Available online at www.sciencedirect.com
SCIENCE DIRECT®
Pattern Recognition 38 (2005) 1857–1874
PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/patcog

Clustering of time series data—a survey

T. Warren Liao*

Industrial & Manufacturing Systems Engineering Department, Louisiana State University, 3128 CEBA, Baton Rouge, LA 70803, USA

Received 16 September 2003; received in revised form 21 June 2004; accepted 7 January 2005

Abstract

Time series clustering has been shown effective in providing useful information in various domains. There seems to be an increased interest in time series clustering as part of the effort in temporal data mining research. To provide an overview, this paper surveys and summarizes previous works that investigated the clustering of time series data in various application domains. The basics of time series clustering are presented, including general-purpose clustering algorithms commonly used in time series clustering studies, the criteria for evaluating the performance of the clustering results, and the measures to determine the similarity/dissimilarity between two time series being compared, either in the forms of raw data, extracted features, or some model parameters. The past researchs are organized into three groups depending upon whether they work directly with the raw data either in the time or frequency domain, indirectly with features extracted from the raw data, or indirectly with models built from the raw data. The uniqueness and limitation of previous research are discussed and several possible topics for future research are identified. Moreover, the areas that time series clustering have been applied to are also summarized, including the sources of data used. It is hoped that this review will serve as the steppingstone for those interested in advancing this area of research.

Keywords: Time series data; Clustering; Distance measure; Data mining

1. Introduction

The goal of clustering is to identify structure in an unlabeled data set by objectively organizing data into homogeneous groups where the within-group-object similarity is minimized and the between-group-object dissimilarity is maximized. Clustering is necessary when no labeled data are available regardless of whether the data are binary, categorical, numerical, interval, ordinal, relational, textual, spatial, temporal, spatio-temporal, image, multimedia, or mixtures of the above data types. Data are called static if all their feature values do not change with time, or change negligibly. The bulk of clustering analyses has been performed on static data. Most, if not all, clustering programs developed as an independent program or as part of a large suite of data analysis or data mining software to date work only with static data. Han and Kamber [1] classified clustering methods developed for handing various static data into five major categories: partitioning methods, hierarchical methods, density-based methods, grid-based methods, and model-based methods. A brief description of each category of methods follows.

Given a set of n unlabeled data tuples, a partitioning method constructs k partitions of the data, where each partition represents a cluster containing at least one object and $k \leqslant n$. The partition is crisp if each object belongs to exactly one cluster, or fuzzy if one object is allowed to be in more than one cluster to a different degree. Two renowned heuristic methods for crisp partitions are the *k-means* algorithm [2], where each cluster is represented by the mean value of the objects in the cluster and the *k-medoids*

* Tel.: +1 225 578 5365; fax: +1 225 578 5109.
E-mail address: ieliao@lsu.edu.

0031-3203/$30.00
doi:10.1016/j.patcog.2005.01.025

algorithm [3], where each cluster is represented by the most centrally located object in a cluster. Two counterparts for fuzzy partitions are the *fuzzy c-means* algorithm [4] and the *fuzzy c-medoids* algorithm [5]. These heuristic algorithms work well for finding spherical-shaped clusters and small to medium data sets. To find clusters with non-spherical or other complex shapes, specially designed algorithms such as Gustafson–Kessel and adaptive fuzzy clustering algorithms [6] or density-based methods to be introduced in the sequel are needed. Most genetic clustering methods implement the spirit of partitioning methods, especially the *k-means* algorithm [7,8], the *k-medoids* algorithm [9], and the *fuzzy c-means* algorithm [10].

A hierarchical clustering method works by grouping data objects into a tree of clusters. There are generally two types of hierarchical clustering methods: agglomerative and divisive. Agglomerative methods start by placing each object in its own cluster and then merge clusters into larger and larger clusters, until all objects are in a single cluster or until certain termination conditions such as the desired number of clusters are satisfied. Divisive methods do just the opposite. A pure hierarchical clustering method suffers from its inability to perform adjustment once a merge or split decision has been executed. For improving the clustering quality of hierarchical methods, there is a trend to integrate hierarchical clustering with other clustering techniques. Both Chameleon [11] and CURE [12] perform careful analysis of object "linkages" at each hierarchical partitioning whereas BIRCH [13] uses iterative relocation to refine the results obtained by hierarchical agglomeration.

The general idea of density-based methods such as DBSCAN [14] is to continue growing a cluster as long as the density (number of objects or data points) in the "neighborhood" exceeds some threshold. Rather than producing a clustering explicitly, OPTICS [15] computes an augmented cluster ordering for automatic and interactive cluster analysis. The ordering contains information that is equivalent to density-based clustering obtained from a wide range of parameter settings, thus overcoming the difficulty of selecting parameter values.

Grid-based methods quantize the object space into a finite number of cells that form a grid structure on which all of the operations for clustering are performed. A typical example of the grid-based approach is STING [16], which uses several levels of rectangular cells corresponding to different levels of resolution. Statistical information regarding the attributes in each cell are pre-computed and stored. A query process usually starts at a relatively high level of the hierarchical structure. For each cell in the current layer, the confidence interval is computed reflecting the cell's relevance to the given query. Irrelevant cells are removed from further consideration. The query process continues to the next lower level for the relevant cells until the bottom layer is reached.

Model-based methods assume a model for each of the clusters and attempt to best fit the data to the assumed model. There are two major approaches of model-based methods: statistical approach and neural network approach. An example of statistical approach is AutoClass [17], which uses Bayesian statistical analysis to estimate the number of clusters. Two prominent methods of the neural network approach to clustering are competitive learning, including ART [18] and self-organizing feature maps [19].

Unlike static data, the time series of a feature comprise values changed with time. Time series data are of interest because of its pervasiveness in various areas ranging from science, engineering, business, finance, economic, health care, to government. Given a set of unlabeled time series, it is often desirable to determine groups of similar time series. These unlabeled time series could be monitoring data collected during different periods from a particular process or from more than one process. The process could be natural, biological, business, or engineered. Works devoting to the cluster analysis of time series are relatively scant compared with those focusing on static data. However, there seems to be a trend of increased activity.

This paper intends to introduce the basics of time series clustering and to provide an overview of time series clustering works been done so far. In the next section, the basics of time series clustering are presented. Details of three major components required to perform time series clustering are given in three subsections: clustering algorithms in Section 2.1, data similarity/distance measurement in Section 2.2, and performance evaluation criterion in Section 2.3. Section 3 categories and surveys time series clustering works that have been published in the open literature. Several possible topics for future research are discussed in Section 4 and finally the paper is concluded. In Appendix A, the application areas reported are summarized with pointers to openly available time series data.

2. Basics of time series clustering

Just like static data clustering, time series clustering requires a clustering algorithm or procedure to form clusters given a set of unlabeled data objects and the choice of clustering algorithm depends both on the type of data available and on the particular purpose and application. As far as time series data are concerned, distinctions can be made as to whether the data are discrete-valued or real-valued, uniformly or non-uniformly sampled, univariate or multivariate, and whether data series are of equal or unequal length. Non-uniformly sampled data must be converted into uniformed data before clustering operations can be performed. This can be achieved by a wide range of methods, from simple down sampling based on the roughest sampling interval to a sophisticated modeling and estimation approach.

Various algorithms have been developed to cluster different types of time series data. Putting their differences aside, it is far to say that in spirit they all try to modify the existing algorithms for clustering static data in such a way that