Degree Project

Measuring association between heterogeneous time-series

Alexandre GUINAUDEAU
alegui@kth.se

Supervisor: Pawel Herman

Examiner: Hedvig Kjellström

April 2, 2018

Contents

Introduction

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.

Finding other data that is associated with the failures and can contribute to explaining them 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.

Interest and assignment 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.

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 use other 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.

1 Pilot study

1.1 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 be applied to any system that has a logging system generating frequent debug messages and a few error events independently. Microsoft researched suggested a solution to correlate continuous and event time series [21] but focused more on the temporal correlation - so the causal correlation - rather than the actual association between timeseries. For non-stationary time series, we assume they have been transformed to be stationary - or more exactly, we will show some ways to transform stationary timeseries to stationary ones.

1.2 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 [5, 18]. 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.

1.3 Association metrics

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 [4, 7], 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 [15]. 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 [15], P-N Tan et al. compared many existing norms on binary classification, and showed their was a way to normalize the data in order to make all the norms they studied equivalent. Such techniques will be important to pre-process the data accurately and to generalize norms to heterogeneous time series.

1.4 Application to heterogeneous data

Several tools exist to analyze the correlation between continuous time series [20], or between events [19], 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 [21]. Therefore, we have to find other ways of defining association. We can also map continuous or categorical time series to event time series - by detecting outliers or change points - and apply existing tools to the transformed data [2, 8].

2 Defining association for heterogeneous time series

In order to find time series which can help us troubleshoot failures, we must first define *association* in a clear way. To do so, we will define a set of properties that a good association metric should abide by, and then define very simple examples where we know how the association metric should behave. Then, we will compare candidate metrics based on these properties and examples in order to select the ones that correctly measure association.

2.1 Fundamental properties

2.1.1 Notations and definitions

Given two time series x and y, we note $A(x, y)$ the association between them. Without loss of generality, we can assume that our metric gives a score from 0 to 1.

x and y are associated if the knowledge of one improves the understanding of the other, in other words if the posterior model forecasts values better than the prior model. More formally, we note $\mathcal{P}(x_t|x_0, \ldots, x_{t-1})$ the forecast of x_t at the instant t, given all previous observations, and $\mathcal{P}_y(x_t|x_0, \ldots, x_{t-1})$ the forecast of x_t at the instant t given all previous observations of x and all observations of y.

We also define a loss metric $\mathcal{L}$ which measures the surprise:

$$\mathcal{L}(x)_t = dist(\mathcal{P}(x_t|x_0, \ldots, x_{t-1}), x_t)$$
$$\mathcal{L}(x|y)_t = dist(\mathcal{P}_y(x_t|x_0, \ldots, x_{t-1}), x_t)$$

where *dist* is a distance metric.

2.1.2 Properties

With these notations, a good metric to measure association between time series should follow the following properties:

P1. (Independence) $A(x, y) = 0 \Leftrightarrow x \perp\!\!\!\perp y$

P2. (Information gain) $A(x, y) > 0 \Leftrightarrow \mathcal{L}(x|y) < \mathcal{L}(x)$

P3. (Perfect knowledge) $A(x, y) = 1 \Leftrightarrow \mathcal{L}(x|y) = 0$

P4. (Symmetry) $A(x, y) = A(y, x)$: A should be symmetric

2.2 Expected behavior for simple examples

We also define simple examples where we can know how the association should behave.

2.2.1 Categorical and event data

Our first example compares an event time series and a categorical one. This could for instance be useful to determine whether an installed part tends to fail more often than another. We will start with the case of two categories and then generalize to any number of categories. Let x be a time series of events, that occur with probability p_1 during t_1 and then at frequency p_2 during t_2. Let $t_{tot} = t_1 + t_2$ the total time, $p_{tot} = \frac{p_1 n_1 + p_2 n_2}{n_1 + n_2}$ the average probability of failure. Let y be a categorical time series with value C_1 during t_1 and C_2 during t_2.

Figure 1: Example of associated categorical and event series

Figure 1 shows an example of such time series x and y. Let's consider other candidate time series y_k with value C_1 during $t'_1 = k$ and C_2 during $t'_2 = t_{tot} - k$. A good association metric should rank y_i time series such that y_{t_1} has the highest association with x.

2.2.2 Numerical and event data

For numerical data, we cannot expect to have a metric that computes the association in the right way on any raw dataset, because the association depends on what the data represents. Events can be triggered by extreme values or by sudden changes, they can depend on the original value, or on the de-trended or de-seasoned value. Events could also have an impact in the future, or for a specific duration. Therefore, we assume the data is preprocessed (potentially de-trended, de-seasoned, shifted, smoothed, derived, normalized) such that large values have a higher risk of triggering an event. This is not a real loss of generality, because one can derive multiple preprocessed

series out of the original one, for instance using all the methods described above, and the algorithm will only find candidates with high association with the event series.

Let x be a time series of events, and $y \sim \mathcal{N}(0, \sigma)$ be a series of Gaussian white noise, which could be for instance the residual of a de-seasoned and de-trended series. When an event occurs, y has a higher value in average:

$$y \sim \begin{cases} \mathcal{N}(0, \sigma) & if\ x = 0 \\ \mathcal{N}(1, \sigma) & if\ x = 1 \end{cases}$$

In this case, we expect x and y to be highly associated. Furthermore, if we define other numerical series that are impacted by a fraction of the events, or by additional events, they should have a lower association with x.

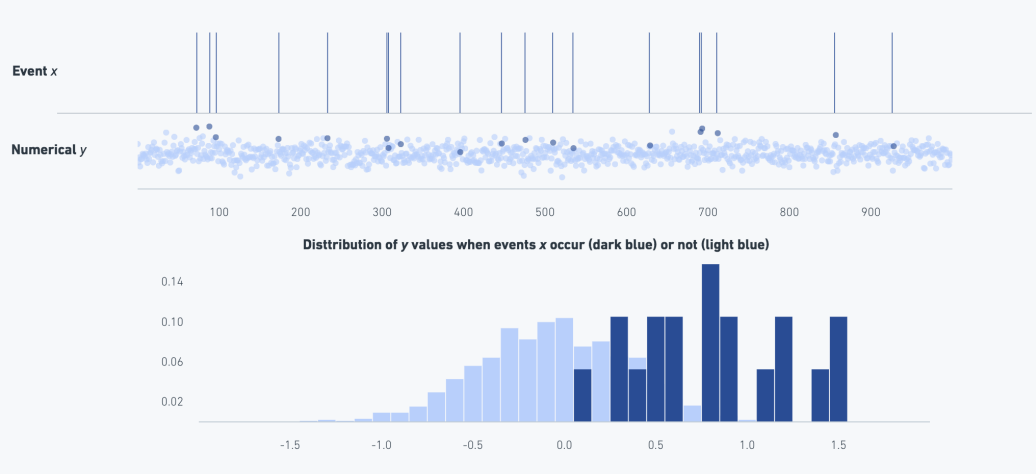

Figure 2: Example of associated numerical and event series ($\sigma = 0.4$)

Figure 2 shows and example of such series x and y. Let's consider other numerical time series y_k having larger values for either a subset or a superset of the events x. A good association metric should rank these time series such that y has the highest association with x.

2.3 Comparison of different metrics

We will start by studying usual metrics that could capture the association between time series. We will compare them based on the properties and examples defined in sections 2.1 and 2.2, and see how they can be generalized to apply to any time series.

2.3.1 Pearson Correlation Coefficient

The most common measure of correlation is the one defined by Pearson ρ defined by:

$$\rho_{x,y} = \frac{cov(x,y)}{\sigma_x \sigma_y} = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2}\sqrt{\sum_i (y_i - \overline{y})^2}}$$

To capture both correlated and anti-correlated time series we consider its square value ρ^2.

Properties It verifies all implications ($\Rightarrow$) of the above properties: a high Pearson correlation coefficient indicates that the knowledge of a series increases our knowledge of the other. However, it only captures linear correlation, so we could miss association between series with non-linear correlation (**P2.** $\not\Leftarrow$ and **P3.** $\not\Leftarrow$).

Categorical data In order to compute the Pearson correlation coefficient for the first example, we need the categorical data to have numerical values, we set $C_1 = 1$ and $C_2 = 0$.

Numerical data For the second example, the correlation score will be quite low even when there is a high association because the correlation is not linear. While the series are correctly ranked, the value of the correlation is hard to interpret and does not indicate how significant the association is.

2.3.2 Spearman and Kendall Correlation Coefficients

Spearman and Kendall both defined correlation metrics based on the ranking of elements rather than their exact value.

The Spearman coefficient is the Pearson correlation between the ranks of the elements in x and y, which for series with distinct elements is equivalent to:

$$r_s = 1 - \frac{6 \sum_i (rank(x_i) - rank(y_i))}{t(t^2 - 1)}$$

The Kendall coefficient computes the proportion of pairs (i, j) for which both series have a concordant order $x_i > x_j$ or $x_i < x_j$:

$$\tau = \frac{1}{t(t-1)} \sum_{i \neq j} sgn(x_i - x_j) sgn(y_i - y_j)$$

Properties While these two coefficients capture non-linear correlations, they still do not capture non-monotonous correlation, for instance if extreme values of x (positive or negative) correspond to high values of y (**P2.** $\not\Leftarrow$ and **P3.** $\not\Leftarrow$).

Categorical data In the first case, Spearman and Kendall correlations are equivalent to Pearson correlation because both series can only take 2 different values, so sorting them gives the same result as taking their absolute value.

Numerical data In the second case however, these coefficients give different results than the Pearson coefficient. They can give better results because they capture non-linear correlations, but might miss the distinction between outliers and relatively high values. They will bias towards series that are impacted by few events, as they will consider the association to be perfect if the events x correspond exactly to the largest values of y.

These two coefficients are interesting to consider in addition to the Pearson coefficient as they capture other kinds of correlations, but they are also hard to interpret and could miss some obvious associations.

2.3.3 Null hypothesis

Properties

Categorical data We can formulate the examples as null hypothesis problems. For the first example, the null hypothesis is:

H_0: Events in the time intervals $[0, t_1]$ and $[t_1, t]$ were drawn with the same probability p

The z-test associated with this hypothesis is

$$Z = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{t_1} + \frac{1}{t_2}\right)}}$$

We reject the null hypothesis if $Z > \alpha$. We note that Z is proportional to the Pearson factor as $Z = \sqrt{t_{tot}} * \rho$.

Numerical data For the second example, a hypothesis we could refute is that both groups (the values of y when x occurs and its values when x does not occur) were drawn from the same distribution. This should give good

results but requires a prior kind of distribution to be defined, in this case we can consider :

H_0: $y_{\|x=1}$ and $y_{\|x=0}$ were drawn from the same distribution $\mathcal{N}(\mu, \sigma)$

The z-test associated with this hypothesis is

$$Z = \frac{\mu_1 - \mu_2}{\sigma}\sqrt{t_{tot}}$$

where μ_1 and μ_2 are the average values of $y_{\|x=1}$ and $y_{\|x=0}$ respectively.

In both cases, the null hypothesis definition requires a prior knowledge of the data, but gives us a good way to interpret the association score. For instance, when $Z > \alpha_{0.99}$, we can reject the null hypothesis with 99% confidence, in other words the probability of occurrence if very likely to have changed at some point.

Mutual Information When considering categorical data, measures of association usually involve the information entropy. The metric that measures shared entropy is the mutual information. For categorical or discrete data, the mutual information is defined by

$$I_{x,y} = H_x + H_y - H_{x,y} = \sum_i \sum_j p(x_i, y_j) \log\left(\frac{p(x_i, y_j)}{p(x_i)p(y_j)}\right)$$

where H denote the entropies and p denotes probabilities. More specifically $p(x, y)$ is a joint probability function, and $p(x)$ and $p(y)$ are marginal probability distribution functions.

Its definition can be generalized to continuous data (see A):

The mutual information is

$$I = H_1 + H_2 - H_{tot}$$

where $H_i = -t_i \left(p_i \log p_i + (1 - p_i) \log(1 - p_i)\right)$

Properties

Categorical data

Numerical data

2.3.4 Conclusion

Ranking candidate time series

Theory Let's consider other candidate time series y_i with value C_1 during t'_1 and C_2 during $t'_2 = t_{tot} - t'_1$. A good association metric should rank y_i time series such that y_{t_1} has the highest association with x. In this example, all metrics studied here find the same optimal associated time series in theory, because $p'_1 - p'_2$ decreases for values of t'_1 close to t'_2, and $\frac{1}{t'_1}$ or $\frac{1}{t'_2}$ is very large for extreme values of t'_1, so $z' < z$.

Practice In practice, we used the Monte-Carlo method to compute the error between the expected best candidate and the actual one. We note that results are pretty good, even for small differences between p_1 and p_2 where it's even hard for a human eye to detect the change in probability. The main bias is towards extreme values of t'_1, when the random samples had an unusually high (or low) frequency at the beginning or at the end. However, using the Null Hypothesis method we can ignore the best candidate when the hypothesis is rejected, in which increases the accuracy of the model.

3 Application to real data

3.1 Preprocessing

Generalization

We can generalize the results above to any categorical data, and to events with any number of change of frequency. Using the null hypothesis, we can recursively split the event time series into phases where the frequency of occurrence is statistically constant. This can be used to preprocess event time series. The mutual information formula remains valid for any number of categories. It biases towards many categories, as the mutual information is always strictly positive. To balance this, we could penalize the mutual information such as defined in [A penalized mutual information criterion for blind separation of convolutive mixtures Mohammed El Rhabi, Guillaume Gelle, Hassan Fenniri, Georges Delaunay]. To keep our algorithm efficient enough to compare thousands of sub-second time series, we decided not to penalize the mutual information.

3.2 Data

3.3 Results

Conclusion

Appendix

A Mutual information generalization

We use Kraskov et al.'s idea [16] to estimate the mutual information between numerical time series drawn from an unknown distribution. If we used the same definition as for discrete data, we would systematically have a mutual information of 1, because no pair of point would have exactly the same value, so the knowledge of a time series would enable us to know the value of the other one. However, this behavior is not expected, as learning these correlation

The previous definition of mutual information was defined for categorical data. For numerical data, the joint probability of two variables is also well defined if the underlying distribution is known:

[insert formula here]

However, it is possible to approximate this underlying distribution by considering the k-nearest neighbors of each point. Given k, for each point P, we compute the distance ε to its k-nearest neighbor. If the underlying distribution μ was uniform, then the probability of being in the ball of center P and of diameter ε would be $\mu \times V$, where V is the volume of this ball. By deriving this expression, we get an approximation of the entropy:

[insert here]

And by extension of the mutual information between two numerical series. This expression is only valid for numerical data where all values are different (or at least where there are never more than $k-1$ elements with the same value), which can be circumvented by adding a small random noise.

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

[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

[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] Smyth, P. and Goodman, R.M. (1992)
An Information Theoretic Approach to Rule Induction from Databases, IEEE Transactions on Knowledge and Data Engineering, 4 (Aug. 1992), 301-316

[5] 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.

[6] 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.

[7] Goodman, R.M. and Smyth, P. (1998)
An information-theoretic model for rule-based expert systems, Int. Symposium in Information Theory, Kobe, Japan, 1988

[8] 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.

[9] 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

[10] 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.

[11] 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

[12] 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)

[13] 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

[14] 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

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

[16] Alexander Kraskov, Harald StÃ¶gbauer, and Peter Grassberger (2004)
Estimating mutual information, Phys. Rev. E 2004, 69: 066138. 10.1103/PhysRevE.69.066138

[17] 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

[18] 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

[19] 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.

[20] 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.

[21] 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