Credit Card Fraud Detection and Concept-Drift Adaptation with Delayed Supervised Information

Andrea Dal Pozzolo, Giacomo Boracchi, Olivier Caelen,
Cesare Alippi, and Gianluca Bontempi

15/07/2015

IEEE IJCNN 2015, Killarney, Ireland

INTRODUCTION

Fraud Detection is notably a challenging problem because of

- concept drift (i.e. customers' habits evolve)
- class unbalance (i.e. genuine transactions far outnumber frauds)
- uncertain class labels (i.e. some frauds are not reported or reported with large delay and few transactions can be timely investigated)

INTRODUCTION II

Fraud-detection systems (FDSs) differ from a classification tasks:

- only a small set of supervised samples is provided by human investigators (they check few alerts).
- the labels of the majority of transactions are available only several days later (after customers have report unauthorized transactions).

Problem formulation

We formalise FD as a classification problem:

- At day t, classifier $\mathcal{K}_{t-1}$ (trained on $t-1$) associates to each feature vector $\mathbf{x} \in \mathbb{R}^n$, a score $P_{\mathcal{K}_{t-1}}(+|\mathbf{x})$.
- The k transactions with largest $P_{\mathcal{K}_{t-1}}(+|\mathbf{x})$ define the alerts A_t reported to the investigators.
- Investigators provide feedbacks F_t about the alerts in A_t, defining a set of k supervised couples $(\mathbf{x}, y)$

$$F_t = \{(\mathbf{x}, y),\ \mathbf{x} \in A_t\}, \tag{1}$$

F_t are the only immediate supervised samples.

Problem formulation II

- At day t, delayed supervised couples $D_{t-\delta}$ are transactions that have not been checked by investigators, but their label is assumed to be correct after that δ days have elapsed.

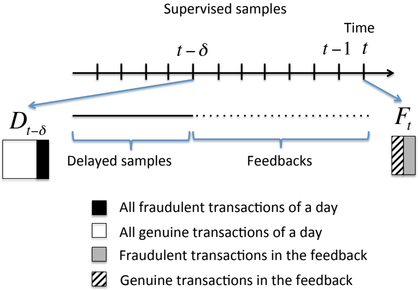

Supervised samples
Time
$t-\delta$
$t-1$
t
$D_{t-\delta}$
F_t
Delayed samples
Feedbacks
■ All fraudulent transactions of a day
□ All genuine transactions of a day
▩ Fraudulent transactions in the feedback
▨ Genuine transactions in the feedback

Figure : The supervised samples available at day t include: i) feedbacks of the first δ days and ii) delayed couples occurred before the δ^{th} day.

- F_t are a small set of risky transactions according the FDS.
- $D_{t-\delta}$ contains all the occurred transactions in a day ($\approx 99\%$ genuine transactions).

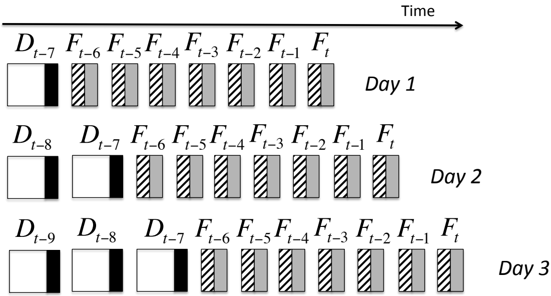

Time
D_{t-7} F_{t-6} F_{t-5} F_{t-4} F_{t-3} F_{t-2} F_{t-1} F_t
Day 1
D_{t-8} D_{t-7} F_{t-6} F_{t-5} F_{t-4} F_{t-3} F_{t-2} F_{t-1} F_t
Day 2
D_{t-9} D_{t-8} D_{t-7} F_{t-6} F_{t-5} F_{t-4} F_{t-3} F_{t-2} F_{t-1} F_t
Day 3

Figure : Everyday we have a new set of feedbacks $(F_t, F_{t-1}, \ldots, F_{t-(\delta-1)})$ from the first δ days and a new set of delayed transactions occurred on the δ^{th} day ($D_{t-\delta}$). In this Figure we assume $\delta = 7$.

ACCURACY MEASURE FOR A FDS

The goal of a FDS is to return accurate alerts, thus the highest precision in A_t. This precision can be measured by the quantity

$$\mathrm{p}_k(t) = \frac{\#\{(\mathbf{x}, y) \in F_t \text{ s.t. } y = +\}}{k} \tag{2}$$

where $\mathrm{p}_k(t)$ is the proportion of frauds in the top k transactions with the highest likelihood of frauds ([1]).

LEARNING STRATEGY

Learning from feedbacks F_t is a different problem than learning from delayed samples in $D_{t-\delta}$:

- F_t provides recent, up-to-date, information while $D_{t-\delta}$ might be already obsolete once it comes.
- Percentage of frauds in F_t and $D_{t-\delta}$ is different.
- Supervised couples in F_t are not independently drawn, but are instead selected by $\mathcal{K}_{t-1}$.
- A classifier trained on F_t learns how to label transactions that are most likely to be fraudulent.

Feedbacks and delayed transactions have to be treated separately.

Concept Drift adaptation

Two conventional solutions for CD adaptation are $\mathcal{W}_t$ and $\mathcal{E}_t$ [6, 5]. To learn separately from feedbacks and delayed transactions we propose $\mathcal{F}_t$, $\mathcal{W}_t^D$ and $\mathcal{E}_t^D$.

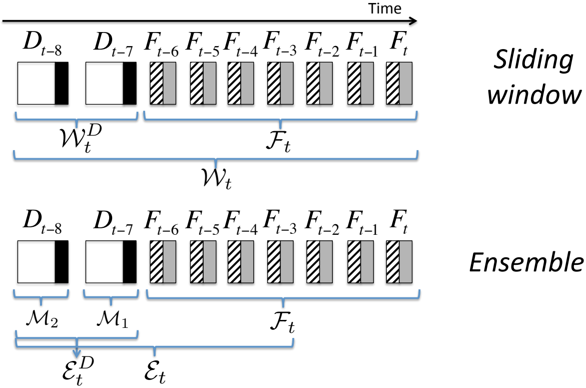


Figure : Supervised information used by different classifiers in the ensemble and sliding window approach.

Classifier aggregations

$\mathcal{W}_t^D$ and $\mathcal{E}_t^D$ have to be aggregated with $\mathcal{F}_t$ to exploit information provided by feedbacks. We combine these classifiers by averaging the posterior probabilities.

Sliding window:

$$P_{\mathcal{A}_t^W}(+|\mathbf{x}) = \frac{P_{\mathcal{F}_t}(+|\mathbf{x}) + P_{\mathcal{W}_t^D}(+|\mathbf{x})}{2}$$

Ensemble:

$$P_{\mathcal{A}_t^E}(+|\mathbf{x}) = \frac{P_{\mathcal{F}_t}(+|\mathbf{x}) + P_{\mathcal{E}_t^D}(+|\mathbf{x})}{2}$$

$\mathcal{A}_t^E$ and $\mathcal{A}_t^W$ give larger influence to feedbacks on the probability estimates w.r.t $\mathcal{E}_t$ and $\mathcal{W}_t$.

Two Random Forest

We used two different Random Forests (RF) classifiers depending on the fraud prevalence in the training set.

- for classifiers on delayed samples we used a Balanced RF [3] (undersampling before training each tree).
- for $\mathcal{F}_t$ we adopted a standard RF [2] (no undersampling).

DATASETS

We considered two datasets of credit card transactions:

Table : Datasets

Id	Start day	End day	# Days	# Instances	# Features	% Fraud
2013	2013-09-05	2014-01-18	136	21,830,330	51	0.19%
2014	2014-08-05	2014-10-09	44	7,619,452	51	0.22%

In the 2013 dataset there is an average of 160k transaction per day and about 304 frauds per day, while in the 2014 dataset there is a daily average of 173k transactions and 380 frauds.

EXPERIMENTS

Settings:

- We assume that after $\delta = 7$ days all the transactions labels are provided (delayed supervised information)
- A budget of $k = 100$ alerts that can be checked by the investigators ($\mathcal{F}_t$ is trained on a window of 700 feedbacks).
- A window of $\alpha = 16$ days is used to train $\mathcal{W}_t^D$ (16 models in $\mathcal{E}_t^D$)

Each experiments is repeated 10 times and the performance is assessed using p_k.

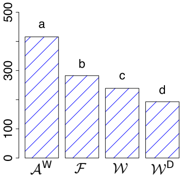
In both 2013 and 2014 datasets, aggregations $\mathcal{A}_t^W$ and $\mathcal{A}_t^E$ outperforms the other FDSs in terms of $\mathtt{p}_k$.

Table : Average $\mathtt{p}_k$ in all the batches for the **sliding window**

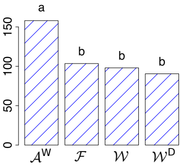
	Dataset 2013		Dataset 2014	
classifier	mean	sd	mean	sd
$\mathcal{F}$	0.609	0.250	0.596	0.249
$\mathcal{W}^D$	0.540	0.227	0.549	0.253
$\mathcal{W}$	0.563	0.233	0.559	0.256
$\mathcal{A}^W$	0.697	0.212	0.657	0.236

Table : Average $\mathtt{p}_k$ in all the batches for the **ensemble**

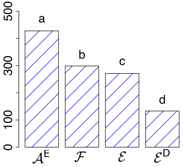
	Dataset 2013		Dataset 2014	
classifier	mean	sd	mean	sd
$\mathcal{F}$	0.603	0.258	0.596	0.271
$\mathcal{E}^D$	0.459	0.237	0.443	0.242
$\mathcal{E}$	0.555	0.239	0.516	0.252
$\mathcal{A}^E$	0.683	0.220	0.634	0.239



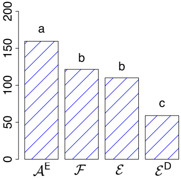
(a) **Sliding window** 2013



(b) **Sliding window** 2014



(c) **Ensemble** 2013



(d) **Ensemble** 2014

Sum of ranks from the Friedman test [4], classifiers having the same letter are not significantly different (paired t-test based upon on the ranks).

EXPERIMENTS ON ARTIFICIAL DATASET WITH CD

In the second part we artificially introduce CD in specific days by juxtaposing transactions acquired in different times of the year.

Table : Datasets with Artificially Introduced CD

Id	Start 2013	End 2013	Start 2014	End 2014
CD1	2013-09-05	2013-09-30	2014-08-05	2014-08-31
CD2	2013-10-01	2013-10-31	2014-09-01	2014-09-30
CD3	2013-11-01	2013-11-30	2014-08-05	2014-08-31

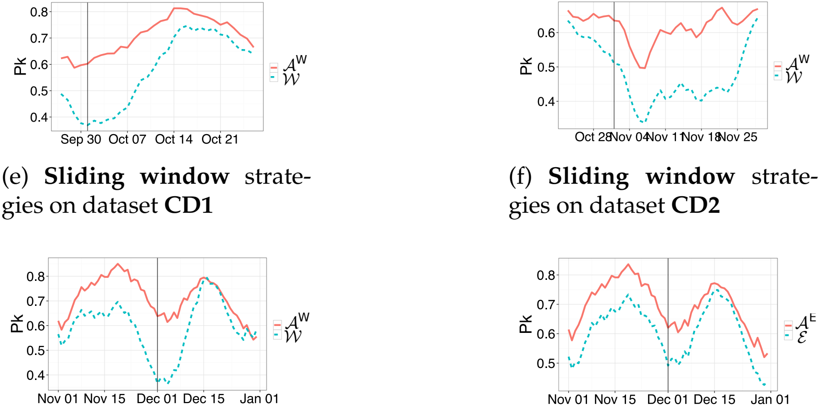
Table : Average p_k in the month before and after CD for the **sliding window** approach

(a) **Before** CD

	CD1		CD2		CD3	
classifier	mean	sd	mean	sd	mean	sd
$\mathcal{F}$	0.411	0.142	0.754	0.270	0.690	0.252
$\mathcal{W}^D$	0.291	0.129	0.757	0.265	0.622	0.228
$\mathcal{W}$	0.332	0.215	0.758	0.261	0.640	0.227
$\mathcal{A}^W$	0.598	0.192	0.788	0.261	0.768	0.221

(b) **After** CD

	CD1		CD2		CD3	
classifier	mean	sd	mean	sd	mean	sd
$\mathcal{F}$	0.635	0.279	0.511	0.224	0.599	0.271
$\mathcal{W}^D$	0.536	0.335	0.374	0.218	0.515	0.331
$\mathcal{W}$	0.570	0.309	0.391	0.213	0.546	0.319
$\mathcal{A}^W$	0.714	0.250	0.594	0.210	0.675	0.244



(e) **Sliding window** strategies on dataset **CD1**

(f) **Sliding window** strategies on dataset **CD2**

(g) **Sliding window** strategies on dataset **CD3**

(h) **Ensemble** strategies on dataset **CD3**

Figure : Average p_k per day (the higher the better) for classifiers on datasets with artificial concept drift smoothed using moving average of 15 days. The vertical bar denotes the date of the concept drift.

Concluding remarks

We notice that:

- $\mathcal{F}_t$ outperforms classifiers on delayed samples (trained on obsolete couples).
- $\mathcal{F}_t$ outperforms classifiers trained on the entire supervised dataset (dominated by delayed samples).
- Aggregation gives larger influence to feedbacks.

CONCLUSION

- We formalise a real-world FDS framework that meets realistic working conditions.
- In a real-world scenario, there is a strong alert-feedback interaction that has to be explicitly considered
- Feedbacks and delayed samples should be separately handled when training a FDS
- Aggregating two distinct classifiers is an effective strategy and that it enables a prompter adaptation in concept drifting environments

FUTURE WORK

Future work will focus on:

- Adaptive aggregation of $\mathcal{F}_t$ and the classifier trained on delayed samples.
- Study the sample selection bias in $\mathcal{F}_t$ introduced by alert-feedback interaction.

BIBLIOGRAPHY

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

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

[3] C. Chen, A. Liaw, and L. Breiman.
Using random forest to learn imbalanced data.
University of California, Berkeley, 2004.

[4] M. Friedman.
The use of ranks to avoid the assumption of normality implicit in the analysis of variance.
Journal of the American Statistical Association, 32(200):675–701, 1937.

[5] J. Gao, B. Ding, W. Fan, J. Han, and P. S. Yu.
Classifying data streams with skewed class distributions and concept drifts.
Internet Computing, 12(6):37–49, 2008.

[6] D. K. Tasoulis, N. M. Adams, and D. J. Hand.
Unsupervised clustering in streaming data.
In *ICDM Workshops*, pages 638–642, 2006.