzer, Y., Bontempi, G.: Scarff: a scalable framework for streaming credit card fraud detection with spark. Inf. Fus. **41**, 182–194 (2018)
6. Carcillo, F., Le Borgne, Y.A., Caelen, O., Bontempi, G.: An assessment of streaming active learning strategies for real-life credit card fraud detection. In: DSAA-The 4th IEEE International Conference on Data Science and Advanced Analytics, vol. 7, pp. 783–790 (2017)
7. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: a survey. ACM Comput. Surv. (CSUR) **41**(3), 15 (2009)
8. Chapelle, O., Scholkopf, B., Zien, A.: Semi-supervised learning (Chapelle, o. et al., eds.; 2006) [book reviews]. IEEE Trans. Neural Netw. **20**(3), 542 (2009)
9. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002)
10. Chen, C., Liaw, A., Breiman, L.: Using Random Forest to Learn Imbalanced Data, vol. 110. University of California, Berkeley (2004)
11. Cohn, D., Atlas, L., Ladner, R.: Improving generalization with active learning. Mach. Learn. **15**(2), 201–221 (1994)
12. Dal Pozzolo, A., Caelen, O., Le Borgne, Y.A., Waterschoot, S., Bontempi, G.: Learned lessons in credit card fraud detection from a practitioner perspective. Expert Syst. Appl. **41**(10), 4915–4928 (2014)
13. Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., Bontempi, G.: Credit card fraud detection: a realistic modeling and a novel learning strategy. IEEE Trans. Neural Netw. Learn. Syst. **PP**(99), 1–14 (2017). https://doi.org/10.1109/TNNLS.2017.2736643
14. Dasgupta, S.: Two faces of active learning. Theoret. Comput. Sci. **412**(19), 1767–1781 (2011)
15. Dorronsoro, J.R., Ginel, F., Sgnchez, C., Cruz, C.: Neural fraud detection in credit card operations. IEEE Trans. Neural Netw. **8**(4), 827–834 (1997)
16. Drews, P., Núñez, P., Rocha, R.P., Campos, M., Dias, J.: Novelty detection and segmentation based on gaussian mixture models: a case study in 3d robotic laser mapping. Robot. Auton. Syst. **61**(12), 1696–1709 (2013)
17. Ertekin, S., Huang, J., Bottou, L., Giles, L.: Learning on the border: active learning in imbalanced data classification. In: Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management, pp. 127–136. ACM (2007)
18. Fan, W., Huang, Y.A., Wang, H., Yu, P.S.: Active mining of data streams. In: Proceedings of the 2004 SIAM International Conference on Data Mining, pp. 457–461. SIAM (2004)
19. Ilonen, J., Paalanen, P., Kamarainen, J.K., Kalviainen, H.: Gaussian mixture pdf in one-class classification: computing and utilizing confidence values. In: 18th International Conference on Pattern Recognition, ICPR 2006, vol. 2, pp. 577–580. IEEE (2006)
20. Jacobusse, G., Veenman, C.: On selection bias with imbalanced classes. In: International Conference on Discovery Science, pp. 325–340. Springer, New York (2016)
21. Japkowicz, N., Stephen, S.: The class imbalance problem: a systematic study. Intell. Data Anal. **6**(5), 429–449 (2002)
22. Jurgovsky, J., Granitzer, M., Ziegler, K., Calabretto, S., Portier, P.E., He-Guelton, L., Caelen, O.: Sequence classification for credit-card fraud detection. Expert Syst. Appl. (2018)
23. Kriegel, H.P., Kröger, P., Schubert, E., Zimek, A.: Loop: local outlier probabilities. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, pp. 1649–1652. ACM, New York, NY, USA (2009). https://doi.org/10.1145/1645953.1646195
24. Lewis, D.D., Catlett, J.: Heterogeneous uncertainty sampling for supervised learning. In: In Proceedings of the Eleventh International Conference on Machine Learning, pp. 148–156. Morgan Kaufmann (1994)
25. Lewis, D.D., Gale, W.A.: A sequential algorithm for training text classifiers. In: Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 3–12. Springer, New York (1994)
26. Li, L., Hansman, R.J., Palacios, R., Welsch, R.: Anomaly detection via a Gaussian mixture model for flight operation and safety monitoring. Transp. Res. Part C Emerg. Technol. **64**, 45–57 (2016)

27. Liu, F.T., Ting, K.M., Zhou, Z.H.: Isolation forest. In: Eighth IEEE International Conference on Data Mining. ICDM'08, pp. 413–422. IEEE (2008)
28. Palau, C., Arregui, F., Carlos, M.: Burst detection in water networks using principal component analysis. J. Water Resour. Plan. Manag. **138**(1), 47–54 (2011)
29. Pang, G., Cao, L., Chen, L., Liu, H.: Learning homophily couplings from non-iid data for joint feature selection and noise-resilient outlier detection. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence, pp. 2585–2591. AAAI Press (2017)
30. Pang, G., Cao, L., Chen, L., Liu, H.: Unsupervised feature selection for outlier detection by modelling hierarchical value-feature couplings. In: 2016 IEEE 16th International Conference on Data Mining (ICDM), pp. 410–419. IEEE (2016)
31. Pichara, K., Soto, A., Araneda, A.: Detection of anomalies in large datasets using an active learning scheme based on dirichlet distributions. In: Ibero-American Conference on Artificial Intelligence, pp. 163–172. Springer (2008)
32. Pimentel, M.A., Clifton, D.A., Clifton, L., Tarassenko, L.: A review of novelty detection. Signal Process. **99**, 215–249 (2014)
33. Pinto da Costa, J.F., Alonso, H., Roque, L.: A weighted principal component analysis and its application to gene expression data. IEEE/ACM Trans. Comput. Biol. Bioinform. **8**(1), 246–252 (2011)
34. Ren, D., Wang, B., Perrizo, W.: Rdf: a density-based outlier detection method using vertical data representation. In: Fourth IEEE International Conference on Data Mining, ICDM'04, pp. 503–506. IEEE (2004)
35. Rokach, L.: Decision forest: twenty years of research. Inf. Fus. **27**, 111–125 (2016)
36. Sahin, Y., Bulkan, S., Duman, E.: A cost-sensitive decision tree approach for fraud detection. Expert Syst. Appl. **40**(15), 5916–5923 (2013)
37. Schohn, G., Cohn, D.: Less is more: active learning with support vector machines. In: ICML, pp. 839–846. Citeseer (2000)
38. Schölkopf, B., Williamson, R.C., Smola, A.J., Shawe-Taylor, J., Platt, J.C.: Support vector method for novelty detection. In: Advances in Neural Information Processing Systems, pp. 582–588 (2000)
39. Seeja, KR., Zareapoor, M.: Fraudminer: a novel credit card fraud detection model based on frequent itemset mining. Sci. World J. **2014**, 1–10 (2014)
40. Sethi, N., Gera, A.: A revived survey of various credit card fraud detection techniques. Int. J. Comput. Sci. Mobile Comput. **3**(4), 780–791 (2014)
41. Settles, B., Craven, M., Ray, S.: Multiple-instance active learning. In: Advances in Neural Information Processing Systems, pp. 1289–1296 (2008)
42. Settles, B.: Active learning literature survey. Univ. Wis. Madison **52**(55–66), 11 (2010)
43. Shimpi, P.R., Kadroli, V.: Survey on credit card fraud detection techniques. Int. J. Eng. Comput. Sci. **4**(11), 15010–15015(2015)
44. Shyu, M.L., Chen, S.C., Sarinnapakorn, K., Chang, L.: A novel anomaly detection scheme based on principal component classifier. Technical report, Miami Univ Coral Gables FL Dept of Electrical and Computer Engineering (2003)
45. Srivastava, A., Kundu, A., Sural, S., Majumdar, A.: Credit card fraud detection using hidden Markov model. IEEE Trans. Dependable Secur. Comput. **5**(1), 37–48 (2008)
46. Tang, J., Chen, Z., Fu, A., Cheung, D.: Enhancing effectiveness of outlier detections for low density patterns. Adv. Knowl. Discov. Data Min. 535–548 (2002)
47. Van Vlasselaer, V., Bravo, C., Caelen, O., Eliassi-Rad, T., Akoglu, L., Snoeck, M., Baesens, B.: Apate: a novel approach for automated credit card transaction fraud detection using network-based extensions. Decis. Support Syst. **75**, 38–48 (2015)
48. Van Vlasselaer, V., Eliassi-Rad, T., Akoglu, L., Snoeck, M., Baesens, B.: Afraid: fraud detection via active inference in time-evolving social networks. In: IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 659–666. IEEE (2015)
49. Vijayanarasimhan, S., Jain, P., Grauman, K.: Far-sighted active learning on a budget for image and video recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3035–3042. IEEE (2010)
50. Wang, W., Guan, X., Zhang, X.: A novel intrusion detection method based on principle component analysis in computer security. Adv. Neural Netw. ISNN **2004**, 88–89 (2004)
51. Wei, W., Li, J., Cao, L., Ou, Y., Chen, J.: Effective detection of sophisticated online banking fraud on extremely imbalanced data. World Wide Web **16**(4), 449–475 (2013)
52. Xie, J., Xiong, T.: Stochastic semi-supervised learning on partially labeled imbalanced data. In: Active Learning and Experimental Design workshop In conjunction with AISTATS 2010, pp. 85–98 (2011)
53. Zareapoor, M., Shamsolmoali, P.: Application of credit card fraud detection: Based on bagging ensemble classifier. Proc. Comput. Sci. **48**, 679–685 (2015)
54. Zhang, Y., Bingham, C., Martínez-García, M., Cox, D.: Detection of emerging faults on industrial gas turbines using extended Gaussian mixture models. Int. J. Rotating Mach. **2017**, 1–9 (2017)
55. Zhang, K., Hutter, M., Jin, H.: A new local distance-based outlier detection approach for scattered real-world data. Adv. Knowl. Discov. Data Min. **5476**, 813–822 (2009)
56. Zhu, J., Hovy, E.H.: Active learning for word sense disambiguation with methods for addressing the class imbalance problem. EMNLP-CoNLL **7**, 783–790 (2007)
57. Žliobaite, I., Bifet, A., Pfahringer, B., Holmes, G.: Active learning with evolving streaming data. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 597–612. Springer (2011)