Data & Knowledge Engineering 68 (2009) 318–337

Contents lists available at ScienceDirect

Data & Knowledge Engineering

journal homepage: www.elsevier.com/locate/datak

Establishing relationships among patterns in stock market data

Dietmar H. Dorr, Anne M. Denton *

Department of Computer Science and Operations Research, North Dakota State University, Fargo, ND, USA

ARTICLE INFO

Article history:
Received 6 September 2007
Received in revised form 1 October 2008
Accepted 6 October 2008
Available online 17 October 2008

Keywords:
Knowledge discovery
Pattern mining
Financial applications
Stock market
Time series data

ABSTRACT

Similarities among subsequences are typically regarded as categorical features of sequential data. We introduce an algorithm for capturing the relationships among similar, contiguous subsequences. Two time series are considered to be similar during a time interval if every contiguous subsequence of a predefined length satisfies the given similarity criterion. Our algorithm identifies patterns based on the similarity among sequences, captures the sequence–subsequence relationships among patterns in the form of a directed acyclic graph (DAG), and determines pattern conglomerates that allow the application of additional meta-analyses and mining algorithms. For example, our pattern conglomerates can be used to analyze time information that is lost in categorical representations. We apply our algorithm to stock market data as well as several other time series data sets and show the richness of our pattern conglomerates through qualitative and quantitative evaluations. An exemplary meta-analysis determines timing patterns representing relations between time series intervals and demonstrates the merit of pattern relationships as an extension of time series pattern mining.

1. Introduction

Time series data are ubiquitous in fields as diverse as economics, science, and industry; hence, it is not surprising that there has been a strong interest in applying data mining techniques to time series data. Time series can be very long, and users are often interested in similarities that extend over a comparatively short time interval, which suggests the use of sliding-window techniques. An approach that is based on sliding windows starts with all possible fixed length, contiguous subsequences of the time series under consideration. Note that the term "subsequence" has multiple meanings in the literature. We use subsequence in the sense of a contiguous section of a sequence that is also sometimes called "substring". In order to address the properties of time series data, special similarity measures have been devised that are defined over variable-length subsequences, as well as making other generalizations [45,7,6,12]. With well-established similarity measures in place, researchers have pursued pattern mining, clustering and classification tasks, as they are common in data mining.

The richness of temporal data is, however, not alone captured in modified similarity measures. In sequential data, strong reasons may be given as to why it can be beneficial to revise even the concept of pattern mining itself: conventionally pattern mining is seen as returning isolated, frequent occurrences in the data. Although relationships among patterns have been extensively used as a basis for pruning through closure properties [1], these set–subset relationships do not normally contribute much to the expressiveness of the result when time series are considered. In comparison to record data, time series data inherently provides an additional dimension (time) for each data item. The time dimension can be utilized not only for mining patterns but also for capturing the relationships among patterns. In our interpretation, a revised concept of pattern mining should include the interrelations among patterns.

* Corresponding author. Tel.: +1 701 231 6748; fax: +1 701 231 8255.
E-mail address: anne.denton@ndsu.edu (A.M. Denton).
URL: http://www.cs.ndsu.nodak.edu/adenton/ (A.M. Denton).

0169-023X/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.datak.2008.10.001

For example, knowing that a group of stock series shares a pattern over a long period of time, while other stock series show a related pattern over a much shorter interval can provide valuable insights into the price developments of stocks. The relationships among patterns have important information content by themselves. It is our goal to capture the similarities among stock market time series such that their sequence–subsequence relationships are preserved. We identify patterns representing collections of contiguous subsequences that share the same shape for a particular time interval. Patterns are defined on the basis of contiguous sections of normalized sliding windows that show pairwise similarities among sequences. The relationships among sliding-window patterns are represented using a directed acyclic graph (DAG) that is constructed based on the overlap between patterns. Leaf nodes within the DAG denote entire sequences, internal nodes represent patterns, and the sequence–subsequence relationships among patterns are represented by the edges. In a directed graph, an internal node, in contrast to a leaf node, has at least one directed edge to another node. The information contained within the DAG, as well as timing information, is represented using a pattern conglomerate notation that constitutes a new level of abstraction. The pattern conglomerate concept is designed to allow meta-analyses. In the context of this paper, a meta-analysis is an analysis applied to the results of another analysis, i.e., our pattern conglomerates (result of the first analysis) can be used as input to another, second analysis (meta-analysis). A pattern conglomerate incorporates the structure of the DAG and the order of clustered sequences, as well as the extent of the subsequences considered during the execution of our algorithm (Section 3.3). The panel (a) of Fig. 1 depicts an example of four time series that shows a total of three characteristic shapes. The sliding-window pattern that is signified by × is shared by all four sequences. Sequences A and B show a longer pattern that extends as far as the section with a □. Time series C and D have a different extended pattern comprised of × and ○. The corresponding DAG representation is shown in panel (b) of Fig. 1. Each time series is represented by a leaf node, and all three patterns are represented as internal nodes. The root node, ×, connects to the two other internal nodes, which represent the longer patterns. Note that the DAG is different from similarity-based representations that are common in hierarchical clustering, where degrees of similarities are used to group sequences. In our case, length of overlap determines the position in the DAG and similarity is defined through a single window-based threshold. Accordingly, the × node is created based on the overlap between patterns A/B (□×) and C/D (×○) rather than the degree of the similarity between the sequences. The third panel (c) of Fig. 1 depicts the abstraction of the DAG in form of a pattern conglomerate. The structure of the DAG is represented using parentheses, and the beginning and ending of regions of similarity between pairs of sequences are indicated by braces with subscripts.

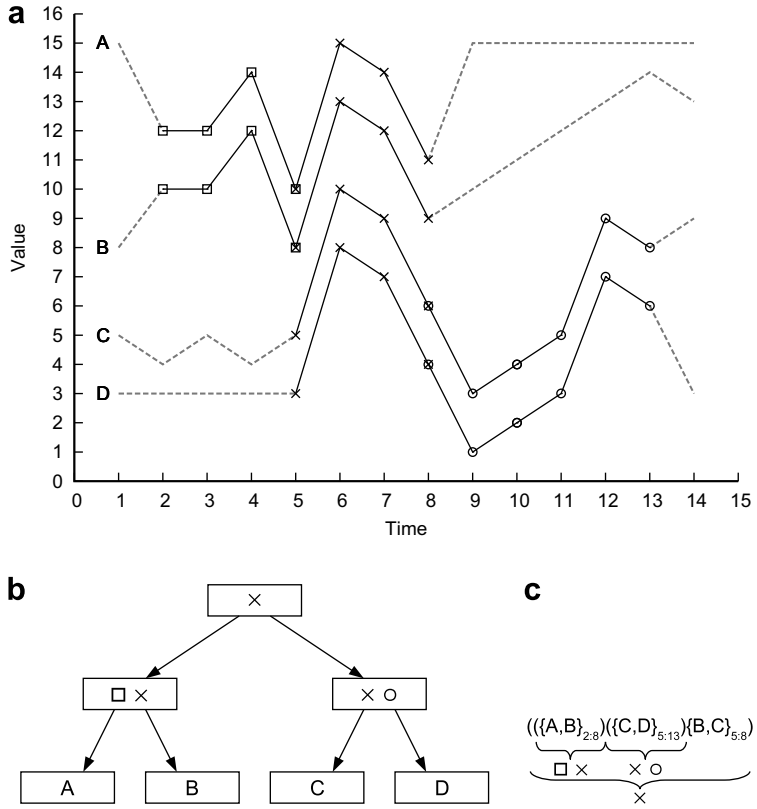

Fig. 1. An example of four time series that are similar to each other over different time intervals (a). The clustering result of the time series is shown in panel (b) using a DAG and (c) by the corresponding pattern conglomerate. In all three panels, the similarities among the time series are denoted by the symbols □, ×, and ○.