Multivariate Anomaly Detection in Time-Series Data

A Goal Document for a Master's Thesis Work

Student:	Supervisor:	Assistant supervisor:	Examiner:
Emil Andersson +46738071334 emil.andersson207@gmail.com	Maria Sandsten Dept. of Mathematical Statistics +462224953 maria.sandsten@matstat.lu.se	Mattias Jönsson Sigma ITC +46704019975 mattias.p.jonsson@sigma.se	Andreas Jakobsson Dept. of Mathematical Statistics +462224520 aj@maths.lth.se

Preliminary start and end date: *12 August 2019 — 27 December 2019*

Background and motivations

Sensor systems appear in many real-world applications such as space shuttles, health care monitoring and factories. In those systems, a lot of multivariate time series data is generated continuously, and it is often of relevance to find unexpected events, also called anomalies.

Anomaly detection has historically been a common topic for research, although this has mostly been with univariate data in smaller amounts. However, with the rise of the Internet of Things (IoT) and the rapid increase of computational power, this has changed. The amount of data is usually very big and in many cases the data is unlabeled. Hence the models must be unsupervised. Sigma ITC's division for IoT is often encountering multivariate anomaly detection in their projects and want to research the topic methodically.

Overall objectives and issues/research questions

The overall objective of the thesis is to evaluate approaches for minimizing the number of false positives when detecting anomalies in multivariate time-series data from sensor networks. Since the number of anomalies often are a small subset of all the values, a larger amount of non-anomalous data is checked leading to a high amount of false positives. The explored data could be non-stationary and may contain frequencies. Furthermore, the data is supposed to be unlabeled, or just partly labelled. Although evaluation will be made on generated labelled datasets as well.

Approach/methodology and methods

The work will be carried out as an explorative analysis of methods based on classical statistics and machine learning. For getting an overview of the subject a book on Outlier Analysis [Aggarwal, 2013], with specific chapters discussing techniques for multivariate time-series data will be studied. At the company earlier work has focused on using isolation forests and autoencoders to solve the problem. Isolation forests have recently been proven to be effective at finding anomalies, however, further exploration of how well the technique performs on multivariate time series is needed. Methods based on state-of-the-art machine learning techniques such as generative adversarial networks (GANs) and long-short term memory (LSTM) have recently been developed as well, the possibility of using those techniques for solving the problem will be evaluated.

Data from a rolling process will be used for testing the models, however, the labelling of real-world data is rarely perfect. Hence, ARIMA techniques will be used to simulate multivariate data with anomalies. The methods will be evaluated on these datasets with techniques such as ROC curves and precision and recall.

Related work and proven experience

Aggarwal, C. C. (2013). Outlier analysis (book). Outlier Analysis. https://doi.org/10.1007/978-1-4614-6396-2

Li, D., Chen, D., Shi, L., Jin, B., Goh, J., & Ng, S.-K. (2019). MAD-GAN: Multivariate Anomaly Detection for Time Series Data with Generative Adversarial Networks. Retrieved from http://arxiv.org/abs/1901.04997

Hundman, K., Constantinou, V., Laporte, C., Colwell, I., & Soderstrom, T. (n.d.). Detecting Spacecraft Anomalies Using LSTMs and Nonparametric Dynamic Thresholding. KDD, 18. https://doi.org/10.1145/3219819.3219845

Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008). Isolation forest. Proceedings - IEEE International Conference on Data Mining, ICDM, 413–422. https://doi.org/10.1109/ICDM.2008.17

Expected contribution to the development of knowledge

Methods for analyzing univariate data has been extensively researched. However, there are shortcomings in the research of methods for finding anomalies in multivariate time series data.

Preliminary resources

Sigma ITC will provide a workplace and a computer that is capable to analyze the data. They will also provide the student with multivariate time-series data from different projects they have taken part in.