2012

Change Points Detection of Vector Autoregressive Model using SDVAR Algorithm

Fatimah Saaid
University of Newcastle

Darfiana Nur
University of Newcastle

Robert King
University of Newcastle

Change Points Detection of Vector Autoregressive Model using SDVAR Algorithm

Abstract

Part of a larger research project to detect fraudulent acts using the telecommunications call details record (CDR) is to locate the change points which could lead to detecting suspicious (fraudulent) calls. The capability of sequential discounting for autoregressive (SDAR) model learning algorithm (as proposed by [6]) to detect change points in time series data is explored. The algorithm is extended to multivariate time series by employing vector autoregressive model using SDVAR. Simulation and real data experiments to illustrate the new algorithm are discussed in this paper.

Keywords

Multivariate time series, Change points, Fraud, Vector autoregressive, SDVAR

Publication Details

Saaid, Fatimah Almah; Nur, Darfiana; and Robert King, Change Points Detection of Vector Autoregressive Model using SDVAR Algorithm, Proceedings of the Fifth Annual ASEARC Conference - Looking to the future - Programme and Proceedings, 2 - 3 February 2012, University of Wollongong.

Change Points Detection of Vector Autoregressive Model using SDVAR Algorithm

Fatimah Almah Saaid, Darfiana Nur, Robert King
The University of Newcastle, Callaghan, NSW, 2308, AUSTRALIA

Abstract

Part of a larger research project to detect fraudulent acts using the telecommunications call details record (CDR) is to locate the change points which could lead to detecting suspicious (fraudulent) calls. The capability of sequential discounting for autoregressive (SDAR) model learning algorithm (as proposed by [6]) to detect change points in time series data is explored. The algorithm is extended to multivariate time series by employing vector autoregressive model using SDVAR. Simulation and real data experiments to illustrate the new algorithm are discussed in this paper.

Key words: Multivariate time series, Change points, Fraud, Vector autoregressive, SDVAR

1. Introduction

Telecommunication companies generate large amounts of call data, known as the call details record (CDR) that operated in real-time basis. Despite the sophisticated technologies, this industry is facing a huge impact due to fraudulent acts. Part of a larger research project to detect fraudulent acts using the telecommunications CDR is to locate the change points which could lead to detecting suspicious (fraudulent) calls. The aim of this paper is to detect change points from the CDRs (as indicative of fraudulent acts) by incorporating unified detection scheme introduced by [6] where the learning model algorithm is extended to a multivariate time series. The algorithm, called Sequential Discounting for Vector Autoregressive (SDVAR), is proposed to detect fraud as soon as it occurs. The remainder of this paper is organized as follows. The following section reviews some previous works in change points detection. Section 3 discusses the method designed for the study. The discussion on the simulation and case study are presented in Section 4. The last section gives some discussions and conclusions.

2. Change Points Analysis

Change points detection has been used in diverse fields. [3] proposed a geometric method for estimating linear state-space models for identifying change points in time-series data. Whilst Bayesian change points is applied by [2] to detect regions of genetic alteration in cancer research. It has also been used in detecting change points of the number of annual tropical cyclone [1]. In recent study, [9] introduced the combination of wavelet denoising and sequential approach to detect change points on mobile phone based on the CDR. Network faulty monitoring is studied by [6]. They introduced a two-learning stage to detect outliers and change points in a unifying framework, ChangeFinder. The scheme is applied by employing autoregressive process where the model is learned using Sequential Discounting for Autoregressive (SDAR) algorithm,

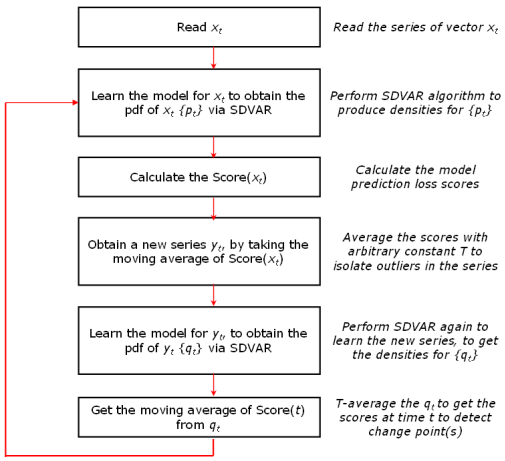

Figure 1: Unifying scheme with ultimate aim to detect change points

also being used by [8]. Adaptive to non-stationary time series is the key advantage of the algorithm. In this paper we study the call behaviour from the CDR by developing *growth profiles* for unique subscribers. The profiles are considered as the *referenced profiles* for normal callers where deviation (change) from these normal behaviours would lead to the identification of suspicious call (act of fraud).

3. Method

The detection flow, as displayed in Figure 1, shows the unified scheme in change detection. It presents two-stage learning scheme of the data in detecting outliers and change points in a single framework. We describe the unified scheme in change detection using the two-stage learning framework but with an extension of employing multivariate autoregressive representative of the time series.

3.1. Unified Detection Scheme

Unified detection scheme is to detect multiple outliers and change points in a time series.The estimation of VAR(p) parameters is done after a series is observed. The SDVAR algorithm in the flow involves online estimation of a time series data by introducing discounting parameter, r. The value of r is between 0 and 1 where the smaller r indicates higher influence of the past data.

3.2. Sequentially Discounting VAR (SDVAR)

Autoregressive (AR) model is the most typical time series model to predict the current value from the past values in a same univariate time series. The number of the past values (or lag values) is referred to the *order* of the model. However, with the increase interest in modelling a series with more than one variable, multivariate time series model is required. In a VAR, or also known as multivariate AR (MAR), the value of each variable at each time point is predicted from the values of the same series and those of all other time series, depending on the variables used in the model.

Sequentially Discounting VAR (SDVAR) is introduced in this paper to learn the model with VAR process. Consider the VAR(p) model where N be the length of m series. Let $\boldsymbol{x_t} = [x_{1,t}, ..., x_{m,t}]^T$ denote $(mx1)$ vectors of time series variables. Then VAR(p) is given by:

$$\boldsymbol{x_{i,t}} = \boldsymbol{\mu} + \boldsymbol{\Phi_1 x_{t-1}} + ... + \boldsymbol{\Phi_p x_{t-p}} + \boldsymbol{\varepsilon_{i,t}}, \tag{1}$$

where $t = 1, ..., N$, $i = 1, ..., m$, Φ_k, $k = 1, ..., p$ is (mxm) coefficient matrices, μ is $(mx1)$ vector and $\varepsilon_{i,t}$ is the $(mx1)$ vectors of i.i.d Gaussian noise with mean 0 and covariance matrix $\mathbf{\Gamma}$. The mean for the i-series is given as $E[x_{i,t}] = \mu_i$. Hence the mean vector is given by

$$\boldsymbol{\mu} = [\mu_1, \mu_2, , ..., \mu_m]^T. \tag{2}$$

The covariance matrix function, $\Gamma(h)$ of the vector process $\mathbf{x}_t$, given by [4] is:

$$\begin{aligned}\mathbf{\Gamma}(h) &= Cov\{\mathbf{x}_t, \mathbf{x}_{t+h}\} = E[(\mathbf{x}_t - \boldsymbol{\mu})(\mathbf{x}_{t+h} - \boldsymbol{\mu})^T] \\ &= E[x_{1,t} - \mu_1, \cdots, x_{m,t} - \mu_m][x_{1,t} - \mu_1, \cdots, x_{m,t} - \mu_m]^T \\ &= \begin{pmatrix} \gamma_{11}(h) & \gamma_{12}(h) & \cdots & \gamma_{1m}(h) \\ \gamma_{21}(h) & \gamma_{22}(h) & \cdots & \gamma_{2m}(h) \\ \vdots & \vdots & \vdots & \\ \gamma_{m1}(h) & \gamma_{m2}(h) & \cdots & \gamma_{mm}(h) \end{pmatrix}. \end{aligned} \tag{3}$$

For $i = j$, $\gamma_{ii}(h)$ is the autocovariance function for the i^{th} components of $\mathbf{x}_t$. While for $i \neq j$, $\gamma_{ij}(h)$ is the *cross-covariance* function between $x_{i,t}$ and $x_{j,t}$. The

cross-covariance $\gamma_{ij}(h)$ is calculated by

$$\gamma_{ij}(h) = E[(x_{i,t} - \mu_i)(x_{j,t+h} - \mu_j)] \quad (4)$$

By mapping with the SDAR, the SDVAR algorithm is developed to estimate the model parameters using VAR estimations by [4] and [7]. Let the discounting parameter $0 < r < 1$, the subsequent discussion describes the SDVAR algorithm, followed by section presenting the results of the simulation and real data experiment.

1. Initialize the parameter estimates, $\hat{\boldsymbol{\mu}}_{\mathbf{0}}$, $\boldsymbol{\Gamma}_j$, $\hat{\boldsymbol{\Phi}}_j$, $\hat{\boldsymbol{\Sigma}}$. The VAR model consists of parameters for each series $\boldsymbol{x_t}$ where the estimations of these parameters are all represented by matrices.

2. Parameter update: For each new data arrives (t=k+1, k+2, ...), read $\boldsymbol{x_t}$ and update the mean, the variances-covariances (and the cross-covariances) and the model's parameters: $\hat{\boldsymbol{\mu}}$=$(1-r)\hat{\boldsymbol{\mu}}_{\mathbf{0}} + r * \boldsymbol{x_t}$, which produces the mean vectors, $\boldsymbol{\mu_t}$, t=1,2,... The parameter updates to produce the variance-covariance matrix (Γ_0) as well as variance-cross-covariance matrix ($\boldsymbol{\Gamma}(h)$ for $h > 0$): $\boldsymbol{\Gamma}(h)$=$(1 - r) * \boldsymbol{\Gamma}(h) + r * (\mathbf{x}_t - \hat{\boldsymbol{\mu}})(\mathbf{x}_{t+h} - \hat{\boldsymbol{\mu}})^T$, h=0,1,...,p.

3. Calculate $\hat{\boldsymbol{x_t}}$ and $\hat{\boldsymbol{\Sigma}}$ where $\hat{\mathbf{x}}_t = \hat{\boldsymbol{\Phi}}_i(\mathbf{x}_{t-i} - \hat{\boldsymbol{\mu}}) + \hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}} = (1-r)\hat{\Sigma} + r(\boldsymbol{x_t} - \hat{\boldsymbol{x_t}})(x_t - \hat{\boldsymbol{x_t}})^T$.

4. Results

4.1. Simulation Study

A simulation sample of VAR(1) process with bivariate system (two variables $x1$ and $x2$) is generated using *mAr.sim* function of *mAr* package in R, as given in Eq. (5) and Eq. (6). The 2-dimensional VAR(1) with $N = 10,000$ is induced with Kullback-Leibler (KL) divergence, as displayed in Figure 2, where the change points are set to occur at each time x x 1000 for $x = 1, ..., 9$.

$$x_{1,t} = c_1 + \phi_{11}x_{1,t-1} + \phi_{12}x_{2,t-1} + \varepsilon_{1,t} \quad (5)$$
$$x_{2,t} = c_2 + \phi_{21}x_{2,t-1} + \phi_{22}x_{2,t-1} + \varepsilon_{2,t} \quad (6)$$

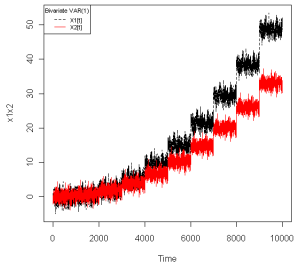

Figure 2: The simulated series with mean jumps created at each thousand

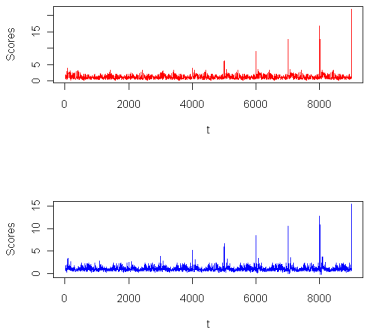

Figure 3: Change points of the simulated series with $r = 0.02$

where $\boldsymbol{\varepsilon}$ is the normally distributed white noise and $\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{\Sigma}_\varepsilon)$. The parameters used in the simulation model are as follows:

$$\boldsymbol{\Phi} = \begin{pmatrix} 0.4 & -0.5 \\ -0.3 & 0.4 \end{pmatrix}, \mathbf{c} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \text{ and } \boldsymbol{\Sigma}_\varepsilon = \begin{pmatrix} 1.0 & 0 \\ 0 & 1.0 \end{pmatrix}.$$

Using $r = 0.02$ and $T = 5$, R took 9.75 seconds to run the SDVAR algorithm for processing the 10,000 data (the speed is also depending on the processor of the computer machine). In Figure 3, the plot in the top is the change points detected for $x1$ and the bottom for $x2$, respectively. The sudden jumps are appropriately detected by the scheme using SDVAR learning model. The sudden jumps are prominent for $x1$ starting at $t = 5000$ where 5 change points are appropriately identified ($K(x1) \geq 15.125$). As sudden jump for $x2$ is detected as early as at time $t = 4000$ and 6 change points detected that corresponds to $K(x2) \geq 6.480$.

4.2. Case Study

A sample of 6-day CDR data is used as a training data in obtaining the user profiles of the subscribers. The

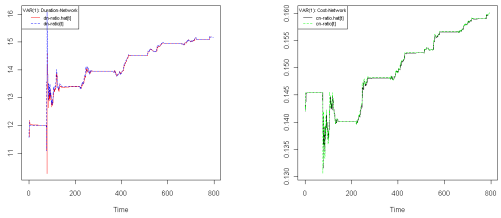

Figure 4: Duration-Network growth (left-hand side) and Cost-Network growth model learned using SDVAR

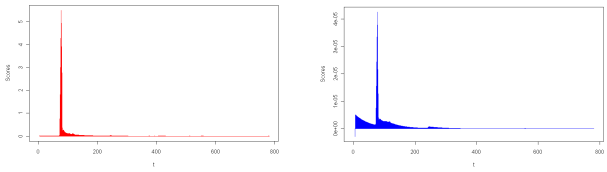

Figure 5: Change points detected for both Duration-Network growth (left-hand side) and Cost-Network growth profiles

CDR is collected from a PBX at one of telecommunications companies in Malaysia. The network growth profiles used as instantiations to detect change points in the CDR for the unique subscribers are duration-to-network and cost-to-network ratios (the network growth is the growth of destinations). One user account's profiles is used for this case study where the growth profiles are developed in 10 minutes interval for the whole six days. Using bivariate VAR(1) process of the two selected profile measurements, *mAr.est* function from *mAr* package in R is applied. Figure 4 exhibits the plot of the series using SDVAR algorithm as the model learning module. The 78^{th} of the 10-minute interval is identified as change point (Figure 5). It shows a sudden deviation from the normal behaviour which may indicate a suspicious act of fraudulent has occurred. The call duration growth in the 78^{th} interval is found to be two times higher than in the 77^{th} interval. In contrast, cost growth profile shows a strange behaviour which is not reflecting the long call made at t=78 of the corresponding user (Figure 4) and a sudden reduction of cost-to-network growth is detected as change point at time t=78 (Figure 5).

5. Conclusions

In this paper, an algorithm for learning a vector autoregressive process is proposed. The SDVAR algorithm is used as a learning module for change point analysis from nonstationary time series data in online manner. The change detection framework is based on the works of [6]. The approach is part of research in detecting fraudulent acts from a CDR. The algorithm is validated using simulation and case study. Findings from the simulation and case studies indicate that multiple (and single) change points can be detected which may lead to an alarming stage of detecting suspicious calls in the CDR.

References

[1] P. Chu and X. Zhao, Bayesian Change-Point Analysis of Tropical Cyclone Activity: The Central North Pacific Case, Journal of Climate, 17(24), 2004:4893-4901.

[2] C. Erdman and J.W. Emerson, A fast Bayesian change point analysis for the segmentation, Bioinformatics, 24(19), 2008: 2143-2148

[3] Y. Kawahara, T. Yairi and K. Machida,Change-Point Detection in Time-Series Data Based on Subspace Identification, Seventh IEEE International Conference on Data Mining (ICDM) 2007, pp.559-564.

[4] H. Lutkepohl, Introduction to Multiple Time Series Analysis, Berlin: Springer-Verlag, 1993.

[5] F.A. Saaid, R. King and D. Nur, Development of Users Call Profiles using Unsupervised Random Forest, Third Annual ASEARC Conference, December 78, 2009, Newcastle, Australia.

[6] J. Takeuchi and K. Yamanishi, A Unifying Framework for Detecting Outliers and Change Points from Time Series, IEEE Transactions on Knowledge and Data Engineering, 18(4), 2006:482-492.

[7] W.W.S. Wei, Time series analysis : univariate and multivariate methods, Redwood City, California: Addison-Wesley Pub., 1990.

[8] B.-K. Yi, N.D.Sidiropoulos, T. Johnson, H.V. Jagadish, C. Faloutsos and A. Biliris, Online Data Mining for Co-Evolving Time Sequences, Proceedings of the 16th International Conference in Data Engineering, 2000.

[9] H. Zhang, R. Dantu and J.W. Cangussu, Change Point Detection based on Call Detail Records, IEEE International Conference on Intelligence and Security Informatics (ISI '09), 2009.