Anomaly Detection and Localisation using Mixed Graphical Models
Romain Laby, François Roueff, Alexandre Gramfort

To cite this version:
Romain Laby, François Roueff, Alexandre Gramfort. Anomaly Detection and Localisation using Mixed Graphical Models. ICML 2016 Anomaly Detection Workshop, Jun 2016, New York, United States. <hal-01347167>

HAL Id: hal-01347167
https://hal-imt.archives-ouvertes.fr/hal-01347167
Submitted on 20 Jul 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Anomaly Detection and Localisation using Mixed Graphical Models

Romain Laby
CNRS LTCI, Télécom ParisTech, Université Paris-Saclay, 46 Rue Barrault, 75013 Paris
Thales Airborne Systems, 2 Avenue Gay Lussac, 78990 Élancourt

François Roueff
CNRS LTCI, Télécom ParisTech, Université Paris-Saclay, 46 Rue Barrault, 75013 Paris

Alexandre Gramfort
CNRS LTCI, Télécom ParisTech, Université Paris-Saclay, 46 Rue Barrault, 75013 Paris

Abstract

We propose a method that performs anomaly detection and localisation within heterogeneous data using a pairwise undirected mixed graphical model. The data are a mixture of categorical and quantitative variables, and the model is learned over a dataset that is supposed not to contain any anomaly. We then use the model over temporal data, potentially a data stream, using a version of the two-sided CUSUM algorithm. The proposed decision statistic is based on a conditional likelihood ratio computed for each variable given the others. Our results show that this function allows to detect anomalies variable by variable, and thus to localise the variables involved in the anomalies more precisely than univariate methods based on simple marginals.

1. Introduction

Anomaly detection refers to the task of detecting anomalous samples within a dataset described by $N$ variables, also called features. The localisation is the task that aims at identifying the subset of variables that are at the origin of the detected anomalies. While the problem of detection has been extensively studied in the machine learning literature (see (Hodge & Austin, 2004)), the problem of localisation in the presence of dependant variables remains a challenge.

In this paper, we propose to address this question using undirected probabilistic graphical models. Such models are particularly useful to represent the joint distribution over a set of $N$ random variables $X_1, \ldots, X_N$, in an efficient and compact way. Undirected graphical models are commonly tied to Gaussian random variables yet recent works have studied the possibility of building models over heterogeneous variables: (Yang et al., 2014) proposes a general class of graphical models where each node-conditional distribution is a member of a univariate exponential distribution and (Lee & Hastie, 2015; Laby et al., 2015) investigate the problem of learning the structure of pairwise graphical model over both discrete and continuous variables. This is done by optimizing the likelihood or the pseudo-likelihood, penalized with a Lasso or group Lasso regularisation.

A standard approach to perform online anomaly detection on temporal data such as signals, is to use the CUSUM algorithm (see (Page, 1954) and (Basseville et al., 1993)). In this work we propose a two-sided test with an adapted CUSUM algorithm to detect anomalies that occur in the conditional distributions rather than in the marginal. The resulting algorithm allows to perform change-point detection and to detect, variable by variable, continuous or categorical, the time when the distribution of the data changes from the “normal” distribution.

2. Mixed Model Presentation

The definition of graphical models relies on the factorisation of the joint distribution. Pairwise models form a particular class of models where the features are grouped in sets of one or two variables. Such models have been widely studied and have a number of practical advantages (Schmidt, 2010). In this paper, we focus on mixed models mixing binary variables $X_C = \{X_i, i \in C\}$ (called categorical thereafter), and continuous variables $X_Q = \{X_u, u \in Q\}$ (called quantitative thereafter). We have
Figure 1. I.d. samplings of $X_Q$ in two dimensions. The values of $X_Q$ are represented by 2^3 different colors. On the left, $\Phi = 0$ and on the right, $\Phi \neq 0$.

(Wolff, 1989), and then to sample $X_Q$ from the conditional Gaussian density (2). This procedure will be used in our numerical experiments below.

In this paper, we do not aim at learning a graphical model but rather at exploiting one for anomaly detection and localisation. See (Yang et al., 2014), (Lee & Hastie, 2015) and (Laby et al., 2015) for recent works that investigate the task of learning the parameters of a mixed undirected graphical model.

3. Anomaly detection and localisation

In this section, we present a method to detect and localise anomalies from a sequence of new data $(X_C^{(t)}, X_Q^{(t)})$, $t = 1, 2, \ldots$, assuming a reference model $\Omega$ that has already been learned using normal data.

The idea to localise anomalies is to monitor each term of the log-pseudo-likelihood (Besag, 1975) as a function of time. The CUSUM algorithm (Page, 1954) has been introduced to sequentially detect a change in the mean of a random variable. Since we want to detect an increase or a decrease, we use the two-sided CUSUM algorithm as proposed in (Basseville et al., 1993). For each $t$ and each variable $X_i$, we define the instantaneous conditional log-likelihood ratio

$$s_i(t) = \log \left( \frac{p(X_i^{(t)}|X_{-i}^{(t)})}{p_{\Omega}(X_i^{(t)}|X_{-i}^{(t)})} \right).$$

(4)

where $X_{-i} = \{X_j, j \in C \cup Q, j \neq i\}$, and a decision statistic defined recursively by $S_i^{(0)} = 0$ and

$$S_i^{(t)} = (S_i^{(t-1)} + s_i^{(t)})^+, \ t = 1, 2, \ldots,$$

(5)

where $(z)^+ = \max(z, 0)$. Here $p$ denotes the density of the alternative hypothesis, that is, the conditional density of the targeted anomalous behaviour.

We focus first on the quantitative variables. By (2), the conditional distribution of $X_Q^{(t)}$ given $X_C^{(t)}$ is the multivariate

$$p_{\Omega}(x_C, x_Q) = \prod_{i \in C} p_{\Omega}(x_C^{(i)}),$$

(3)

where $C$ is the set of quantitative variables. For each $t$ and each variable $X_i$, we define the conditional log-likelihood ratio

$$s_i(t) = \log \left( \frac{p(X_i^{(t)}|X_{-i}^{(t)})}{p_{\Omega}(X_i^{(t)}|X_{-i}^{(t)})} \right).$$

(4)

where $X_{-i} = \{X_j, j \in C \cup Q, j \neq i\}$, and a decision statistic defined recursively by $S_i^{(0)} = 0$ and

$$S_i^{(t)} = (S_i^{(t-1)} + s_i^{(t)})^+, \ t = 1, 2, \ldots,$$

(5)

where $(z)^+ = \max(z, 0)$. Here $p$ denotes the density of the alternative hypothesis, that is, the conditional density of the targeted anomalous behaviour.

We focus first on the quantitative variables. By (2), the conditional distribution of $X_Q^{(t)}$ given $X_C^{(t)}$ is the multivariate
We focus now on the categorical variables. Each variable $X^k$ follows a Bernoulli distribution with mean $p_i = \mathbb{E}_{\Omega}[X^k_i | X^k_{-i}] = \frac{e^{q_i(X^k_i, \Delta)}}{1 + e^{q_i(X^k_i, \Delta)}}$.

In the case of categorical variables, we define the conditional distribution of the alternative hypothesis as a Bernoulli distribution with mean $a_i$. The instantaneous log-likelihood ratio is then given by

$$s_i^t = X_i^t \log \frac{a_i}{p_i} + \left(1 - X_i^t\right) \log \left(\frac{1 - a_i}{1 - p_i}\right).$$

We choose $a_i$ such as the drift of the decision function (5) under the null hypothesis is set to the same value $-\frac{d^2}{2}$, as for quantitative variables in (6). This drift is given by computing $\mathbb{E}_{\Omega}[s_i^t | X_{-i}^t]$ with $s_i^t$ as in (6), yielding the equation

$$p_i \log \frac{a_i}{p_i} + \left(1 - p_i\right) \log \left(\frac{1 - a_i}{1 - p_i}\right) = -\frac{d^2}{2}. \quad (10)$$

It is easy to show that this equation in $a_i$ (with $d$ and $p_i$ fixed) has two distinct solutions $a_i^{(\uparrow)} \in [p_i, 1]$ (associated to the statistic $S_i^{(\uparrow)}$) and $a_i^{(\downarrow)} \in [0, p_i]$ (associated to $S_i^{(\downarrow)}$), detecting respectively increase and decrease of the mean $p_i(t)$, with a conditional negative drift $-\frac{d^2}{2}$ under the null hypothesis. For the same reasons as with quantitative variables, we will consider the sum $S_i^t = S_i^{(\uparrow)} + S_i^{(\downarrow)}$ in the experiments.

Under the null hypothesis each decision statistic $S_i^{(\uparrow)}$ or $S_i^{(\downarrow)}$ evolves with a negative drift $-\frac{d^2}{2}$. Hence, because of the positive part in (6), it remains close to zero with high probability. In contrast, under the alternative, the conditional drift becomes positive and the decision statistic $S_i^t$ eventually increase above any arbitrarily high threshold $h$. We thus label as a change time the first times $t$ when $S_i^t > h$. The choice of $h$ selects how sensitive the test is to a change in both possible directions.

Note that, by (6), the conditional negative drift under the null hypothesis (when no changes occur) of the decision statistic (5) is given by $\mathbb{E}_{\Omega}[s_i^t | X_{-i}^t] = -\frac{d^2}{2}$.

We focus now on the categorical variables. Each variable $X_i, i \in \mathcal{C}$ has a conditional Bernoulli distribution with mean

$$p_i = \mathbb{E}_{\Omega}[X_i | X_{-i}] = \frac{e^{q_i(X_i) + \Phi X_{-i}}}{1 + e^{q_i(X_i) + \Phi X_{-i}}},$$

where

$$q_i(X_i, \Delta) = \theta_{i0,i0} + 2\theta_{i0,-i0}X_{-i0} + \Phi_{i0,\Delta}X_{\Delta}.$$ (8)

In the case of categorical variables, we define the conditional distribution of the alternative hypothesis as a Bernoulli distribution with mean $a_i$. The instantaneous log-likelihood ratio is then given by

$$s_i^t = X_i^t \log \frac{a_i}{p_i} + \left(1 - X_i^t\right) \log \left(\frac{1 - a_i}{1 - p_i}\right).$$ (9)

We choose $a_i$ such as the drift of the decision function (5) under the null hypothesis is set to the same value $-\frac{d^2}{2}$.
We compare our method to the Wilcoxon test presented in (Lung-Yut-Fong et al., 2011), which is designed to detect changes in the distribution of a set of quantitative variables from batch data. In the following, we thus apply this approach to detect a change of distribution for each quantitative variable. Figure 4 displays the statistic of this test as a function of the possible change times. When only one change occurs in the data, this statistic is expected to approximately have a triangle shape with a maxima or a minima around the true change time. We use the same dataset as for the experiment with the anomalies localised on the second (green) quantitative variable, where \( \mu_1 \) changes from 0 to 3 at time \( t = 50 \). Figure 4 should thus be compared with the top row of Figure 3. In contrast to online methods such as the one we propose, this Wilcoxon statistic cannot be computed recursively as it requires the whole set of data to be computed. Moreover it is not suited to localise the anomaly since a change of \( \mu_1 \), although it only modifies the conditional distribution of \( X_1 \) given \( X_{−1} \), yields a change of all the marginal distributions. This is why in Figure 4, the Wilcoxon statistics display triangle shapes for all the quantitative variables with a more obvious change for the variables directly connected to \( X_1 \).

5. Conclusion

In this paper, we proposed an online method that allows to detect anomalies in a data stream, but more importantly to localise which variables are at the origin of the problem. By using a mixed undirected graphical model learned over a set of normal data, we manage to track changes occurring in the conditional distributions which offers more specific detections than when studying only marginal distributions. This method is based on a two-sided CUSUM algorithm, where decision statistics are computed for every variable and involve the calculation of conditional likelihoods.
References


