Subsequence Time Series Clustering

Jason Chen
Australian National University, Australia

INTRODUCTION

Clustering analysis is a tool used widely in the Data Mining community and beyond (Everitt et al. 2001). In essence, the method allows us to "summarise" the information in a large data set X by creating a very much smaller set C of representative points (called centroids) and a membership map relating each point in X to its representative in C. An obvious but special type of data set that one might want to cluster is a time series data set. Such data has a temporal ordering on its elements, in contrast to non-time series data sets. In this article we explore the area of time series clustering, focusing mainly on a surprising recent result showing that the traditional method for time series clustering is meaningless. We then survey the literature of recent papers and go on to argue how time series clustering can be made meaningful.

BACKGROUND

A time series is a set of data points which have temporal order. That is,

$$X = \{x_t \mid t = 1, ..., n\} \quad (1)$$

where t reflects the temporal order. Two types of clustering of time series has historically been undertaken: whole series clustering and subsequence clustering. In whole series clustering, one generally has a number of time series of equal length (say n) and one forms a vector space of dimension n so that each time series is represented by a single point in the space. Clustering then takes place in the usual way and groupings of similar time series are returned.

Whole series clustering is useful in some circumstances, however, often one has a single long time series data set X and the aim is to find a summary set of features in that time series, e.g. in order to find repeating features or particular repeating sequences of features (e.g. see the rule finding method proposed in (Das et al.1998)). In this case, what was historically done was to create a set Z of subsequences by moving a sliding window over the data in X, i.e.

$$z_{p-(w-1)} = x_{p-(w-1)}, x_{p-(w-2)}, ..., x_{p-2}, x_{p-1}, x_p \quad (2)$$

$z_p \in Z, p = w...n$. Each subsequence z_p (also called more generally a regressor or delay vector; see below) essentially represents a feature in the time series. These features live in a w-dimensional vector space, and clustering to produce a summarising set C of "centroid" features can proceed in the usual way. This technique has historically been called Subsequence Time Series (STS) Clustering, and quite a lot of work using the technique was published (see (Keogh et al. 2003) for a review of some of this literature). In this article we will focus on the area of subsequence time series clustering. For a review of whole time series clustering methods, see (Wang et al. 2004).

Given the widespread use of STS clustering, a surprising result in (Keogh et al. 2003) was that it is meaningless. Work in (Keogh et al. 2003) defined a technique as meaningless if the result it produced was essentially independent of the input. The conclusion that STS clustering was meaningless followed after it was shown that, if one conducted STS clustering on a range of even very distinct time series data sets, then the cluster centroids resulting from each could not be told apart. More specifically, the work clustered each time series multiple times and measured the average "distance" (see (Keogh et al. 2003) for details) between clustering outcomes from the same time series and between different time series. They found on average that the distance between clustering outcomes from the same and different time series were the same. Further, they discovered the strange phenomenon that the centroids produced by STS clustering are smoothed sine-type waves.

After the appearance of this surprising result, there was great interest in finding the cause of the dilemma and a number of papers on the topic subsequently appeared. For example, Struzik (Struzik 2003) proposed that the

"meaningless" outcome results only in pathological cases, i.e. when the time series structure is fractal, or when the redundancy of subsequence sampling causes trivial matches to hide the underlying rules in the series. They suggested autocorrelation operations to suppress the latter, however these suggestions were not confirmed with experiments.

In contrast, Denton (Denton, 2005) proposed density based clustering, as opposed to, for example, k-means or hierarchical clustering, as a solution. They proposed that time series can contain significant noise, and that density based clustering identifies and removes this noise by only considering clusters rising above a preset threshold in the density landscape. However, it is not clear whether noise (or only noise) in the time series is the cause of the troubling results in (Keogh et al. 2003). For example, if one takes the benchmark Cylinder-Bell-Funnel time series data set (see (Keogh et al. 2003)) without noise and applies STS clustering, the strange smoothed centroid results first identified there are still returned.

Another interesting approach to explain the dilemma was proposed by Goldin et. al. (Goldin et al. 2006). They confirmed that the ways (multiple approaches were tried) in which distance between clustering outcomes were measured in (Keogh et al. 2003) did lead to the conclusion that STS-clustering was meaningless. However, they proposed an alternative distance measure which captured the "shape" formed by the centroids in the clustering outcome. They showed that if one calculates the average shape of a cluster outcome over multiple clustering runs on a time series, then the shape obtained can be quite specific to that time series. Indeed if one records all the individual shapes from these runs (rather than recording the average), then in an experiment on a set of ten time series they conducted, one is able to match a new clustering of a time series back to one of the recorded clustering outcomes from the same time series. While these results suggest meaningfulness is possible in STS-clustering, it seems strange that such lengths are required to distinguish between clustering outcomes of what can be very distinct time series. Indeed we will see later that an alternative approach, motivated from the Dynamical Systems literature, allows one to easily distinguish between the clustering outcomes of different time series using the simple distance measure adopted in (Keogh et al. 2003).

Another approach proposed by Chen (Chen 2005, 2007a) to solve the dilemma forms the basis of work which we later argue provides its solution. They proposed that the metrics adopted in (Keogh et al. 2003) in the clustering phase of STS clustering were not appropriate and proposed an alternative clustering metric based on temporal and formal distances (see (Chen 2007a) for details). They found that meaningful time series clustering could be achieved using this metric, however the work was limited in the type of time series to which it could be applied. This work can be viewed as restricting the clustering process to the subset of the clustering space that was visited by the time series; a key tenet of later work that we argue below forms a solution to the STS-clustering dilemma.

Peker (Peker 2005) also conducted experiments in STS-clustering of time series. They identified that clustering with a very large number of clusters leads to cluster centroids that are more representative of the signal in the original time series. They proposed the idea of taking cluster cores (a small number of points in the cluster closest to the centroid) as the final clusters from STS clustering. The findings in this work concur with work in (Chen 2007a) and the work we explore below, since they are compatible with the idea of restricting clustering to the subset of the clustering space visited by the time series.

While each of the works just reviewed show interesting results which shed light on the problems involved with STS-clustering, none provides a clear demonstration for general time series of how to overcome them.

MAIN FOCUS

We now propose our perspective on what the problem with STS clustering is, and on a solution to this problem; based on a number of recent papers in the literature. Let us revisit the problems found in (Keogh et al. 2003) with the STS clustering method. This work proposed that STS-clustering was meaningless because:

A. One could not distinguish between the clustering outcomes of distinct time series, even when the time series themselves were very different, and
B. Cluster representatives were smoothed and generally did not look at all like any part of the original time series