Clustering of Time Series Data

Anne Denton
North Dakota State University, USA

INTRODUCTION

Time series data is of interest to most science and engineering disciplines and analysis techniques have been developed for hundreds of years. There have, however, in recent years been new developments in data mining techniques, such as frequent pattern mining, which take a different perspective of data. Traditional techniques were not meant for such pattern-oriented approaches. There is, as a result, a significant need for research that extends traditional time-series analysis, in particular clustering, to the requirements of the new data mining algorithms.

BACKGROUND

Time series clustering is an important component in the application of data mining techniques to time series data (Roddick & Spiliopoulou, 2002) and is founded on the following research areas:

- **Data Mining:** Besides the traditional topics of classification and clustering, data mining addresses new goals, such as frequent pattern mining, association rule mining, outlier analysis, and data exploration.
- **Time Series Data:** Traditional goals include forecasting, trend analysis, pattern recognition, filter design, compression, Fourier analysis, and chaotic time series analysis. More recently frequent pattern techniques, indexing, clustering, classification, and outlier analysis have gained in importance.
- **Clustering:** Data partitioning techniques such as k-means have the goal of identifying objects that are representative of the entire data set. Density-based clustering techniques rather focus on a description of clusters, and some algorithms identify the most common object. Hierarchical techniques define clusters at arbitrary levels of granularity.
- **Data Streams:** Many applications, such as communication networks, produce a stream of data. For real-valued attributes such a stream is amenable to time series data mining techniques.

Time series clustering draws from all of these areas. It builds on a wide range of clustering techniques that have been developed for other data, and adapts them while critically assessing their limitations in the time series setting.

MAIN THRUST

Many specialized tasks have been defined on time series data. This chapter addresses one of the most universal data mining tasks, clustering, and highlights the special aspects of applying clustering to time series data. Clustering techniques overlap with frequent pattern mining techniques, since both try to identify typical representatives.

Clustering Time Series

Clustering of any kind of data requires the definition of a similarity or distance measure. A time series of length *n* can be viewed as a vector in an *n*-dimensional vector space. One of the best-known distance measures, Euclidean distance, is frequently used in time series clustering. The Euclidean distance measure is a special case of an L_p norm. L_p norms may fail to capture similarity well when being applied to raw time series data because differences in the average value and average derivative affect the total distance. The problem is typically addressed by subtracting the mean and dividing the resulting vector by its L_2 norm, or by working with normalized derivatives of the data (Gavrilov et al., 2000). Several specialized distance measures have been used for time series clustering, such as dynamic time warping, DTW (Berndt & Clifford 1996), and longest common subsequence similarity, LCSS (Vlachos, Gunopulos, & Kollios, 2002).

Time series clustering can be performed on whole sequences or on subsequences. For clustering of whole sequences, high dimensionality is often a problem. Dimensionality reduction may be achieved through Discrete Fourier Transform (DFT), Discrete Wavelet Transform (DWT), and Principal Component Analysis (PCA), as some of the most commonly used techniques. DFT (Agrawal, Falutsos, & Swami, 1993) and DWT have the goal of eliminating high-frequency components that are typically due to noise. Specialized models have been introduced that ignore some information in a targeted way (Jin, Lu, & Shi 2002). Others are based on models for specific data such as socioeconomic data (Kalpakis, Gada, & Puttagunta, 2001).

A large number of clustering techniques have been developed, and for a variety of purposes (Halkidi, Batistakis, & Vazirgiannis, 2001). Partition-based techniques are among the most commonly used ones for time series data. The k-means algorithm, which is based on a greedy search, has recently been generalized to a wide range of distance measures (Banerjee et al., 2004).

Clustering Subsequences of a Time Series

A variety of data mining tasks require clustering of subsequences of one or more time series as preprocessing step, such as Association Rule Mining (Das et al., 1998), outlier analysis and classification. Partition-based clustering techniques have been used for this purpose in analogy to vector quantization (Gersho & Gray, 1992) that has been developed for signal compression. It has, however, been shown that when a large number of subsequences are clustered, the resulting cluster centers are very similar for different time series (Keogh, Lin, & Truppel, 2003). The problem can be resolved by using a clustering algorithm that is robust to noise (Denton, 2004), which adapts kernel-density-based clustering (Hinneburg & Keim, 2003) to time series data. Partition-based techniques aim at finding representatives for all objects. Cluster centers in kernel-density-based clustering, in contrast, are sequences that are in the vicinity of the largest number of similar objects. The goal of kernel-density-based techniques is thereby similar to frequent-pattern mining techniques such as Motif-finding algorithms (Patel et al., 2002).

Related Problems

One time series clustering problem of particular practical relevance is clustering of gene expression data (Eisen, Spellman, Brown, & Botstein, 1998). Gene expression is typically measured at several points in time (time course experiment) that may not be equally spaced. Hierarchical clustering techniques are commonly used. Density-based clustering has recently been applied to this problem (Jiang, Pei, & Zhang, 2003)

Time series with categorical data constitute a further related topic. Examples are log files and sequences of operating system commands. Some clustering algorithms in this setting borrow from frequent sequence mining algorithms (Vaarandi, 2003).

FUTURE TRENDS

Storage grows exponentially at rates faster than Moore's law for microprocessors. A natural consequence is that old data will be kept when new data arrives leading to a massive increase in the availability of the time-dependent data. Data mining techniques will increasingly have to consider the time dimension as an integral part of other techniques. Much of the current effort in the data mining community is directed at data that has a more complex structure than the simple tabular format initially covered in machine learning (Džeroski & Lavrač, 2001). Examples, besides time series data, include data in relational form, such as graph- and tree-structured data, and sequences. When addressing new settings it will be of major importance to not only generalize existing techniques and make them more broadly applicable but to also critically assess problems that may appear in the generalization process.

CONCLUSION

Despite the maturity of both clustering and time series analysis, time series clustering is an active and fascinating research topic. New data mining applications are constantly being developed and require new types of clustering results. Clustering techniques from different areas of data mining have to be adapted to the time series context. Noise is a particularly serious problem for time series data, thereby adding challenges to clustering process. Considering the general importance of time series data, it can be expected that time series clustering will remain an active topic for years to come.

REFERENCES

Banerjee, A., Merugu, S., Dhillon, I., & Ghosh, J. (2004, April). Clustering with Bregman divergences. In *Proceedings SIAM International Conference on Data Mining*, Lake Buena Vista, FL.

Berndt D.J., & Clifford, J. (1996). Finding patterns in time series: A dynamic programming approach. In *Advances in knowledge discovery and data mining* (pp. 229-248). Menlo Park, CA AAAI Press.

Das, G., & Gunopulos, D. (2003). Time series similarity and indexing. In N. Ye (Ed.), *The handbook of data mining* (pp. 279-304). Mahwah, NJ: Lawrence Erlbaum Associates.