Interactive Time Series Clustering with COBRAS^{TS}

Toon Van Craenendonck, Wannes Meert, Sebastijan Dumančić and Hendrik Blockeel

KU Leuven, Department of Computer Science
{firstname.lastname}@kuleuven.be

Abstract. Time series are ubiquitous, resulting in substantial interest in time series data mining. Clustering is one of the most widely used techniques in this setting. Recent work has shown that time series clustering can benefit greatly from small amounts of supervision in the form of pairwise constraints. Such constraints can be obtained by asking the user to answer queries of the following type: *should these two instances be in the same cluster?* Answering "yes" results in a must-link constraint, "no" results in a cannot-link. In this paper we present an *interactive clustering system* that exploits such constraints. It is implemented on top of the recently introduced COBRAS^{TS} method. The system repeats the following steps until a satisfactory clustering is obtained: it presents several pairwise queries to the user through a visual interface, uses the resulting pairwise constraints to improve the clustering, and shows this new clustering to the user. Our system is readily available and comes with an easy-to-use interface, making it an effective tool for anyone interested in analyzing time series data.

Video and code are available at https://dtai.cs.kuleuven.be/software/cobras/

1 Introduction

Clustering is one of the most popular techniques in data analysis, but also inherently subjective [3]: different users might prefer very different clusterings, depending on their goals and background knowledge. Semi-supervised methods deal with this by allowing the user to define constraints that express their subjective interests [4]. Often, these constraints are obtained by querying the user with questions of the following type: *Should these two instances be in the same cluster?* Answering "yes" results in a must-link constraint, "no" in a cannot-link.

In this paper we present an *interactive clustering system* that exploits such constraints. The system is based on COBRAS^{TS} [2], a recently proposed method for semi-supervised clustering of time series. COBRAS^{TS} is suitable for interactive clustering as it combines the following three characteristics: (1) it can present the best clustering obtained so far at *any time*, allowing the user to inspect intermediate results (2) it is *query-efficient*, which means that a good

clustering is obtained with only a small number of queries (3) it is *time-efficient*, so the user does not have to wait long between queries. Given small amounts of supervision COBRAS^{TS} has been shown to produce clusterings of much better quality compared to those obtained with unsupervised alternatives [2].

By making our tool readily available and easy to use, we offer any practitioner interested in analyzing time series data the opportunity to exploit the benefits of interactive time series clustering.

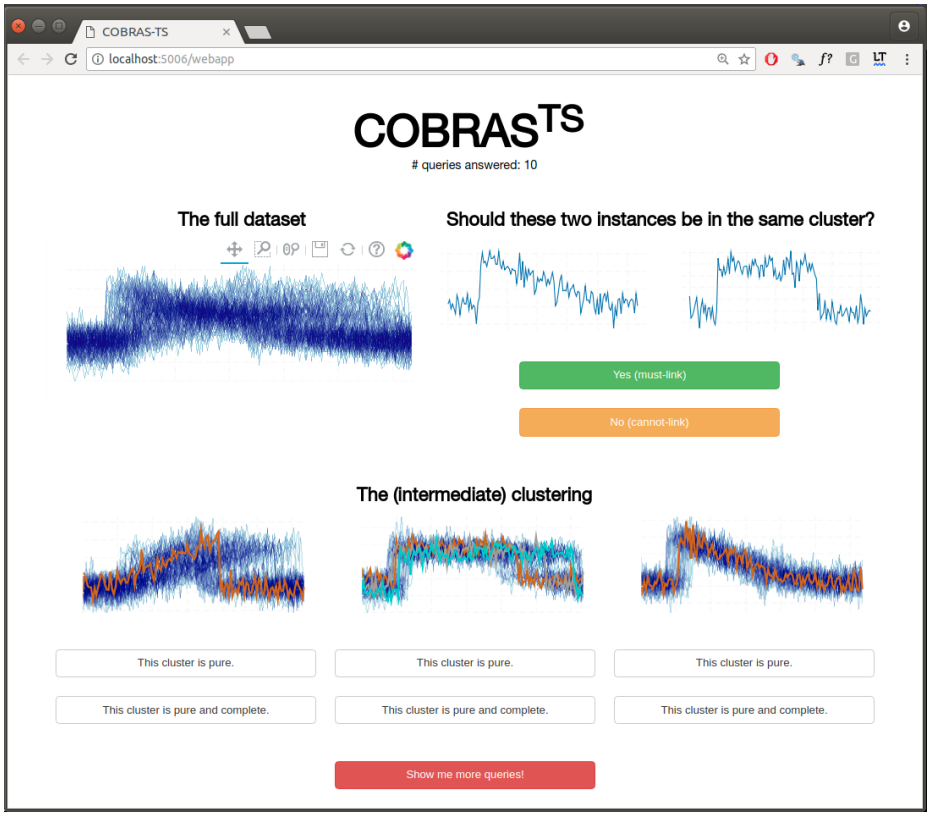


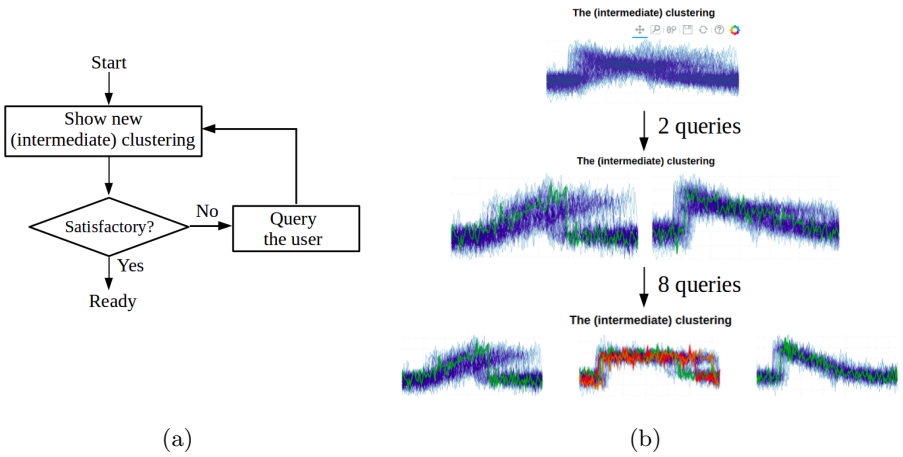
Fig. 1: Screenshot of the web application.

2 System description

The graphical user interface of the system is shown in Figure 1. On the top left the full dataset is shown, as a plot with all time series stacked on top of each other. On the top right, the system shows the querying interface. It presents the instances for which the pairwise relation is being queried, and two buttons that the user can click to indicate that these two instances should (not) be in the same cluster. On the bottom, the system shows the intermediate clustering.

This clustering is updated after every couple of queries. The main loop that is executed is illustrated in Figure 2(a): the system repeatedly queries several pairwise relations and uses the resulting constraints to improve the clustering, until the user is satisfied with the produced clustering.

Each time an updated clustering is presented, the user can optionally indicate that a cluster is either pure, or pure and complete. If a cluster is indicated as being pure, the system will no longer try to refine this cluster. It is still possible, however, that other instances will be added to it. If a cluster is indicated as being pure and complete, the system will no longer consider this cluster in the querying process: it will not be refined, and other instances can no longer be added to it. This form of interaction between the user and the clustering system (i.e. indicating the purity and/or completeness of a cluster) was not considered in the original COBRAS^{TS} method, but experimentation with the graphical user interface showed that that it helps to reduce the number of pairwise queries that is needed to obtain a satisfactory clustering.



(a) (b)

Fig. 2: (a) The interactive clustering loop. (b) Demonstration of clustering improvement as queries are answered.

The COBRAS^{TS} system is implemented as a web application that is run locally. It is open source and available online[1]. It is also available on PyPI, allowing installation with a single command[2].

[1] https://dtai.cs.kuleuven.be/software/cobras/

[2] `pip install --find-links https://dtai.cs.kuleuven.be/software/cobras/datashader.html pip cobras_ts[gui]`

3 Example run

Figure 2(b) shows the sequence of clusterings that is generated by the application for a sample of the CBF dataset [1]. It starts from a single cluster that contains all instances. After the user has answered two pairwise queries, the system presents an updated clustering containing two clusters. The first cluster contains mainly upward clusters, whereas the second cluster contains a mixture of downward and horizontal patterns. As this clustering is not satisfactory yet, more pairwise queries are answered. After 8 more queries, the system again presents an improved clustering. This time, the clustering clearly separates three distinct patterns (upward, horizontal and downward). While distinguishing between these three types of patterns is easy for a user, it is difficult for most existing clustering systems; none of COBRAS$^{\text{TS}}$'s competitors is able to produce a clustering that clearly separates these patterns [2].

4 Conclusion

The proposed demo will present a readily available and easy-to-use web application for interactive time series clustering. Internally, it makes use of the recently developed COBRAS$^{\text{TS}}$ approach. The application enables users to exploit minimal supervision to get clusterings that are significantly better than those obtained with traditional approaches.

Acknowledgements

Toon Van Craenendonck is supported by the Agency for Innovation by Science and Technology in Flanders (IWT). This research is supported by Research Fund KU Leuven (GOA/13/010), FWO (G079416N) and FWO-SBO (HYMOP-150033).

References

1. Y. Chen, E. Keogh, B. Hu, N. Begum, A. Bagnall, A. Mueen, and G. Batista. The UCR time series classification archive, July 2015. http://www.cs.ucr.edu/~eamonn/time_series_data/.
2. T. Van Craenendonck, W. Meert, S. Dumančić, and H. Blockeel. COBRAS-TS: A new approach to Semi-Supervised Clustering of Time Series. *https://arxiv.org/abs/1805.00779, under submission*, May 2018.
3. U. von Luxburg, R. C. Williamson, and I. Guyon. Clustering: Science or Art? In *Workshop on Unsupervised Learning and Transfer Learning, 2014*.
4. K. Wagstaff, C. Cardie, S. Rogers, and S. Schroedl. Constrained K-means clustering with background knowledge. In *Proc. of ICML 2001*.