Chapter X

Exploratory Time Series Data Mining by Genetic Clustering

T. Warren Liao
Louisiana State University, USA

ABSTRACT

In this chapter, we present genetic-algorithm (GA)-based methods developed for clustering univariate time series with equal or unequal length as an exploratory step of data mining. These methods basically implement the k-medoids algorithm. Each chromosome encodes, in binary, the data objects serving as the k-medoids. To compare their performance, both fixed-parameter and adaptive GAs were used. We first employed the synthetic control-chart data set to investigate the performance of three fitness functions, two distance measures, and other GA parameters such as population size, crossover rate, and mutation rate. Two more sets of time series with or without a known number of clusters were also experimented: one is the cylinder-bell-funnel data and the other is the novel battle simulation data. The clustering results are presented and discussed.

INTRODUCTION

Before prediction models can be built in data mining or knowledge discovery, it is often advisable to first explore the data. Clustering is known to be a good exploratory data-mining tool. The goal of clustering is to create structure for unlabeled data by objectively forming data into homogeneous groups, where the within-group object similarity and the between-group object dissimilarity are optimized. The bulk of clustering analyses has been performed on data associated with static features, that is, feature values that do not change with time, or the changes are negligible.

Two major classes of clustering methods are partitioning and hierarchical clustering. Well-known partitioning-based clustering methods include k-means (MacQueen, 1967), k-medoids (Kaufman & Rousseeuw, 1990), and the corresponding fuzzy versions: fuzzy c-means...
Exploratory Time Series Data Mining by Genetic Clustering

Hierarchical clustering methods are either of the agglomerative type or the divisive type. Lately, soft computing technologies, including neural networks and genetic algorithms, have emerged as another class of clustering techniques. Two prominent methods of the neural network approach to clustering are competitive learning and self-organizing feature maps. Most genetic clustering methods implement the spirit of partitioning methods, especially the \textit{k}-means algorithm (Krishna & Murty, 1999; Maulik & Bandyopadhyay, 2000), and the fuzzy \textit{c}-means algorithm (Hall, Özyurt, & Bezdek, 1999).

Just like static feature data, forming groups of similar time series given a set of unlabeled time series is often desirable. These unlabeled time series could be monitoring data collected during different periods from a particular process, or from several different processes. These processes could be natural, engineered, business, economical, or medical related. Unlike static feature values, the time series of a feature consists of dynamic values, that is, values changed with time. This greatly increases the dimensionality of the problem and calls for somewhat different, and often more complicated, clustering methods. This study will focus only on time-series data.

In surveying work related to time-series clustering, Liao (2005) distinguished three different time-series clustering approaches: those working with full data either in the time or frequency domain, those working with extracted features, and model-based approaches with models built from the raw data. An example of the first approach is Golay \textit{et al.} (Golay, Kollias, Stoll, Meier, Valavanis, & Boesiger, 1998). They applied the fuzzy \textit{c}-means algorithm to provide the functional maps of human brain activity on the application of a stimulus. In their study, three different distances (the Euclidean distance and two cross-correlation-based distances) were alternately used for comparison purposes. Goutte \textit{et al.} (Goutte, Toft, & Rostrup, 1999) and Fu \textit{et al.} (Fu, Chung, Ng, & Luk, 2001) took the feature-based approach. Goutte \textit{et al.} clustered functional magnetic resonance imaging (fMRI) time series in groups of voxels with similar activations using two algorithms: \textit{k}-means and Ward’s hierarchical clustering. The cross-correlation function, instead of the raw fMRI time series, was used as the feature space. Fu \textit{et al.} described the use of self-organizing maps for grouping similar temporal patterns dispersed along the time series. Two enhancements were made: consolidating the discovered clusters by a redundancy removal step, and introducing the perceptually important point-identification method to reduce the dimension of the input data sequences.

Three model-based time-series clustering methods are described next. Li and Biswas (1999) described a clustering methodology for temporal data using hidden Markov model representation with a sequence-to-model likelihood distance measure. The temporal data was assumed to have the Markov property. Time series were considered similar when the models characterizing individual series were similar. Policker and Geva (2000) presented a model for nonstationary time series with time varying mixture of stationary sources, comparable to the continuous hidden Markov model. Fuzzy clustering methods were applied to estimate the continuous drift in the time-series distribution, and the resultant membership matrix was given an interpretation as weights in a time varying, mixture probability distribution function. Kalpakis \textit{et al.} (Kalpakis, Gada, & Puttagunta, 2001) studied the clustering of ARIMA time-series by using the Euclidean distance between the Linear Predictive Coding Cepstra of two time-series as their dissimilarity measure and the Partition around Medoids (PAM) method as the clustering algorithm.

To the best of our knowledge, the only study that applied genetic algorithms to cluster time