MESSY: Multivariate Event Sequence SummarY

Roel Bertens* and Jilles Vreeken°

Abstract

Recently, mining patterns from discrete time series data has seen increased interest [1, 2]. So far, however, although most real data of that type is multivariate, almost all research attention has solely been invested into univariate sequential data. Suppose that we want to analyze the data collected by multiple sensors, e.g. monitoring the movement and temperature of a bridge over time. With existing techniques for serial episode mining we will not find any correlation between the sensors when they are analyzed separately.

In this work we propose a framework that enables us to obtain succinct descriptions of multivariate event sequence data in terms of rich multivariate sequential patterns. To this end we allow sequence patterns to span over the alphabets of multiple sensors. With this strategy we aim to describe the whole dataset, consisting of multiple data streams, with only a small set of patterns (also called a code table). We follow the approach of [1], i.e. we employ the Minimum Description Principle [3] to identify that set of patterns that together describe the data most succinctly.

Next to the formalization of the properties of multivariate sequential patterns we will answer the following questions in more detail: how can we use MDL to score the quality of a set of patterns, how can we find good descriptions of a dataset given a set of patterns – allowing gaps in their occurrences – and how can we find good sets of patterns directly from data without having to first mine the complete set of frequent patterns.

The resulting framework can, besides a high level analysis of the data, also be used for other data mining tasks. We can, for example, use it for anomaly detection. That is, we can use a code table that is build for a data set to compute for each possible window in the data an anomaly score based on its compressed size using the code table. The proposed method only needs to cover the data once in order to be able to compute the outlier scores for all windows of all sizes. When we relate a window's outlier score to the average outlier score for all windows we can determine whether to regard this window as an outlier or not. Setting an outlier score threshold will then yield all corresponding outliers.

References

[1] Tatti, Nikolaj and Vreeken, Jilles. The Long and the Short of It: Summarising Event Sequences with Serial Episodes, In *ACM SIGKDD*, 2012.

[2] H. T. Lam, F. Mrchen, D. Fradkin, and T. Calders. Mining compressing sequential patterns. In *SDM*, 2012.

[3] Grünwald, Peter. The minimum description length principle. MIT Press, Cambridge, 2007.

*Speaker, PhD student at Universiteit Utrecht with the Algorithmic Data Analysis group of Arno Siebes
°Max Planck Institute for Informatics, Saarland University, and Cluster of Excellence MMCI