A Matter of Time: how to visually analyze multivariate (and multidimensional) data, irregularly sampled and having multiple granularities

Alessio Bertone

Department of Information and Knowledge Engineering (ike),

Danube University Krems,

dr.karl-dorrek-straße 30, 3500 krems (austria),
alessio.bertone@donau-uni.ac.at

Abstract. Multi Relational Data Mining searches for patterns that involve multiple tables from a relational database. In order to avoid the generation of a huge relation involving all of the attributes and the loss of information, including essential semantic information represented by the links in the database design, it aims to discover knowledge directly from relational data. Time is an intrinsic data dimension in many domains, which is different from any other dimension. However, the complexity of dealing with many temporal data at a time and to contemporary provide a visual aid to users, represents a challenge still open. This work represents an attempt to exploit the ideas coming from Multi Relational Data Mining (and Temporal Data Mining) in order to be able to analyze multivariate and multigranular time-oriented data, and to discover patterns, trends and relations among data previously unknown. Moreover it presents the preliminary steps performed so far and the steps to be performed to present, interact and communicate visual overviews of the available data and the results of the performed analysis.

Keywords: *temporal data, temporal data mining, multi relational data mining, information visualization*

Description of the Research Problem

Time is an intrinsic data dimension in several domains such as medical records, financial data, or biographical data, and plays a central role for the tasks of revealing trends, as well as identifying patterns and relationships in the data. As a matter of fact, exploring trends, patterns, and relationships are particularly important tasks when dealing with time-oriented data and information. The main reason is that in contrast to other quantitative data dimensions that are usually “flat”, time has an internal semantic structure which increases its complexity. Time is highly structured in the sense of different granularities, different calendar systems, different time zones, and irregularities. Moreover, the hierarchical structure of granularities in time (e.g., minutes, hours, days, weeks, months) comprises different forms of divisions (e.g., one hour is composed by 60 minutes, one day is composed by 24 hours, and so on) which is in contrast to the decimal system, and granularities are combined to form calendar systems. Thus a multigranular dataset contains different granularities and this characteristic has to be taken into account during every kind of analysis.

Data Mining can be defined as a step of the Knowledge Discovery process which consists of the search for patterns of interest in a particular representational form or a set of such representations, such as classification rules and trees, clustering or association rules [1]. The adoption of time in the data mining tasks leads to what is known as Temporal Data Mining.

To incorporate time in the data mining process gives us the ability to detect activities rather than just states, i.e. the ability to find behavioural aspects of communities of objects rather then just describing their states at a certain point in time.

Temporal Data Mining covers several different topics, such as the discovery of frequent, or interesting, sequences in a time series (e.g., finding customers whose spending pattern over time are similar to a given spending profile), the detection of relations between different sequences (e.g., in the stock market analysis we are interested in finding rules that describe relations between different stocks) and so on [2], [3].

(Temporal) Data are mainly stored into (temporal) relational databases. In a relational database, multiple relations are linked together via entity-relationship links. Many classification approaches (e.g., neural networks, support vector machines) can only be applied to data represented in single, "flat" relational form, that is, they expect data in a single table. However, many real-world applications, such as fraud detection, financial applications, and medical data analysis, require to take decisions based on information stored in multiple relations in a relational database. Thus, the importance of Multi Relational Data Mining has been recently growing.

Multi Relational Data Mining methods [4], [5] search for patterns that involve multiple tables (relations) from a relational database. In order to avoid the generation of a huge, undesirable universal relation involving all of the attributes and the loss of information, including essential semantic information represented by the links in the database design, Multi Relational Data Mining aims to discover knowledge directly from relational data (differently from propositional approach which assumes that the data reside in a single table, and, using joins and aggregations, converts multiple relational data into a single flat data relation).

Several Information Visualization methods support the visualization of temporal data [6], [7]. However, apart from few exceptions (for instance, [8]), these methods are especially tailored for a specific application domain, are bound to a particular feature of time, and they mostly treat time just as any other data dimension (attribute).

As the application of data mining techniques to time series databases is a relatively new field [9], the combination of automatic/ semi-automatic data mining techniques, time series and visualization, appears as even newer and challenging. Some contributions give more emphasis to time series and data mining [10], [11], others also concentrate on the visualization part (e.g., VizTree [12]) or propose a combination of visual data mining and time series (e.g. , Parallel Bar Chart [13], TimeSearcher technique [14]). However each of them presents some lack in the visualization part, leaves the mining task up to the user, or requires a strong expertise in the application field.

This work aims to exploit the ideas coming from Multi Relational Data Mining (and Temporal Data Mining) in order to be able to analyse multivariate, time-oriented data (namely several time series) having different temporal granularities. Moreover, it aims to investigate how to adapt existing visualization techniques (or possibly creating new ones) in order to provide visual overviews of the available data, to allow the user to intuitively interact with the given visualizations and to communicate the results of the analysis performed, plausibly adjusting the parameters of the analysis.

Research Questions

Main Questions

How can Multi Relational Data Mining and Temporal Data Mining be adapted and used to visually analyze multivariate time-oriented data having multiple granularities in order to discover of patterns, trends and relations among data (previously unknown)?

Sub-Questions

- Does it make sense to introduce temporal aspects into Multi Relational Data Mining?
- How to deal with different time granularities of multiple time series?
- What constraints are needed to perform a meaningful granularity alignment / synchronization among the time series?

- How to graphically represents not only the time series (overview), but also the granularity synchronization (parameter tuning), allowing an easy interaction (interaction and presentation of the results)?
- How to communicate results and allow easy interaction with the proper visualizations?

The Chosen Approach

Differently from having a single flat data relation, Multi Relational Data Mining tasks (e.g., multi relational clustering, multi relational classification) search for patterns that involve multiple tables and relations from entity relational databases.

As the user is usually interested in finding unknown relations among data (time series), relevant trends and patterns, a similar approach can be adopted to deal with several time series, having multiple granularities. However, differently from the entity relational case, no relations are available to navigate this huge amount of time series, thus the only possibility to have a coherent analysis seems to be the navigation by exploiting the intrinsic temporal features of the data themselves.

Basically, the idea is to provide the user with instruments to perform a sort of alignment or synchronization of the available granularities: once a reference granularity (target granularity) has been chosen, all of the other granularities will be converted to the coarser or finer one, in order to provide a coherent temporal overview of the available data into an “alignment table” (that is, each time series will be expressed according to the chosen one, e.g. days, hours).

In order to perform the granularities’ synchronization, many constraints are needed: for instance, even if this could sound simplifying, one can assume that the patient temperature sampled each hour is constant second by second and minute by minute. This is valid in the case of conversion to a finer granularity. On the contrary, in case of conversion to a coarser granularity, other assumptions are needed: for instance, if a time series T1 is sampled at a level of hours, and a time series T2 at a level of days (which in this case is the “target granularity”), one cannot assume to directly convert hours to days, since it does not make sense. However, what can be applied is a sort of analysis of the trend within the finer granularity in order to obtain a “summarization” of the values (e.g., the mean) according to the target granularity. Moreover to avoid the loss of information, for instance, some visual hints (e.g., colours) to the graphical representation can be added to indicate relevant trends within the time series.

A similar idea could be adopted to merge the potentialities of existing Temporal Data Mining and Information Visualization techniques, in order to perform an analysis of multivariate and multigranular time-oriented data.

From a mere visualization point of view, the adopted methods (in case both of Multi Relational Data Mining and of Temporal Data Mining), on the one hand, should result transparent to the user as they are within a visualization, on the other hand, they should be easily interpretable in order to better address the user needs.

Work performed so far

The first step of my thesis work was to investigate the State of the Art of Temporal Data Mining and Multi Relational Data Mining for temporal data. In the former case, a lot of literature was available, and has been organized to provide an overview of the main aspects of the field. On the contrary, in the latter case, some recent works tried to face the above problem, but not using the proposed approach, rather trying for instance to create regression model to capture the relations between customers and products, or building predictive model (e.g. decision trees), above all in medical applications.

Parallel to the investigation of Multi Relational Data Mining methods, many databases containing datasets have been explored. The aim is to find some interesting examples of multivariate (and multidimensional) data, having multiple granularities on which to test the reliability of the proposed approach, to better address some theoretical hypothesis and possibly to improve it.

Next working steps

The work to be performed can be organized into tasks or phases as follows (see the Framework in Fig. 1):

- *(Formalization of) Granularity Alignment:* In the first phase of the work, the formalization of the rules to perform a correct granularity alignment, from a coarser granularity to a finer one and vice versa will be developed. These rules represent the constraints which have to be taken into account to compare many time series at a time, having different granularities.

- *Pattern Extraction*: Once the alignment has been performed, there is the need to retrieve all the relevant patterns within the alignment table. In this case, there are several possibilities, such as the adoption of a technique to extract the so called “tentative motifs”, that is, the most relevant sub-sequences, used as input for the identification of frequent patterns (the “motifs”) [15], [16], [17].

- *Pattern Selection*: In order to compare the patterns, the definition of a similarity measure is required. This similarity measure has to deal with the qualitative and quantitative features of the data, therefore it should be composed of two different parts: the former to evaluate the qualitative features (e.g., Hamming distance) and the latter to evaluate the quantitative features (e.g., Longest Common Subsequence). A step to choose the most relevant patterns could then be performed. In this case, a clustering algorithm could group all the patterns and then provide a small amount of relevant patterns, one representative for each cluster, according, for instance, to the number of clusters, the algorithm chosen, or possibly to some user requirements.

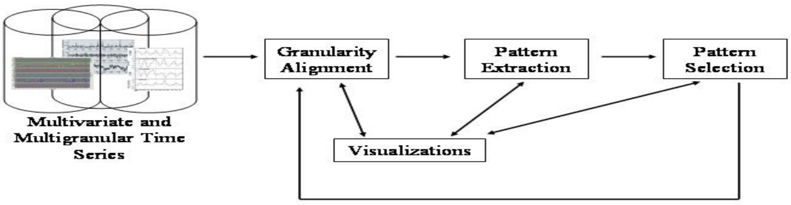

Fig. 1: The proposed framework

- *Visualization*: Finally, parallel to most of the above steps, one or more existing visualizations will be extended or new ones will be designed to ease all the phases of the analysis. For instance, a visualization may facilitate the task of the granularity alignment, by allowing to select the "target granularity" and to display the results obtained from the conversion from a finer granularity to a coarser one. Similarly, another visualization may display the retrieved relevant patterns in order to clearly outline the discovered and previously unknown relations and possibly to (re-)iterate partially or completely the process from the very beginning.

- *Evaluation*: Concerning the evaluation of the approach, firstly, a sample test set will be used to evaluate results obtained by the adoption of the proposed approach. Secondly, the approach will be tested with real data previously chosen.

Conclusion

The adoption of Multi Relational Data Mining (and Temporal Data Mining) in order to be able to analyse multivariate, time-oriented data having different temporal granularities is a challenging task and the proposed approach represents an attempt to cope with it.

An important aspect of my thesis work is the ambition to provide the instruments to deal with many multigranular, multidimensional temporal data at a time and, at the same time, to visually facilitate the users during the whole analysis task. Even if this thesis is at the early stages, the possibilities of successful results are very promising.

References

[1] U. Fayyad, P. Smyth, G. Piatetetsky-Shapiro, R. Uthurusamy, Advances in knowledge discovery and data mining, AAAI Press/The MIT Press, 1996.

[2] C. M. Antunes and A. L. Oliveira. Temporal data mining: An overview. In KDD Workshop on Temporal Data Mining, pages 1--13, San Francisco, CA, 26 August 2001.

[3] S. Laxman and P.S. Sastry, A survey of temporal data mining, Sadhana Vol. 31, Part 2, April 2006, pp. 173–198.

[4] S. Dzeroski. Multi-relational data mining: An introduction. SIGKDD Explorations, 5: 1-16, 2003.

[5] J. Knobbe, H. Blockeel, A. Siebes, and Van der Wallen D. Multi-relational data mining. In Proceedings of Benelearn 99, 1999.

[6] Müller, W. and Schumann, H., Visualization Methods for Time-Dependent Data - an Overview, Proceedings of Winter Simulation (WSC'03), New Orleans, 2003.

[7] Silva, S.F. and Catarci, T., Visualization of Linear Time-Oriented Data: A Survey, Proceedings of the First International Conference on Web Information Systems Engineering, Hong Kong, 2000.

[8] Aigner, W., Miksch, S., Müller, W., Schumann, H., Tominski, C.: Visualizing Time-Oriented Data - A Systematic View, Computers & Graphics, in press, 2007.

[9] E. Keogh, S. Lonardi, and W. Chiu, Finding Surprising Patterns in a Time Series Database In Linear Time and Space. In the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. July 23 - 26, 2002. Edmonton, Alberta, Canada, pp.550-556, 2002.

[10] F Guil, JM Juarez, R Marin, Mining Possibilistic Temporal Constraint Networks: A Case Study in Diagnostic Evolution at Intensive Care Units, Proceedings of IDAMAP Workshops, August 25-26, 2006, Verona (Italy).

[11] S Badaloni, M Falda, Mining Temporal Characterization of Ill-known Diseases, Proceedings of IDAMAP Workshops, August 25-26, 2006, Verona (Italy)

[12] J. Lin, E. Keogh, S. Lonardi, J.P. Lankford, D.M. Nystrom, Visually Mining and Monitoring Massive Time Series. In proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Seattle, WA, Aug 22-25, 2004.

[!3] Chittaro L., Combi C., Trapasso G., Data Mining on Temporal Data: a visual Approach and its Clinical Application to Hemodialysis, Journal of Visual Languages and Computing, vol.14, no.6, pp.591-620, December 2003

[14] H. Hochheiser, Interactive Querying of Time Series Data, in CHI ’02, extended abstracts on Human Factors in Computing Systems. ACM Press, pp. 552–553, 2002

[15] Staden, R. (1989). Methods for discovering novel motifs in nucleic acid sequences. Comput. Appl. Biosci., Vol. 5(5). pp 293-298.

[16] Tompa, M. and Buhler, J. (2001). Finding motifs using random projections. In proceedings of the 5th Int’l Conference on Computational Molecular Biology. Montreal, Canada, Apr 22-25. pp 67-74.

[17] Lin, J., Keogh, E., Patel, P. & Lonardi, S. (2002). Finding Motifs in Time Series . In proceedings of the 2nd Workshop on Temporal Data Mining, at the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Edmonton, Alberta, Canada. July 23-26.