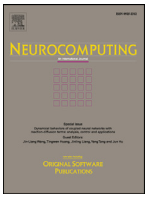
Neurocomputing 000 (2018) 1–13

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



Anomaly detection in smart card logs and distant evaluation with Twitter: a robust framework

E. Tonnelier[a,*], N. Baskiotis[a], V. Guigue[a], P. Gallinari[a]

[a] *UPMC - Sorbonne Universités - LIP6 - CNRS, 4 place Jussieu, Paris 75005, France*

ARTICLE INFO

Article history:
Received 9 July 2017
Revised 20 November 2017
Accepted 27 December 2017
Available online xxx

Keywords:
Anomaly detection
NMF
Machine learning
Smart card logs
Transportation network analysis
Twitter evaluation

ABSTRACT

Smart card logs constitute a valuable source of information to model a public transportation network and characterize normal or abnormal events; however, this source of data is associated to a high level of noise and missing data, thus, it requires robust analysis tools. First, we define an anomaly as any perturbation in the transportation network with respect to a typical day: temporary interruption, intermittent habit shifts, closed stations, unusual high/low number of entrances in a station. The Parisian metro network with 300 stations and millions of daily trips is considered as a case study. In this paper, we present four approaches for the task of anomaly detection in a transportation network using smart card logs. The first three approaches involve the inference of a daily temporal prototype of each metro station and the use of a distance denoting the compatibility of a particular day and its inferred prototype. We introduce two simple and strong baselines relying on a differential modeling between stations and prototypes in the raw-log space. We implemented a raw version (sensitive to volume change) as well as a normalized version (sensitive to behavior changes). The third approach is an original matrix factorization algorithm that computes a dictionary of typical behaviors shared across stations and the corresponding weights allowing the reconstruction of denoised station profiles. We propose to measure the distance between stations and prototypes directly in the latent space. The main advantage resides in its compactness allowing to describe each station profile and the inherent variability within a few parameters. The last approach is a user-based model in which abnormal behaviors are first detected for each user at the log level and then aggregated spatially and temporally; as a consequence, this approach is heavier and requires to follow users, at the opposite of the previous ones that operate on anonymous log data. On top of that, our contribution regards the evaluation framework: we listed particular days but we also mined RATP[1] Twitter account to obtain (partial) ground truth information about operating incidents. Experiments show that matrix factorization is very robust in various situations while the last user-based model is particularly efficient to detect small incidents reported in the twitter dataset.

© 2018 Published by Elsevier B.V.

1. Introduction

Transportation networks have become a crucial urbanization planning tool: in dense urban areas, most people rely only on public transportation system to move, to go to work, to visit friends or for entertainment trips. Understanding, predicting and characterizing transportation network failures is critical to improve the whole system and provide a reliable service. It has been shown that a good transportation system increases public health, boosts the economy, saves space and time inside a city [1]. Decision makers have to rely on strong indicators to pursue coherent development policies. Until last decade, expert knowledge and population surveys were the usual and most accurate techniques to apprehend the behavior of a transportation network [2]. Smart cards revolutionized the field with the opportunity to obtain massive accurate data on the user's mobility and, thus, to analyze the use of a transportation network. Research exploiting this data has been conducted on numerous applications, such as detection of meteorological events [3], prediction of congestion [4], characterization of users habits [5] and prediction of individual trips [6]. Exploiting log flows enables to catch habits on a mid/long-term basis, it provides strong dynamic mobility flow information and it gives a new view on service quality and customer satisfaction [7]. However, log

[*] Corresponding author.
E-mail addresses: emeric.tonnelier@lip6.fr (E. Tonnelier), nicolas.baskiotis@lip6.fr (N. Baskiotis), vincent.guigue@lip6.fr (V. Guigue), patrick.gallinari@lip6.fr (P. Gallinari).

[1] Parisian transport authority

https://doi.org/10.1016/j.neucom.2017.12.067
0925-2312/© 2018 Published by Elsevier B.V.

flows often contain a high level of noise and missing data[2], they occur on a complex graph structure and have multi-scale aspects (the density of logs highly depends on the time of the day and the overall number of logs varies from one station to another).

This article focuses on the task of anomaly detection on smart card logs in an unsupervised setting, i.e., without having any knowledge on the time periods of anomalies. Anomaly detection is a widely studied topic in several application domains [8], for instances in computer security [9], fraud detection [10], ...From a machine learning perspective, there are two main contexts depending on the availability of labels indicating the anomalies. If labels are available, usual supervised machine learning algorithms are relevant to extract meaningful patterns from the data (Neural Networks [11], Shapelets-type [12]). When no supervision is available -like in our application-, methods are generally based on clustering algorithms to identify outliers (K-Means [13], Density-based clustering [14], KNN [15]).

We define an anomaly as a spatiotemporal event -attached to a given station and during a certain time window- corresponding to the deviation of the station activity from its regular model. We focus on short anomalies occurring inside a usual day; thus, the proposed approaches will use daily reference models describing the station behavior during 24 hours; we simply separate the seven days of the week for every station. According to [3], the weekly regularity hypothesis is relevant at the season scale and we mainly work on a 3 months dataset. In this study, four different approaches are explored to detect network anomalies. The first three proposals rely on averaged models for every day-station; indeed, log flows are very noisy and aggregating several samples corresponding to a same phenomenon is a standard noise reduction strategy. The first two proposed approaches naively use those daily averaged models as prototypes for each couple of station and day of the week. Anomalies are detected by using a threshold on the L_1 distance between a given raw station daily log and the corresponding prototype. Two variants are considered: the first one using the raw number of check-in while the second one using the normalized signal, less subject to volumetric differences. Those two first approaches can be seen as a strong kernel method which tackles anomaly detection as outliers identification [16]. The third approach relies on a Non-Negative Matrix Factorization (NMF) algorithm to learn a denoised and compact representation of the daily temporal station profiles. NMF algorithms learn simultaneously the atom dictionary and the parsimonious weights. Both are used to reconstruct the original signal, thus, encouraging the emergence of atoms shared across a number of profiles and avoiding overfitting. Our contribution consists in an NMF improvement that enforces time consistency during decomposition and increases the robustness of the latent representation. Then, an anomaly detection procedure operating directly in the latent space is proposed; both (station, day) couples and references are mapped in the latent space and compared using a L_1 distance. The fourth proposed approach is based on user modeling. Anomalies are measured for each log based on an individual spatial model. The problem is turned into an anomaly aggregation problem: detecting real events in an anomaly flow. We perform a temporal convolution on the station graph that tackles this issue. This heavier but more accurate modeling has advantages and drawbacks which will be discussed in the experimental part. Note that the first three approaches can be applied with few information, as only the entrance count for each station is required, where the fourth approach requires user identification and the history of the users for a large time period to establish accurate individual models.

A last contribution of this work resides in the evaluation framework: the lack of supervision regarding network failures is critical to evaluate the proposed approach. To tackle this classical issue and to apprehend the advantages of each approach, three different protocols are proposed: the first one compares the ability of the approaches to detect *vanishing signals*, i.e., when no entrances are recorded for a small period of time; the second protocol explores qualitative results comparing the detected anomalies to a list of all particular days in the studied period (Christmas, Parisian terrorist attacks, Conference Of Parties -COP 21-, ...). A quantitative study is presented using the messages emitted by the official Parisian transport authority (RATP) twitter account which reports many operating incidents. Those messages are preprocessed and used as a distant evaluation [17]. Finally, we propose a series of experiments on a toy dataset. This approach allows us to study our proposals strengths and weaknesses on a controlled environment.

The paper is organized as follows: Section 2 presents the context of the proposed work and usual anomaly detection approaches in Machine Learning; Section 3 introduces notations and the proposed models; Section 4 describes the protocol settings, the used datasets and the experiments conducted to assess the performances of the proposed models.

2. Related work

We first list interesting applications relying on smart card logs before focusing on anomaly detection. Finally, we propose a short review on nonnegative matrix factorization, demonstrating its interest for log analysis.

2.1. Smart card logs

Historically, ground survey was the only tool to monitor the use of a transportation system. But surveys have three main caveats: they are expensive, especially in the case of a wide transportation network covering a large urban space; due to the cost, they are rarely conducted and thus, are not able to reflect quick shifts of habits, temporary phenomena, or the changing dynamic of the network; at last, they are able to capture frequent recurrent habits efficiently but lack of precision for occasional trips due to the limits of the statistical tools. Since last decades, the growing use of smart cards in transportation networks offers a rich alternative to characterize transportation networks by exploiting the log data of the cards indicating when and where card owners use the network [18]. However, as other data coming from sensors (individual GPS traces, road sensors for traffic analysis, [4,19]), smart card logs are massive, noisy and incomplete. Some category of users are not equipped with cards: tourist or occasional visitors may use tickets which are not included in those data. Sensor failures and fraudsters are other common sources of noise. Moreover, log exploitation to construct accurate models is hard due to the randomness in the user's behavior and the lack of semantic information in the log : we know neither the reason of a trip, nor the destination[3]. Those difficulties have been first tackled by engineering robust statistical features designed with domain experts to demonstrate the relevance of log data to perform classical tasks as Origin/Destination matrices modeling the public urban transportation dynamic [20]. A common approach consists to aggregate data spatially, temporally and/or by groups of users to exhibit robust profiles: for example, [21] extracts various indicators (entrance time, number of bus stops use...) to model the use of Gatineau's (Quebec) transportation system. Such models with handcrafted features are competitive with ground survey techniques but requires

[2] STIF Parisian authority estimates that about 30% of data are missing.

[3] Parisian system is tap-in only, exits are not logged.

Please cite this article as: E. Tonnelier et al., Anomaly detection in smart card logs and distant evaluation with Twitter: a robust framework, Neurocomputing (2018), https://doi.org/10.1016/j.neucom.2017.12.067

Download English Version:

https://daneshyari.com/en/article/6863924

Download Persian Version:

https://daneshyari.com/article/6863924

Daneshyari.com