Semana de Engenharia 2010
Guimarães, 11 a 15 de Outubro

TIME SERIES DATA MINING

Nuno Constantino Castro
CCTC – Department of Informatics
University of Minho
E-mail: castro@di.uminho.pt

KEYWORDS
Data Mining, Time series, Motif discovery, Motif evaluation, Statistical Significance.

INTRODUCTION

Data Mining or Knowledge Discovery in Databases (KDD) is an important area of computer sciences. The relevance of this area is due to the enormous quantity of information daily produced by different sources, for instance the web, biological processes, finance, the aeronautic industry, retail, and telecommunications data. A considerable amount of this information represents temporal events which are typically stored in the form of time series. There are several phenomena expected to be identified among databases of this type, namely through motif (pattern) discovery, classification, clustering, query by content, abnormality detection, and forecast of property values.
We focus particularly on the area of time series motif discovery (Lin and Keogh 2002) , also known as the extraction of recurrent patterns. These patterns are relevant because they summarise the time series of a domain and help the domain expert understand the database at hand (Ferreira et al. 2006). Figure 1 shows one example of such type of pattern in the context of electroencephalogram (EEG) time series. This specific motif is detected in three different time series in the database.

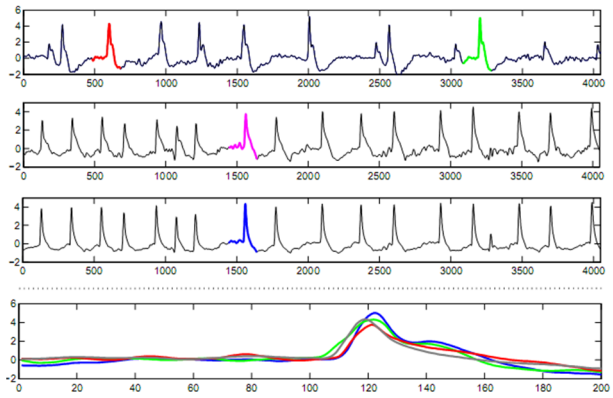

Figure 1: Example of a motif in EEG time series. Above: in its original context. Below: aligned in the same referential.

The remainer of this document is organized as follows. First, we propose a new algorithm for the task of motif discovery in time series databases. Then, we introduce a novel framework that allows to evaluate the statistical significance of time series motifs.

MOTIF DISCOVERY IN TIME SERIES

The proposed algorithm attempts to tackle the limitations we have found in the existing methods for the task of motif discovery in time series databases. They only find motifs in univariate time series; do not scale well due to the expensive random disk accesses; assume the data can fit into main memory; and present a high number of unintuitive parameters. Our algorithm can find motifs in multivariate time series; is a disk-aware algorithm - performs only one disk pass over the time series database (and therefore is suitable for streaming datasets); allows to adjust the amount of memory to use according to the devices that will run the algorithm; has a small number of parameters; and allows to find motifs with different granularities, which allows to navigate in the top frequent motifs structure. We perform one traversal over the time series database and discretise each time series using the iSAX representation technique (Shieh and Keogh 2008). During the previous step we maintain and update the count of each iSAX word in a hash table. After the traversal, we aim to have the top-K frequent patterns in main memory. Then, starting with a low motif resolution, we expand the motifs to a higher resolution and update the counts accordingly. That means to consecutively double the iSAX resolution of each top discovered pattern, in order to achieve a motif hierarchical structure. This step allows us to navigate in the top-K motifs structure, providing insight on the time series database to the user. The algorithm has the option to limit the amount of memory to use in case memory restrictions exist. Devices such as sensors or mobile devices are examples of this scenario. In that case, we provide an approximation of the Top-K patterns using the Space Saving algorithm (Metwally et al. 2005) applied to the frequent time series patterns scenario. Our approach as been experimentally tested in both real and synthetic datasets scenarios.

STATISTICAL SIGNIFICANCE OF TIME SERIES MOTIFS

Since the motif discovery problem formulation in 2002 (Lin et al. 2003), a large number of proposals on how to extract motifs from a time series database have been introduced. Surprisingly, most of these proposals don't focus on how to evaluate the extracted motifs. Returned motifs are typically (subjectively) evaluated by humans because they are application dependent and not previously labeled (motif discovery is unsupervised). In practice, this is unfeasible. Datasets are often large and motif mining algorithms return a prohibitively large number of patterns. To only present to the expert the most frequent motifs is not an interesting approach, as frequent patterns are not necessarily the most interesting ones. Many frequent patterns are spurious, trivial or simply expected: they are not meaningful to the user. For example, in a randomly generated database containing 10000 time series of length 1024, the top motif has 164 instances, and the motif count average is 5 (Castro and Azevedo 2010).

Statistical tests have been successfully applied to other pattern mining problems. For example, in bioinformatics they have been used to detect DNA segments with significantly unexpected frequency (Ferreira and Azevedo 2007); in networks, to find significant subgraphs (Ferreira et al. 2006). They aim to answer the following question: "Can this pattern be observed so many (or few) times just by chance?" The observed count (frequency) of a pattern is typically compared to its expected count. This difference is then statistically analysed. However, this approach can not be directly applied to time series data. It is not clear how to calculate the expected frequency of a given section of the series.

To leverage the wealth of algorithms available for symbolic data (DNA sequences, proteins, text, etc.), we use a symbolic definition of time series motifs. We estimate the probability of occurrence of a word (motif) using Markov Chain models. In these models, the probability of a motif is estimated according to its subword count. Given a motif, we compare the difference between its observed count and estimated expected count in terms of statistical significance. Namely, we calculate the p-value of this difference, aiming to answer whether we can observe such a count solely by chance. After performing this calculating on the extracted motifs, they can be ranked according to their p-value in order to assess their significance.

Our contributions are twofold: to provide a framework to assess the statistical significance of time series motifs, and compare the performance of several statistical hypothesis tests on motif extracted from real datasets. This allows time series data mining practitioners to properly evaluate the motifs extracted from their data, and researchers to properly evaluate the output of motif discovery algorithms using statistical significance.

FUTURE WORK

Further work includes using time series motifs as "building-blocks" to other time series data mining tasks. Our intuition is that motifs are meaningful patterns which are characteristic of a particular application/domain. Since they can be used to describe/summarize the time series database, they can be used as the basis for other data mining tasks such as classification, abnormality detection and forecasting. We plan to propose a method to perform one of the previously referred tasks, using the best motifs in the database as features.

REFERENCES

Shieh, J. and E. Keogh. 2008. "iSAX: Indexing and Mining Terabyte Sized Time Series.", in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 623–631.

Metwally, A., Agrawal, D. and A. Abbadi. 2005. "Efficient Computation of Frequent and Top-k Elements in Data Streams.", in *Proceedings of the 10th International Conference on Database Theory*, pp. 398-412.

Lin, J., Keogh, E., Lonardi, S. and P. Patel. 2002. "Finding Motifs in Time Series.", in *Proceedings of the 2nd Workshop on Temporal Data Mining*, pp. 53-68.

Ferreira, P., Azevedo, P., Silva, C. and R. Brito. 2006. "Mining approximate motifs in time series.", in Discovery Science. Springer, pp. 89–101.

Ferreira, P. and P. Azevedo. 2007. "Evaluating Protein Motif Significance Measures: A case study on Prosite Patterns.", in *IEEE Symposium on Computational Intelligence and Data Mining*. CIDM, pp. 171–178.

Castro, N. and P. Azevedo. 2010. "Multiresolution Motif Discovery in Time Series", in *Proceedings of the Tenth SIAM International Conference on Data Mining*. SIAM, pp. 665–676.

He, H. and A. Singh. 2006. "Graphrank: Statistical modeling and mining of significant subgraphs in the feature space", in *Proceedings of the Sixth International Conference on Data Mining*, pp. 885–890.

Boeva, V., Clément, J., R´egnier, M., Roytberg, M. and V. Makeev. 2007. "Exact p-value calculation for heterotypic clusters of regulatory motifs and its application in computational annotation of cis-regulatory modules.", in *Algorithms for molecular biology*, vol. 2, no. 1, p. 13.

AUTHOR BIOGRAPHY

NUNO C. CASTRO was born in Porto, Portugal and went to the University of Minho, where he studied

computer science and obtained his degree in 2006. He worked for a couple of years for Nokia Siemens Networks in data mining R&D before starting his PhD work in Time Series Data Mining in 2007.