Subsequence Time Series Clustering

Jason R. Chen*
Department of Information Engineering
Research School of Information Science and Engineering
College of Engineering and Computer Science
Australian National University
Canberra, ACT
Australia
voice: +61 2 6125 8687
fax: +61 1 6125 8660
email: jason.chen@anu.edu.au

(* Corresponding author)

Subsequence Time Series Clustering

Jason Chen, Australian National University

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,

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

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.

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

$z_p \in Z,\ p = w \ldots n$. 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

They proposed that these two problems were one and the same, i.e. that one could not distinguish between cluster centres of different time series because they were all smoothed, and hence alike. This presumption turns out to be false, i.e. really these are two separate problems which need to be addressed and solved separately. For example, (Chen 2007b) showed how a time series clustering technique could produced distinguishable cluster centres (i.e. overcome (A)) but still produce centres that were smoothed (i.e. not overcome (B)). Hence, (Chen 2007b) proposed a new set of terminologies to reflect this fact. They proposed that a time series clustering method which overcomes problem (A) should be called meaningful, and one that overcomes problem (A) and (B) should be called useful. We will expand later on the motivation behind why these terms were adopted in each case.

Recall how it was shown in (Keogh et al. 2003) that STS-clustering is meaningless; the work clustered each time series multiple times and then measured the distance between clustering outcomes from the same time series and between different time series. Work in both (Chen 2007b) and (Simon et al. 2006) proposed that what was required to make the STS-clustering method meaningful was to introduce a lag *q* into the window forming process. That is, form subsequences as,

$$z_{p-(w-1)q} = x_{p-(w-1)q}, x_{p-(w-2)q}, \ldots, x_{p-2q}, x_{p-q}, x_p \tag{3}$$

$z_p \in Z, p = (w-1)q + 1 \ldots n$ (i.e. so that now adjacent points in the subsequence are separated by q data points in the time series) where we call z_p a regressor or delay vector. The inspiration of both works was from the field of Dynamical Systems (Sauer et al. 1991, Ott et al. 1994) where it is well known that introducing a lag is required for the embedding of any real world (i.e. noisy and represented with limited precision) time series in a vector space using a sliding windows type process. Geometrically, not using a lag means subsequence vectors will be clumped along the diagonal of the space, and hence, even with a small amount of noise present in the time series, and reasonable precision, the "information" in the embedding that distinguishes one time series from another is lost. Work in (Simon et al. 2006) went on to conduct the same experiment as in (Keogh et al. 2003) (albeit with different time series), but using a lag, and found that cluster centres produced from distinct time series were then indeed distinguishable. Work in (Chen 2007b) confirmed the result in (Simon et al. 2006) using basically the same time series as used in (Keogh et al. 2003). For clarity, and to distinguish between what follows, we follow the terminology adopted in (Chen 2007b) and denote the STS-clustering technique where a lag is introduced into the sliding windows process as Unfolded Time Series (UTS) clustering.

According to the “meaningful” and “useful” terminology introduced above, the UTS clustering method produces meaningful clustering outcomes. That is, if we cluster two distinct time series using the method, then UTS clustering produces centroid sets in each case which are distinct from one another. In essence, the “information” existing in the original time series which made them distinct has been retained in the clustering outcome, and so the clustering outcome really can be described as meaningful. Hence, the problem of achieving meaningful time series clustering would seem solved, i.e. one must introduce a lag into the subsequence vector construction process.

This could mark the end of the dilemma. However, recall the second problem ((B) above) observed by Keogh with STS clustering; that centroids are smoothed and do not look like, or retain the properties of, the original time series. Work in (Chen 2007b) noted that the UTS-clustering method, although meaningful, was still prone to this second problem, i.e. according to our adopted terminology it is not a useful time series clustering method. The term “useful” was adopted in (Chen 2007b) based on the observation that one clusters a time series to produce a summary set of features in the time series. If these features do not look like any part of the time series, then the outcome, although meaningful, is not useful. Why should UTS clustering be meaningful, but not produce centroids representative of the time series?

To answer this question, (Chen 2007b) proposed that we need to look more fundamentally at what we are asking when we UTS (or STS) cluster a time series. If we UTS cluster with a sliding window length of *d*, then we form a *d* dimensional clustering space $\Re^d$. In its entirety, $\Re^d$ represents the full range of possible subsequence (i.e. feature) shapes and magnitudes that can exist. However, (Chen 2007b) noted that the underlying system producing the time series almost certainly will not live on all of $\Re^d$, or indeed even on a convex subset of $\Re^d$ (something assumed by typical clustering algorithms like k-means and Expectation Maximization used in STS clustering to date). What sense does it make to include in the clustering process parts of $\Re^d$ that cannot be realised in the underlying system? Work proposing methods for clustering on subspaces (Haralick & Harpaz 2005), and manifolds (Breitenbach & Grundic 2005) exists and is motivated by exactly this line of thinking. Some simple experiments were conducted in (Chen 2007b) to show that this unrestricted approach to clustering in UTS (STS) clustering is the root cause of the smoothed centroid problem.

So we should cluster only in the subset of $\Re^d$ where valid outcomes from the underlying system exist. Unfortunately, given only a finite time series produced by the system, one cannot know the extent of this subset. However, this need not matter if the aim of clustering a time series is to (a) summarise the time series that was seen, rather than (b) to summarise the possible time series outcomes of an underlying system. (Chen 2007b) proposed that (a) is generally what we want to do when clustering a time series, and corresponds to asking the question: given the features observed in a time series, which *k* (for *k* clusters) of these features best "summarises" the time series. Given this observation, (Chen 2007b) went on to propose a method that restricts the clustering process to the region in $\Re^d$ visited by the time series. They proposed that this approach corresponds to the correct way to apply the clustering technique if indeed we want to ask the question corresponding to (a). They called the approach the Temporal-Formal (TF) clustering algorithm.

Details of the results of applying the technique can be found in (Chen 2007b), however in summary, the technique was applied on key time series data sets adopted from (Keogh et al. 2003) and (Simon et al. 2006). The results for all time series in the data set were,

(i) centroids were produced which remained in among data points in the cluster they represented, i.e. centroids looked like features from the original time series

(ii) clustering outcomes were meaningful (as per the definition of meaningful above) in all cases.

The conclusion was therefore made that the TF clustering algorithm was a useful time series clustering method. Further, analysis of the clustering outcomes for a number of time series was conducted in (Chen 2007b), including the benchmark Cylinder-Bell-Funnel time series. In each case the TF-clustering algorithm lead to the intuitively correct or (in the case of the benchmark time series) required outcome.

FUTURE TRENDS

Subsequence clustering of time series has been the focus of much work in the literature, and often as a subroutine to higher level motivations such as rule discovery, anomaly detection, prediction, classification and indexing (see (Keogh et al. 2003) for details). With the discovery that STS-clustering is meaningless, much future work will involve revisiting and reviewing the results and conclusions made by this work. Of great importance then is the discovery of a meaningful subsequence time series clustering method. We have argued here that a means for the solution of the problem exists, and while the arguments made in this work seem both clear and cogent, this work is quite recent. The dust still has not settled in the time series clustering area of data mining research.

CONCLUSION

We have reviewed the area of time series clustering, focusing on recent developments in subsequence time series (STS) clustering. Prior to 2003, STS clustering was a widely accepted technique in data mining. With the discovery in (Keogh et al. 2003) that STS-clustering is meaningless, a number of articles were published to explain the dilemma. While interesting results were presented in all these papers, we argued that two papers provide a solution to the dilemma ((Simon et al. 2006) and (Chen 2007b)). Specifically, work in (Keogh et al. 2003) identified two problems with STS clustering: (A) and (B) as described above. Together the papers show that these problems can be solved (respectively) by (a) introducing a lag into the sliding windows part of the STS-clustering process, and (b) restricting the clustering process to only that part of the clustering space pertinent to the time series at hand. Hence, we propose that this work forms a solution to the STS-clustering dilemma first identified in (Keogh et al. 2003).

REFERENCES

Denton, A. (2005), Kernel-density-based clustering of time series subsequences using a continuous random-walk noise model, in *Proceedings of IEEE International Conference on Data Mining*, Houston, USA.

Everitt, B.S., Landau,S., & Leese, M. (2001), *Clustering Analysis*, Wiley.

Goldin,D., Mardales, R., & Nagy, G. (2006), In search of meaning for time series subsequence clustering: matching algorithms based on a new distance measure, in *Proceedings of Conference of Information and Knowledge Management*, Arlington, USA.

Keogh, E., Lin, J., & Truppel, W. (2003), Clustering of time series subsequences is meaningless: Implications for previous and future research, in *Proceedings of the International Conference of Data Mining*.

Ott, E., Sauer, T., & Yorke, J. (1994), *Coping with Chaos*, Wiley.

Das, G., Lin, K., Mannila, H., Renganathan, G., & Smyth, P. (1998), Rule discovery from time series, in *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining*, New York, NY.

Simon, G., Lee, J.A., & Verleysen, M. (2006), Unfolding preprocessing for meaningful time series clustering, *Neural Networks* 19, 877–888.

Chen, J.R. (2005), Making subsequence time series clustering meaningful, in *Proceedings of IEEE International Conference on Data Mining*, Houston, USA, pp. 114–121.

Chen, J.R. (2007a), Making clustering in delay vector space meaningful, *Knowledge and Information Systems* ,11(3), 369–385.

Chen, J.R. (2007b), Useful clustering outcomes from meaningful time series clustering, in *Proceedings of the Australasian Data Mining Conference*, Gold Coast, Australia.

Peker, K. (2005), Subsequence time series (sts) clustering techniques for meaningful pattern discovery, in *International Conference Integration of Knowledge Intensive Multi-Agent Systems (KIMAS)*,Waltham, USA.

Breitenbach, M., & Grundic, G.Z. (2005), Clustering through ranking on manifolds, in *Proceedings of the 22nd International Conference on Machine Learning*, Bonn, Germany.

Haralick, R.M., & Harpaz, R. (2005), Linear manifold clustering, in *Proceedings of the International Conference on Machine Learning and Data Mining*, Leipzig.

Sauer, T., Yorke, J., & Casdagli, M. (1991), Embedology, *Journal of Statistical Physics* 65, 579.

Wang, X., Smith, K.A., Hyndman, R., & Alahakoon, D. (2004), A scalable method for time series clustering, *Technical report of Department of Econometrics and Business Statistics*, Monash University, Victoria, Australia.

Struzik, Z.R. (2003), Time Series Rule Discovery: Tough, Not Meaningless, *Foundations of Intelligent Systems*, *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, pp. 32–39.

KEY TERMS AND THEIR DEFINITIONS

Time Series: a data set containing elements which have a temporal ordering

Whole Time Series Clustering: the process of applying standard clustering techniques to a dataset whose elements are distinct time series of equal length.

Subsequence Time Series (STS) Clustering: the process of applying standard clustering techniques to a dataset whose elements are constructed by passing a sliding window over (usually) a single (long) time series.

Subsequence Vector: elements of the data set obtained in STS clustering, i.e. by using Equation 2 above.

Unfolded Time Series (UTS) Clustering: UTS clustering is STS clustering where a lag greater than unity has been introduced into the sliding windows process.

Regressor: elements of the data set obtained in UTS clustering, i.e. by using Equation 3 above.

Delay Vector: elements of the data set obtained in UTS clustering, i.e. by using Equation 3 above.

Lag: the sliding window used in UTS clustering need not capture, as a delay vector, a sequence of adjacent points in the time series. The lag is the value $q = p+1$ where p is the number of data points in the time series lying between adjacent points in the delay vector. So, for example, a lag $q = 3$ means the first delay vector will be $x_1, x_4, x_7, \ldots$

Temporal-Formal (TF) Clustering: UTS clustering where the clustering process is restricted to the region in the clustering space that was visited by the time series.

Meaningless: in general, an algorithm is said to be meaningless if its output is independent of its input. In the context of time series clustering, a time series clustering algorithm is said to be meaningless if one cannot distinguish between the clustering outcomes of distinct time series.