
Machine Learning, 58, 103–105, 2005

Guest Editorial

Introduction

It is the nature of humans to measure things, and with rare exceptions, these things change over time. Recording these measurements leads to a proliferation of time series data. Time series data are thus ubiquitous; large volumes of time series data are routinely created in virtually all domains of human endeavor; examples include gene expression data, space telemetry, music, electrocardiograms, electroencephalograms, sensors, gait analysis, etc.

Although statisticians have worked with time series for more than a century, many of their techniques hold little utility for researchers working with time series databases. The most obvious explanation of this is the problem of scaling these techniques to the massive datasets routinely encountered. Consider this motivating example, the NASA Earth Observation System (EOS), transmits up to 50 GB of data to earth every hour, and most of this data is time series. A striking fact about these massive datasets is that the vast majority of the data will never be processed or examined by an algorithm, much less a human expert. The central tenet of the emerging field of data mining is that such databases are potentially repositories, not just of data, but also of knowledge. More concretely, data mining has been defined as the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.

This special issue is devoted to papers that advance the state of the art in time series data mining. Three of the papers in this issue consider the classic machine learning problems of classification and clustering.[1–3] Two of the papers describe similarity search in large time series databases.[4,5] The remaining papers span the space of time series data mining applications, including rule mining,[6] anomaly detection,[7] and the discovery of repeated patterns in time series.[8]

It is interesting to note that with the exception of indexing, research into the tasks enumerated above predate not only the decade old interest in data mining, but computing itself. What then, are the essential differences between the classic and the data mining versions of these problems? The key difference is simply one of size and scalability; time series data miners routinely encounter datasets that are gigabytes in size. As a simple motivating example, consider hierarchal clustering. The technique has a long history and well-documented utility. If however, we wish to hierarchically cluster a mere million items, we would need to construct a matrix with 10^{12} cells, well beyond the abilities of the average computer for many years to come. A data mining approach to clustering time series, in contrast, must explicitly consider the scalability of the algorithm. Such scalability is often achieved by using some high level abstraction of the data, and the papers in this special issue illustrate this by using a great variety of time series representations, including Adaptive Piecewise Linear Approximations,[4] bit-level approximations,[2] symbolic approximations,[8] and Piecewise Linear Approximations.[5]

In the rest of this introduction, we describe these research trends in the context of the papers that appear in this special issue.

Time series similarity search

Time series similarity search is an important tool in its own right, allowing users to interactively explore large databases and informally test hypotheses (for example, "After we see a double peak, we are likely to see a sharp spike"). In addition, similarity measurement is an important subroutine in all clustering algorithms and many classification, prediction, association detection, summarization, motif discovery, and anomaly detection algorithms. In this special issue, Vlachos et al. consider techniques to efficiently index multidimensional time series (i.e. trajectories).[5] They introduce a novel similarity measure that allows flexibility in the time axis, and also allows some sections of a query to be don't-cares. Similar measures had been considered before, but the authors are the first to show how such measures can be efficiently indexed. Shou et al. also consider a similarity measure that allows flexibility in the time axis, Dynamic Time Warping (DTW).[4] Here, the authors introduce a technique to speed up DTW similarity search.

Classification

While classic machine learning algorithms can be applied to time series classification. Time series datasets differ from other domains in that we typically encounter high dimensionality, high feature correlation, and high levels of noise. For this reason, much of the research in this area focuses on the extraction of low dimensional, uncorrelated high quality features. Mierswa and Morik consider this problem in the domain of music.[1] They point out that the classic tools of Euclidean distance and DTW[4] have little utility for music classification because "*The* i^{th} *value of a favorite song has no correspondence to the* i^{th} *value of another favorite*, *even if relaxed to the* $(i \pm n)^{\text{th}}$ *value*".[1] To create high quality features they propose a genetic programming approach to the automatic construction of data transformations.

Kadous and Sammut apply the classic machine learning idea of constructive induction to the problem.[3] They present a method aimed at learning tasks involving multivariate time series data, such as sign language recognition. They propose to expand the scope of attribute-value learning to domains with instances that have some kind of recurring substructure, such as strokes in handwriting recognition, or local maxima in time series data. These "metafeatures" are defined by the user, but are extracted automatically and are used to construct attributes.

Bagnall and Janacek consider the problem of clustering and classifying time series.[2] They empirically and theoretically show an original and surprising result. They demonstrate that the simple procedure of clipping the time series (discretizing to a single bit per time instance) reduces memory requirements and significantly speeds up clustering *without* decreasing clustering accuracy. They further show that clipping increases clustering accuracy when there are outliers in the data.

Exploratory data mining

Much of the previous work on time series data mining (including the work considered above) has focused on similarity search and supervised leaning, in other words, the efficient discovery of *known* patterns in a database. However, several papers in this special issue address a more difficult problem, the detection of previously *unknown* patterns and relationships in time series.

Hetland and Satrom consider several different mining tasks, including sequence prediction, unsupervised mining of interesting rules and discovering connections between separate time series.[1] While these tasks have all been (separately) considered before, the authors create models and rules that are (at least in principle) human-readable. This is very important since data mining is often touted as an interactive and iterative process.

Begnum and Burgess consider the problem of anomaly detection in computer systems. They note that subtle (possibly non-linear) correlations between locally averaged host observations, at different times and places, hint at information about the associations between the hosts in a network. These smoothed, pseudo-continuous time-series suggest relationships with entities in the wider environment. For anomaly detection, mining this information can potentially provide a valuable source of observational experience for determining comparative anomalies or rejecting false anomalies.

Finally, this special issue concludes with a paper on finding repeated patterns in time series.[8] Such patterns were christened "time series motifs" by the author of this introduction, because of their close analogy to their discrete counterparts in computational biology. Tanaka and Uehara generalize previous work to allow the discovery of repeated patterns when the pattern length is not known in advance. They also generalize to the problem of discovery motifs in multi dimensional time series.

Acknowledgments

We thank the authors for their contributions, and the reviewers for their comprehensive reviews. We also thank Chotirat Ann Ratanamahatana for her useful comments and suggestions.

Notes

1. Evolutionary rule mining in time series databases. Magnus Lie Hetland and Pal Satrom.
2. Automatic feature extraction for classifying audio data, Ingo Mierswa and Katharina Morik.
3. Clustering time series with clipped data, Anthony Bagnall and Gareth Janacek.
4. Classification of multivariate time series and structured data using constructive induction, Mohammed Waleed Kadous and Claude.
5. Principle components and importance ranking of distributed anomalies, Kyrre Begnum and Mark Burgess.
6. Fast and exact warping of time series using adaptive segmental approximations, Yutau Shou, Nikos Mamoulis and David W. Cheung.
7. Multi-dimensional time-series motif discovery based on MDL principle, Yoshiki Tanaka and Kuniaki Uehara.
8. Elastic Translation Invariant Matching of Trajectories, Michail Vlachos, Dimitrios Gunopulos, George Kollios.

Eamonn Keogh
eamonn@cs.ucr.edu