Semi-Markov Clustering

Matthew Robards and Peter Sunehag(presenting)
NICTA
Locked Bag 8001
Canberra, ACT 2601
Australia
Matthew.Robards@nicta.com.au, Peter.Sunehag@nicta.com.au

February 17, 2009

1 Introduction

Clustering of subsequences of time series is a widely studied and applied class of techniques [3, 4, 2, 6, 1, 5, 7] which aims at interpreting a time series as a shorter sequence of symbols, each of which represent a segment of the series with a pattern that is similar to other segments that has been assigned the same symbol.

The research field of subsequence clustering, which was already a widely applied and studied technique, took a dramatic turn after Keogh, Lin and Truppel [4] published a paper which claimed that clustering subsequences using sliding window approaches is meaningless. They had discovered that techniques such as the standard k-means clustering of the segments of a time series created by a sliding window of a fixed size are often looking the same regardless of the input data. When using a fixed size sliding window one will often create segments that are in different phases with respect to the pattern in the data which makes it difficult to learn meaningful patterns.

Here we extend the objective function of k-means clustering from fixed window segments to segmentations of varying length. In this new clustering framework there is no need for using overlapping windows. Semi-markov clustering does not cluster a fixed set of segments, but works with all possible segmentations to find one way of segmenting the sequence into consistent patterns. Unlike in the fixed window setting, the result is a segmentation into consistent patterns and centroids that look exactly like those patterns. We report results on several dataset including the artificial Cylinder-Bell-Funnel set, ECG and accelerometer traces from body worn sensors.

Topic: estimation, prediction, and sequence modeling
Preference: Oral/Poster are both fine

2 Semi-Markov k-Means

We assume that we are given sequences of observations x, i.e. a multi-dimensional time series, and that there exist an unknown associated sequence of labels y which are jointly drawn from some distribution $\Pr(x, y)$. To keep the notation simple (we will need to index the positions within the sequence and we would like to avoid double indexing) assume that we only have a *single* sequence x. Extensions to multiple sequences are straightforward. We express y in a compressed representation as a *list* of pairs $y = (n_0, l_0), \ldots, (n_m, l_m)$ of segment boundaries n_i and associated labels l_i. Note that the *number* of segments m itself is variable.

We define a feature vector $\phi(x, n_{i-1}, n_i)$ for every segment. The feature vector's dimension does not depend on the length of the segment.

In our semi-Markov k-means clustering algorithm, we find a labeled segmentation y of a given data sequence x by performing the following minimization:

$$y := \arg\min_y F(x, y; \mu) \tag{1}$$

where we would ideally like to use

$$\tilde{F}(x, y; \mu) := \frac{1}{m} \sum_{i=1}^{m} \|\phi(x, n_i, n_{i-1}) - \mu(l_i)\|^2 \tag{2}$$

which is the average square distance of the feature vector for the segments from $y = (n_0, l_0), \ldots, (n_m, l_m)$ to their class centroids $\mu(l_i)$. This ideal formulation (2) is, however, ill suited for dynamic programming which is necessary to perform the minimization in (1). Instead we take advantage of the fact that if all segments have the same length η and the total length of the sequence is M then $\frac{1}{m} = \frac{\eta}{M}$ and we can move η into the summation, resulting in the objective that we do use, namely

$$F(x, y; \mu) := \sum_{i=1}^{m} \|\phi(x, n_i, n_{i-1}) - \mu(l_i)\|^2 \eta_i \tag{3}$$

where $\eta_i = n_i - n_{i-1}$, i.e. the segment length of the i:th segment, and we have removed the irrelevant factor $\frac{1}{M}$. If all η_i are equal this formulation is equivalent to the ideal (2) and therefore using (3) is also a generalization of the objective function used for sliding window k-means to all possible segmentations. Note that if we did not use either an average distance or the factors η_i then we would have a bias towards having as few and long segments as possible since that would yield fewer terms in the sum.

Given an initialization of the centroids $\mu(l_i)$ we perform an iterative optimization procedure that is an easy extension of the standard k-means procedure which interchangable performs a labeled segmentation based on given centroids and calculates new centroids based on an existing labeled segmentation through simply averaging the segments assigned to each class.

References

[1] J. Chen. Making subsequence time series clustering meaningful. pages 114–121. IEEE International Conference on Data Mining, Houston USA, 2005.

[2] G. Das, K. Lin, H. Manilla, H. Renganathan, and P. Smyth. Rule discovery from time series. 4:th International Conference on Knowledge Discovery and Data Mining, New York, NY USA, 1998.

[3] D. Golden, R. Mardales, and G. Nagy. In search of meaning of time series subsequence clustering: matching algorithms based on a new distance measure. Conference of Information and Knowledge Management, Arlington, USA, 2006.

[4] E. Keogh, J. Lin, and W. Truppel. Clustering of time series subsequence is meaningless: Implications for previous and future research. International Conference of Data Mining, 2003.

[5] K. Peker. Subsequence time series (sts) clustering techniques for meaningful pattern discovery. Australasian Data Mining Conference, Gold Coast, Australia, 2005.

[6] G. Simon, J. Lee, and M. Verleysen. Unfolding preprocessing for meaningful time series clustering. volume 19, pages 877–888. Neural Networks, 2006.

[7] Z. Strutzik. Time series rule disovery: Tough not meaningless. pages 32–39. Foundations of Intelligent Systems, Lecture Notes in Computer Science, Springer, 2003.