Degree Project - Specification

Correlating heterogeneous time-series

Alexandre GUINAUDEAU alegui@kth.se

Supervisor: Pawel Herman

Examiner: Hedvig Kjellström

March 31, 2018

Background & objective

As it spreads to many industries and increases in volume, sensor data is a becoming an important field of research in terms of security, storage, and analysis. In manufacturing plants, aircrafts, electronic intensive care units or data centers, sensors continuously monitor many parameters of the environment. The data generated by these sensors is then processed to optimize the way the system works, or to diagnose failures once they occurred. These systems can have hundreds or thousands of sensors with sub-second sampling. With this amount of data, diagnosing failures requires an efficient processing of the data. This is even more complex with "flaky" failures, which do not occur systematically, but only with a certain probability. This could be due to a bad wire connection for instance.

Interest and assignment

Finding the sensors that contain useful information is critical to understand the circumstances of the incident and troubleshoot it. In most cases, the cause of the failures is specific to what the data represents. For instance, the sensor data could be a consequence of the failure rather than a cause, or a third factor could have affected both the sensor and the system. In any case, providing the system engineer with a limited list of sensors that seem to have some correlation with the failure is very valuable to enable them to quickly troubleshoot the problem. Once the engineers know where to look, they are usually able to understand what happened by diving into the specific environment parameters.

The main difficulty to solve this problem is the heterogeneity of the data: it can be continuous data (for instance measures of environment parameters such as a pressure or a temperature), categorical data (for instance the part that is installed on a machine) or events (for instance when an operation is triggered). In many cases a time series can be represented in all of these ways, but in general one of them makes more sense.

In this thesis we do not care whether the one time series is the cause of the other, nor whether there is a positive or linear correlation between them. We only care about is a generalization of the *association* of binary variables, more precisely whether the knowledge of one times series can improve the understanding of the other one.

Objective

The desired outcome of the degree project is a metric that accurately captures the association between heterogeneous time series.

Research question & method

When can a continuous or categorical time series be considered as associated to a binary event time series? More specifically, what metric can be defined to measure association between heterogeneous time series?

I have access to a large set of real data - both sensor and event data. However, given the sensitivity of this data, I will generate data for this report. I will generate fake data, starting with examples based on simple probability rules to understand how correlation metrics work and how they can be generalized. Then I will use notional data from open source examples or generated in a real environment, to confirm the hypothesis scale to real-life examples.

Specified problem definition How can we generalize correlation metrics such as Pearson coefficient to apply to apply to heterogeneous time series?

Examination method To compare models, we will simulate time series that are inter-dependent to one another, as well as time series that are independent. For fake data, we will either use the mathematical definition or Monte-Carlo methods to confirm the validity of the model.

Finally, we will rank real-world series based on their computed association to confirm that highly-associated series have the expected behavior on more complex data.

Expected scientific results We will first study usual correlation metrics, and generalize their definition to categorical, event and continuous time-series. Finally, we will see whether these metrics are comparable, in other words if it is possible to decide if a categorical time series is more associated to the continuous time series than the events one. This would not be compulsory to solve the problem of incident diagnosis, as it would already be very valuable to find the time series candidates for each type.

Evaluation

The objective is to find elements in the data that are associated to the apparition of events.

News value Automatically finding associated sensor time series would be very valuable for all of the systems described above to troubleshoot failures faster and anticipate them. It could also be applied to any system that has a logging system generating frequent debug messages and a few error events independently.

Pilot study

Correlation metrics The most common metric to measure correlation is the *Pearson* coefficient. It can easily be interpreted and visualized [3], and often gives very good results. However, in some case, it can be misleading, especially because it only captures linear correlations and because it assumes normal distribution of the input parameters.

In this case, the *Spearman* coefficient is more robust (although less efficient), as it only compares ranks between elements, and therefore captures monotonic correlations, even if there are not linear. The *Kendall* coefficient is very similar as it also uses the rank of elements to compare them, it is less efficient but more robust. It is also much easier to interpret, as it captures the percentage of elements positively correlated [4, 16]. These correlation metrics are probably more adapted to our problem for numerical data as they capture non-linear correlations, but may be harder to generalize because their definition is not as simple as Pearson's.

All these usual correlation metrics apply to numerical data, and cannot therefore be used for categorical data. In this case, a metric that often works well is the J-measure, which is based on the concept of entropy in information theory. For binary data, it is possible to transform the data to make most association measures equivalent [20]. This is not as clear for more complex data, and requires to clearly define the concept of association for categorical and continuous data as well.

However, it is not clear if this measure can be compared to the previous ones when the data is both numerical and categorical - for instance when it is binary. In [20], P-N Tan et al. compared many existing norms on binary classification by ranking 10 examples from highest to lowest association. Even on this very simple example, they show that different metrics rank

the examples in different orders, but that there is a way to normalize the data where all the norms become equivalent. [Goodman, R.M. and Smyth, P. An information-theoretic model for rule-based expert systems. 1988 Int. Symposium in Information Theory, Kobe, Japan, 1988] / [G. Piateski and W. Frawley. Knowledge discovery in databases. MIT press, 1991.]

Comparison [[Selecting the right objective measure for association analysis Pang-Ning Tan*, Vipin Kumar, Jaideep Srivastava Department of Computer Science, University of Minnesota, 200 Union Street SE, Minneapolis, MN 55455, USA]]

[Croux, C. and Dehon, C. (2010). Influence functions of the Spearman and Kendall correlation measures. Statistical Methods and Applications, 19, 497-515.]

Application to heterogeneous data Several tools exist to analyze the correlation between continuous time series [18], or between events [17], but they do not perform well when it comes to correlating continuous time series and events. However, sensor data is heterogeneous: it can be continuous, discrete, categorical or binary. Usual measures of correlation such as the Pearson and Spearman correlation do not perform well on this kind of data [19]. Therefore, we have to find other ways of defining correlation, or to map continuous time series to event time series to use existing tools.

A lot of work has been done to detect anomalies in continuous time series. Most sensor data - such as Climate [1], Intensive Care Units [7] or computer resources [19] - has natural frequencies. It is therefore possible to find pseudo-periods in the data that facilitate its pre-processing and enable multi-scale analysis [13]. In other words, the time series can be decomposed into scales, in order to detect both local [5, 8, 12] and global anomalies [9, 11]. Shahabi et al. even defined a surprise score that enables the comparison of all of these potential outliers, regardless of the scale [10].

For categorical time series, we could look for change points. These points are the best candidates to induce failures in the system [2, 6].

Finally, a completely different approach to this problem could be to cluster timeseries (for example using the method described in [14]), and then find the best candidates within the event's cluster. Time series that are similar to the failure time series are more likely to be correlated.

Conditions & schedule

Resources needed to solve the problem consist in the access to the data and the infrastructure to store and process it.

Limitations We assume the data has been cleaned. If relevant, it has been de-trended and de-seasonalized.

Collaboration with principal supervisor My supervisor will mostly contribute in general thesis and problem definition.

Schedule The first month will be used to get a grasp of the state of the art and to define correlation in a scientific way.

The two following ones will be used to define metrics and try them on simple examples to confirm they accurately capture the correlation.

During the two following months, we will work on more complex examples and see if the metric generalizes well, and if different metrics can be compared to one another.

Finally, the last weeks of the project will be the opportunity to summarize the work done, write the report and prepare the presentation.

References

[1] Claude Frankignoul, Klaus Hasselmann (1977)
Stochastic climate models, Part II - Application to sea-surface temperature anomalies and thermocline variability, Tellus, 29:4, 289-305
https://doi.org/10.3402/tellusa.v29i4.11362

[2] David S. Stoffer, David E. Tyler, Andrew J. McDougall (1993)
Spectral Analysis for Categorical Time Series: Scaling and the Spectral Envelope, Biometrika Vol. 80, No. 3, pp. 611-622
https://www.researchgate.net/profile/David_Tyler2/publication/239726763_Spectral_Analysis_for_Categorical_Time_Series_Scaling_and_the_Spectral_Envelope/links/0046352d842c990d8d000000.pdf

[3] Joseph Lee Rodgers and W. Alan Nicewander (1987)
Thirteen Ways to Look at the Correlation Coefficient, pp 59-66 Download citation https://doi.org/10.1080/00031305.1988.10475524

[4] Valz P.D. & Thompson M.E. (1994)
Exact inference for Kendall's S and Spearman's rho, Journal of Computational and Graphical Statistics 3: 459-472.

[5] E. M. Knorr and R. T. Ng. (1998)
Algorithms for Mining Distance-Based Outliers, In Proceedings of the 24th International Conference on Very Large Databases (VLDB), pages 392-403, 1998.
http://www.vldb.org/conf/1998/p392.pdf

[6] Valery Guralnik, Jaideep Srivastava (1999)
Event Detection from Time Series Data, In Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'99), pp. 33-42, August 15-18, 1999, San Diego, CA, USA.
http://dmr.cs.umn.edu/Papers/P1999_6.pdf

[7] M.-C. Chambrin, P. Ravaux, D. Calvelo-Aros, A. Jaborska, C. Chopin, B. Boniface (1999)
Multicentric study of monitoring alarms in the adult intensive care unit (ICU): a descriptive analysis, Intensive Care Medicine, Volume 25, Issue 12, pp 1360-1366
https://doi.org/10.1007/s001340051082

[8] S. Ramaswamy, R. Rastogi, and K. Shim. (2000)
Efficient algorithms for mining outliers from large data sets, In SIGMOD '00: Proceedings of the 2000 ACM SIGMOD international conference on Management of data, pages 427-438, 2000. https://dl.acm.org/citation.cfm?id=335437&dl=ACM&coll=DL&CFID=841695128&CFTOKEN=52089307

[9] M. M. Breunig, H. Kriegel, R. T. Ng, and J. Sander (2000)
LOF: Identifying density-based local outlier In Proceedings of the ACM SIGMOD International Conference on Management of Data, pages 93-104, 2000
http://www.dbs.ifi.lmu.de/Publikationen/Papers/LOF.pdf

[10] C. Shahabi, X. Tian, and W. Zhao. TSA-tree (2000)
A wavelet-based approach to improve the efficiency of multilevel surprise and trend queries on time-series data, In Statistical and Scientific Database Management, pages 55-68, 2000.
https://infolab.usc.edu/DocsDemos/pakdd01.pdf

[11] C.Shahabi, S. Chung, M.Safar and G.Ha jj (2001)
2D TSA-tree: A Wavelet-Based Approach to Improve the Efficiency of Multi-Level Spatial Data Mining, Technical Report 01-740, Department of Computer Science, University of Southern California. (2001)
https://pdfs.semanticscholar.org/39c5/5ee09a2c49e736de730e2cc7cc61f789ace1.pdf

[12] F. Angiulli and C. Pizzuti. (2002)
Fast outlier detection in high dimensional spaces,In PKDD '02: Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery, pages 15-26, 2002 https://link.springer.com/chapter/10.1007/3-540-45681-3_2

[13] Costa M, Goldberger AL, Peng CK (2002)
Multiscale entropy analysis of complex physiologic time series, Phys Rev Lett 2002, 89: 068102. 10.1103/PhysRevLett.89.068102
https://dbiom.org/files/publications/Peng_MultiscaleEntropyAnalysisComplexPhysiologicTimeSeries.pdf

[14] Keogh, E., Lonardi, S., Ratanamahatana, C. (2004)
Towards Parameter-Free Data Mining, In proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
http://www.cs.ucr.edu/~eamonn/SIGKDD_2004_long.pdf

[15] Umaa Rebbapragada , Pavlos Protopapas , Carla E. Brodley , Charles Alcock (2009)
Finding anomalous periodic time series Machine Learning, v.74 n.3, p.281-313, March 2009. doi: 10.1007/s10994-008-5093-3
`https://arxiv.org/pdf/0905.3428.pdf`

[16] Xu Weichao, Yunhe Hou, Hung Y. S. & Yuexian Zou (2010)
Comparison of Spearman's rho and Kendall's tau in normal and contaminated normal models, Manuscript submitted to IEEE Transactions on Information Theory
`http://arxiv.org/PS_cache/arxiv/pdf/1011/1011.2009v1.pdf`

[17] J.-G. Lou, Q. Fu, Y. Wang, and J. Li (2010)
Mining dependency in distributed systems through unstructured logs analysis, SIGOPS Operating Systems Review, 41(1):91-96, 2010.

[18] D. Wu, Y. Ke, J. X. Yu, S. Y. Philip, and L. Chen (2010)
Detecting leaders from correlated time series, In Database Systems for Advanced Applications, pages 352-367. Springer.

[19] Zhang, Dongmei and Lou, Jian-Guang and Ding, Justin and Fu, Qiang and Lin, Qingwei (2014)
Correlating Events with Time Series for Incident Diagnosis, SigKDD'14, July 2014
`https://www.microsoft.com/en-us/research/publication/correlating-events-time-series-incident-diagnosis-2/`

[20] Pang-Ning Tan*, Vipin Kumar, Jaideep Srivastava (2004)
Selecting the right objective measure for association analysis