ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1.
Available from http://www.i6doc.com/en/.

Anomaly detection and characterization in smart card logs using NMF and Tweets

Emeric Tonnelier and Nicolas Baskiotis and Vincent Guigue and Patrick Gallinari

UPMC - Sorbonne Universités - LIP6 - CNRS
4 place Jussieu, 75005 Paris

Abstract. This article describes a novel approach to detect anomalies in smart card logs. In this study, we chose to work on a 24h base for every station in the Parisian metro network. We also consider separately the 7 days of the week. We first build a robust averaged reference for (day,station) couples and then, we focus on the difference between particular situations and references. All experiments are conducted both on the raw data and using an NMF denoised approximation of the log flow. We demonstrate the interest and the robustness of the latter strategy. Then we mine RATP[1] Twitter account to obtain ground truth information about operating incidents. This synchronized flow is used to evaluate our models.

1 Introduction

Understanding, predicting and characterizing transportation network failures is critical to improve the whole system. Decision makers have to rely on strong indicators to pursue coherent development policy. Until recently, most information came from expert knowledge and population surveys.

Smart cards change the situation: we get the opportunity to obtain massive accurate data and to follow users. Several references illustrate how we can detect specific events in log flows (e.g. snowy days) [1], understand congestion [2], characterize users habits [3] or predict individual trip [4]. Exploiting log flows enables to catch habits on a mid/long term basis, it provides a supervision for prediction tasks and it gives a new view on service quality and customer satisfaction [5].

This article tackles anomaly detection in smart card logs. The first idea consists in building a robust averaged reference and then to consider distant situation as abnormal. Focusing on outliers is common in transportation data mining; For instance, [6] mines infrequent patterns to detect pickpockets. Early studies also rely on averaged models [1]. Once established this strong baseline, we propose to reconsider the situation from another view; we propose to use Non Negative Matrix Factorization (NMF) to detect anomalies in a latent space. Such representation learning algorithms have been used successfully to tackle habit extraction and user clustering [3]. We are going to demonstrate their robustness for failure detection.

We work on a 24h base, considering every station separately. NMF is known as a source separation algorithm; it enables us to decompose the original (day, station) couple into a new representation space while removing most of the signal

[1]Parisian transport authority

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1.
Available from http://www.i6doc.com/en/.

noise[7]. The 24 atoms of the small dictionary are learned during the training step, over the whole dataset. In order to enforce their interpretability, we add a mono-modal constraint in the algorithm; combined with a regularized framework, we obtain compact atoms with few overlapping, each one describing a specific part of the day. The general paradigm remains unchanged: abnormal situations correspond to the furthest points with respect to the averaged reference, but in the latent space [8].

The major issue resides in the evaluation framework: the lack of supervision regarding network failures is critical for us. It prevents any rigorous comparison between approaches. We crawled the RATP[1] twitter account which describes most operating incidents. We propose to use this piece of information as a distant evaluation [9]. We do not perform an explicit matching between detected incident and ground truth but we propose ROC based metrics to measure correlations between the model and twitter alarms.

First, we define notations and models in depth; the next step consists in an extensive experimental part describing both smart card and twitter dataset, the distant evaluation metrics and the comparison of our models in various contexts.

2 Models and notations

Our study covers the last 3 months of 2015 (91 days, 13 weeks) for 300 metro stations and we choose a discretization step of 1 minute (ie, $T = 1440$ intervals for 24h). We assume that every station and day has a specific behavior; as far as smart cards are concerned, we consider $N = 91 \times 300 = 27300$ objects $\mathbf{x}_{s,i} \in \mathbb{N}^T$ with $T = 1440$. Each cell $x_{s,i}(t)$ corresponds to the number of entry-logs in the period[2]. We denote by $X \in \mathbb{N}^{N \times T}$ the matrix gathering all $\mathbf{x}_{s,i}$.

2.1 Baseline (BL) & Normalized baseline (NBL)

Anomaly detection algorithms for time series are based generally on a distance to a regular regime [1]. We consider in the following a week periodicity, i.e. we assume that the station activity has a similar behavior every Sunday, Monday,... The objective in this context is to learn a reference model per couple (station,day of the week). We compute $\bar{\mathbf{x}}_{s,d} \in \mathbb{N}^T = \frac{1}{13}\sum_i \mathbf{x}_{s,i\times 7+d}$ for $d \in \{1,\dots,7\}$ to learn $N' = 7 \times 300 = 2100$ averaged references corresponding to the days of the week. We denote by $\bar{X} \in \mathbb{R}^{N' \times T}$ the matrix gathering all references. Then, we define a anomaly score function based on the L_1 distance between a couple and its associated reference: $\text{score}(s,i) = \sum_t |x_{s,i}(t) - \bar{x}_{s,d_i}(t)|$.

It seems clear that such a modeling is suitable to particular calendar day detection like bank holiday. In order to detect fine-grained anomalies, we propose a second baseline relying on a normalized version of the $\mathbf{x}$. Thus, we define $\mathbf{x}^\dagger_{s,i} = \mathbf{x}_{s,i}/||\mathbf{x}_{s,i}||_1$ and their associated normalized references $\bar{\mathbf{x}}^\dagger_{s,d} = \frac{1}{13}\sum_i \mathbf{x}^\dagger_{s,i\times 7+d}$. The anomaly score is computed as previously.

[2]The Parisian metro network is equipped with a tap-in smart card system; exits are not logged.

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1.
Available from http://www.i6doc.com/en/.

2.2 Nonnegative Matrix Factorization (NMF)

Our goal is to provide a more robust, efficient and understandable representation of the behavior of the network. The main assumption about the generic behavior of a station is that it can be separated into few weighted standard patterns. Thus, we explore a state of the art robust source separation algorithm adapted to our nonnegative dataset: the NMF [7]. With NMF, we work exclusively on normalized data: we aim at modeling habits, not at detecting days with clear power drops. The idea consists in learning both a dictionary $D \in \mathbb{R}^{Z \times T}$ made of Z atoms $\mathbf{a}_z \in \mathbb{R}^T$ and the associated reconstruction code matrix $W \in \mathbb{R}^{N \times Z}$ so as to obtain $\mathbf{x}^\dagger_{s,i} \approx \sum_z w_{s,i}(z)\mathbf{a}_z$, where $\mathbf{w}_{s,i} \in \mathbb{R}^Z$ is the weight vector associated to $\mathbf{x}^\dagger_{s,i}$. The general formulation of the regularized learning problem is the following one: $\text{argmin}_{W,D} \left\|X^\dagger - WD\right\|_F + \lambda_W \left\|W\right\|_F$.

In order to enforce efficiency, robustness and make atoms more understandable, we introduce slight modifications in the original NMF. First, we divide the learning process into two steps: 1) $\bar{W}$ and D are learned on the reference matrix $\bar{X}^\dagger$ so as to obtain robust atoms quickly. 2) Once the dictionary D fixed, W is learned by considering N independent reconstruction problems corresponding to the $\mathbf{x}^\dagger_{s,i}$ using the $\bar{\mathbf{x}}_{s,d_i}$ representation as initialization to enforce the use of same atoms for same days. Our second proposal consists in a mono-modal constraint added on the atoms; doing this, we enforce every atom to have a single maximum. As a consequence, each atom focuses on a specific compact part of the day (cf Fig 1). In practice, we introduce a smoothing procedure preserving only the strongest maximum for each atom in the gradient descent algorithm.

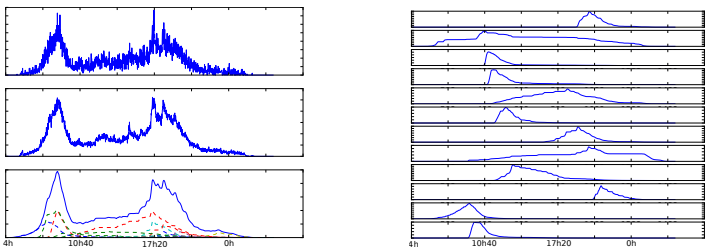


Fig. 1: [left] (a) 09/10/2015 for station *Marcel Sembat*, (b) Averaged Wednesday model for *Marcel Sembat*, (c) NMF reconstruction of the first distribution. [right] Mono-modal atom examples extracted from the dictionary.

The anomaly detection approach is based on the symmetrised KL divergence; indeed, every $\mathbf{w}$ is a distribution (due to the normalization constraint in the NMF procedure). Thus, the anomaly score becomes: $\text{score}(s,i) = \sum_z w_{s,i}(z) \log \frac{w_{s,i}(z)}{\bar{w}_{s,d_i}(z)} + \bar{w}_{s,d_i}(z) \log \frac{\bar{w}_{s,d_i}(z)}{w_{s,i}(z)}$.

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1.
Available from http://www.i6doc.com/en/.

3 Experiments

We conduct our series of experiments exploiting two synchronized datasets. The smart card one counts 520 millions logs made of a time stamp and a location (station). In parallel, we crawl and process a Tweet corpus from the RATP account that gives us information about 255 operating incidents. From the raw material, we extract a time stamp, a duration and a metro line; namely we get time and location characterizations for the incidents. We consider the duration as a strength indicator : it will enable us to compare ranked ground truth with ranked detections (according to the score function defined in sec 2).

Three kinds of anomalies impact the log data. The first one regards *sensors failures*, when no logs are recorded for a station whereas it still works normally. Then, we come to the *operating incidents*. Obviously, the impact of those anomalies depends on the severity of the incident. Finally, anomalies can be induced by a specific context (bank holiday, special event, ...). Unfortunately, those cases overlap. For instance, strong operating incidents correspond to a total interruption of the service of one or several stations; thus, a zero signal is observed in the log data, as for sensor failures. Conversely, slowdowns of the traffic tend to be invisible in the log flow.

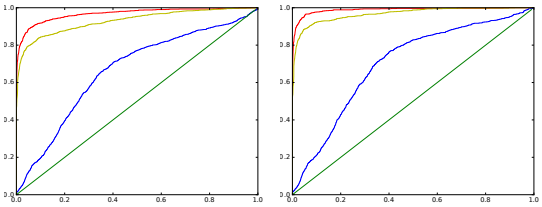
As a consequence, the evaluation difficulty resides both in the lack of supervision and in the heterogeneity of the impacts on the signal. We present two series of experiments to understand the behavior of our models and compare their performances. The first one tackles vanishing signals: we have to know if our models detect those events and what part of the alarms are related to this case. The second one consists in a distant evaluation of operating incident detection studying the correlation between detected anomalies and crawled twitter incidents.

3.1 Vanishing signal detection

Vanishing events may be labeled easily by identifying time windows with a zero signal for each (station, day) couple. To study the ability of our model and baselines to detect vanishing signals, we propose an evaluation close to the bipartite ranking framework [10]. Thus, we compute the following ROC curve: each point corresponds to a threshold α on the anomaly score. Namely, the $\alpha\%$ top ranked couples according to our model are considered as positive - presenting an anomaly - and the rest as negative; For each α, we plot the true positive rate (percentage of positive couples labeled as positive) w.r.t. the false positive rate (percentage of negative couples labeled as positive). The area under the curve gives the overall performance of the model. Fig. 2(a) shows the ROC curves for a minimal vanishing intervals greater than 30 min (left) and 1 hour (right).

The standard baseline (BL) alarms are centered on atypical days and the model is not able to catch vanishing signals efficiently. Then, NMF outperforms the normalize baseline (NBL), especially on shorter interruptions. More than 70% of anomalies due to a vanishing signal of more than 30 minutes are detected at very first ranks for the NMF model compare to the 50% for the NBL. Generally

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1. Available from http://www.i6doc.com/en/.



(a) ROC curves for a minimal vanishing intervals of 30min (left) and 1 hour (right). BL = blue; NBL = yellow; NMF = red.

	NMF	BL	NBL
Geogr.	**73.8**%	72.4%	51.1%
Temp.	60.7%	63.1%	**70.6**%

(b) Accuracies of the 3 models to determine the time and the metro line of an incident in the Twitter dataset.

Fig. 2:

speaking, the denoising ability of the NMF and the sparse representation helps the NMF model to distinguish between irregularities due to the noise and real signal anomalies.

3.2 Distant evaluation with Twitter

We now come to the most important results of the article. We compare our detectors with the Twitter ground truth according to the same metrics as previously. At first, results are aggregated by days, over the whole network. Then we conduct two specific analysis to determine the time location in the day and the metro line where the incident occurs.

The aggregated results, at the couple level are presented in Fig. 3. We demonstrate that our model is always above the two baselines. Namely, our ranking of the incident (in the latent space) matches the most abnormal days according to Twitter (taking into account the length of each incident). We also measure the percentage of common detection w.r.t. the threshold of the detectors (Fig. 3, right). Surprisingly, the top ranked anomalies are different between NMF and BL/NBL. Once passed the 20% more powerful anomalies, NMF and NBL have between 60% and 80% of common detection. As earlier, BL focuses on different days.

Table 2(b) illustrates the results of temporal and geographical projections. Regarding the evaluation, the process is reversed: we simply try to find anomalies from our detectors that explain the tweets. If the anomaly occurs during the tweet incident, then the tweet is temporally explained, if the anomaly is on the same line as the tweet, it is geographically explained. NMF performs well on geographical aspects but it is overcome on the temporal benchmark. Indeed, in order to determine the time location of the incident, we select the peak of the atom with the highest variation w.r.t. to the reference: given the support of the atom, we have a too large approximation.

4 Conclusion and Perspectives

We propose a novel NMF based approach to model smart card logs. It is robust and very compact and we demonstrate its ability to catch anomalies in noisy

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1.
Available from http://www.i6doc.com/en/.

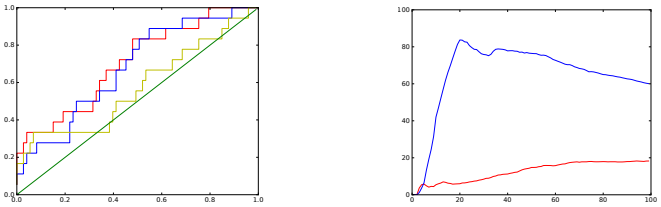


Fig. 3: (left) ROC curves of abnormal day detections, blue: BL; yellow: NBL; red: NMF. (right) % of common detection w.r.t. the threshold of the detectors, red: BL *vs* NMF, blue: NBL *vs* NMF.

signals. The modified NMF is very fast and its only weakness resides in the temporal location of the anomaly. The dictionary is made of few atoms providing a high compression rate of the data but it reduces the temporal accuracy. We also introduce an original distant evaluation scheme that enables us to show quantitative results on an unsupervised task. The perspectives around this work concern the anomaly detection at the user scale, to obtain a finer-grain detector and understand how people react when they face abnormal situations.

Acknowledgment: The authors gratefully acknowledge the STIF regarding the exploitation of the dataset. This work was partially founded by the FUI AWACS grant.

References

[1] M. Trépanier, C. Morency, B. Agard, E. Descoimps, and J.S. Marcotte. Using smart card data to assess the impacts of weather on public transport user behavior. In *Conference on Advanced Systems for Public Transport*, 2012.

[2] I. Ceapa, C. Smith, and L. Capra. Avoiding the crowds: understanding tube station congestion patterns from trip data. In *ACM KDD workshop on urban computing*, 2012.

[3] M. Poussevin, E. Tonnelier, N. Baskiotis, V. Guigue, and P. Gallinari. Mining ticketing logs for usage characterization with nonnegative matrix factorization. *LNCS Big Data Analytics in the Social and Ubiquitous Context*, 2016.

[4] S. Foell, G. Kortuem, R. Rawassizadeh, S. Phithakkitnukoon, M. Veloso, and C. Bento. Mining temporal patterns of transport behaviour for predicting future transport usage. In *Conference on Pervasive and ubiquitous computing adjunct publication*, 2013.

[5] T. Camacho, M. Foth, and A. Rakotonirainy. Pervasive technology and public transport: Opportunities beyond telematics. *IEEE Pervasive Computing*, 12(1), 2012.

[6] B. Du, C. Liu, W. Zhou, Z. Hou, and H. Xiong. Catch me if you can: Detecting pickpocket suspects from large-scale transit records. In *KDD*. ACM, 2013.

[7] P. O. Hoyer. Non-negative matrix factorization with sparseness constraints. *JMLR*, 2004.

[8] L. Xiong, X. Chen, and J. Schneider. Direct robust matrix factorization for anomaly detection. In *International Conference on Data Mining*. IEEE, 2011.

[9] M. Mintz, S. Bills, R. Snow, and D. Jurafsky. Distant supervision for relation extraction without labeled data. In *Joint Conference on Natural Language Processing*. ACL, 2009.

[10] S. Clémençon and N. Vayatis. Adaptive estimation of the optimal roc curve and a bipartite ranking algorithm. In *Algorithmic Learning Theory*. Springer, 2009.