Clustering and Co-clustering of Multivariate Time-series based on Complex Networks

Amirhossein Jazayeri[1], Mandani Ntekouli[1], and Gerasimos Spanakis[1]

Department of Advanced Computing Sciences, Maastricht University, Maastricht, The Netherlands a.jazayeri@student.maastrichtuniversity.nl,{m.ntekouli, jerry.spanakis}@maastrichtuniversity.nl

1 Introduction

Ecological Momentary Assessment (EMA) is a real-time data collection method that captures individuals' mood states, behaviors, and well-being over time. However, the complexity of EMA data, characterized as high-dimensional time series data, presents significant analytical challenges. To exploit the full potential of such data, advanced machine learning techniques, such as clustering and biclustering (also known as co-clustering), offer a robust framework for uncovering hidden patterns and relationships within EMA datasets [1,2]. Clustering facilitates the identification of homogeneous group profiles among individuals, while biclustering extends this analysis by simultaneously uncovering patterns that link both individuals and specific features, revealing commonalities and differences across multiple dimensions [7].

The primary objective of this thesis is twofold: first, to explore and evaluate various clustering and biclustering methods, and second, to investigate the impact of different data transformations on these methods. To apply most of the traditional clustering algorithms, it is essential to represent time-series in a reduced-dimensional space. More specifically, latent vectors through generative models and network-based representations by applying various distance metrics on EMA data. By integrating such time-series transformations, the study aims to evaluate the effectiveness of clustering and biclustering, focusing on both cluster quality and forecasting performance.

2 Methodology

The methodology of this study is structured around three key components, including data transformations, clustering, and biclustering methods, as well as their evaluation. Starting from the data transformations, we specifically aimed at reducing the dimensionality of high-dimensional EMA time-series datasets. To achieve this, first, generative models were utilized to transform high-dimensional data into lower-dimensional latent vectors[5]. Alternatively, complex network-based representations were constructed from original data and latent data, allowing the use of community detection algorithms [10]. These algorithms identify communities or clusters within the network, representing groups of individuals with similar behavioral patterns.

In addition to community detection methods, other traditional clustering methods, such as k-means and hierarchical clustering, were explored. Furthermore, this thesis explores biclustering techniques, including Spectral Biclustering and Bipartite Spectral Graph Partitioning to find biclusters that link specific groups of individuals with distinct sets of features [9,8].

To evaluate the effectiveness of these methods, a combination of intrinsic evaluation metrics and a downstream forecasting task is used. Intrinsic evaluation metrics, such as the silhouette score, the modularity score [12] and the Davies-Bouldin index [11], are used to assess the quality of the derived clusters separately for each method. To, subsequently, provide an overall comparison of all, a downstream forecasting task was conducted. This task evaluates the effectiveness of different approaches by comparing the performance of the cluster-derived models. This forecasting task assesses the practical utility of the (bi)clustering methods by predicting the progression of EMA features.

3 Experimental Results

According to the experiments conducted, various clustering and biclustering methods along with different data transformations were investigated. Among clustering methods, applying traditional clustering algorithms to the generated latent vectors resulted in higher silhouette scores (on average 0.14, and 0.46 in the best experiment). These findings suggest that dimensionality reduction through generative models effectively preserves key features while minimizing noise, leading to the formation of more coherent clusters. In contrast, biclustering methods showed better performance when applied directly to the original multivariate time-series data, reaching 0.05.

In the downstream forecasting task, network-based clustering methods exhibited superior performance, achieving an average MSE score of 0.72 when applied to latent vectors and 0.73 when applied to raw data. On the contrary, the lowest performance was observed with biclustering, yielding an MSE of 0.82, despite offering richer insights on both individual and feature levels.

4 Conclusions

This thesis examines the effectiveness of various clustering and biclustering techniques applied to different transformations of EMA data. According to all the conducted experiments, clustering consistently outperformed biclustering methods, in terms of both intrinsic evaluation metrics and forecasting performance. Moreover, data transformations deriving latent vectors through generative models, as well as complex networks through distance measures demonstrated superior results, further enhancing the quality and predictive power of the clustering outcomes. For future work, more advanced modeling approaches should be explored, taking advantage of the full potential of the biclustering results and enhancing their performance.

References

1. Ntekouli, M., Spanakis, G., Waldorp, L., Roefs, A.: Clustering Individuals Based on Multivariate EMA Time-Series Data. Psychometrika (September 2022)
2. Ntekouli, M., Spanakis, G., Waldorp, L., Roefs, A.: Model-based Clustering of Individuals' EMA Data. Psychometrika (September 2022)
3. Madeira, S.C., Oliveira, A.L.: Biclustering algorithms for biological data analysis: a survey. IEEE/ACM Transactions on Computational Biology and Bioinformatics 1(1), 24-45 (2004)
4. Ferreira, L.N., Zhao, L.: Time series clustering via community detection in networks. Information Sciences 326, 227-242 (2016)
5. Todo, W., Laurent, B., Loubes, J.M., Selmani, M.: Dimension Reduction for time series with Variational AutoEncoders. arXiv (2022)
6. Desai, A., Freeman, C., Wang, Z., Beaver, I.: TimeVAE: A Variational Auto-Encoder for Multivariate Time Series Generation. CoRR (2021)
7. Busygin, S., Prokopyev, O., Pardalos, P.M.: Biclustering in data mining. Computers & Operations Research 35(9), 2964-2987 (2008)
8. Dhillon, I.S.: Co-clustering documents and words using bipartite spectral graph partitioning. Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, 269-274 (2001)
9. Madeira, S.C., Oliveira, A.L.: Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **1**(1), 24-45 (2004). doi:10.1109/TCBB.2004.2
10. Ferreira, L.N., Zhao, L.: Time series clustering via community detection in networks. *Information Sciences*, **326**, 227-242 (2016). doi:10.1016/j.ins.2015.07.046
11. Davies, D.L., Bouldin, D.W.: A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **PAMI-1**(2), 224-227 (1979). doi:10.1109/TPAMI.1979.4766909
12. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. *Physical Review E*, **69**(2), 026113 (2004). doi:10.1103/PhysRevE.69.026113