UNDERTOW: Multi-Level Segmentation of Real-Valued Time Series

Tom Armstrong and **Tim Oates**
Department of Computer Science and Electrical Engineering
University of Maryland Baltimore County
1000 Hilltop Circle, Baltimore, Maryland 21250
{arm1, oates}@umbc.edu

Abstract

The discovery of meaningful change points, finding segments, in both categorical and real-value data time series is a well-studied problem. Prior segmentation algorithms and tasks operate under overly restrictive assumptions (e.g., *a priori* knowledge of the number of segments, trivial inputs) and in singular domains (e.g., finding common regions in images, speaker change detection). We introduce a domain-independent algorithm, UNDERTOW, which discovers segment boundaries in real-valued time series and constructs hierarchies of segments to form macro segments.

Introduction

Segmenting data is a hard problem. In spite of the challenge, adult humans and animals are facile at processing inputs from sensory receptors to discover meaningful change points, yet they begin life with rudimentary abilities. Learning over time and diversity of experience are necessary to identify correct segment boundaries. For example, children require over one year of proximal stimuli to gain adult-level performance in parsing speech across phonological phrase boundaries (Christophe *et al.* 2003).

How to define a segment and what constitutes a *good* segmentation remain open problems. Perceptual organization theories vary in a spectrum from structuralism to the Gestaltists. In vision research, the structuralist state of the art uses graph cuts as segment boundaries and maximizes the homogeneity of each segment globally for a good segmentation (Shi & Malik 2000). This structuralist view fails on other domains when homogeneity is harder to define than simple uniformity of color. On the other end of the spectrum, Gestaltists globally minimize the description to define a good segmentation, but there are no working unsupervised, domain-independent Gestalt methods.

The remainder of the article is organized as follows. First, we review the related work from real-valued time series and categorical data (e.g., natural language text). Then, we introduce our algorithm contribution and preliminary results before concluding.

Related Work

In 1961, Richard Bellman introduced a dynamic programming solution to finding the optimal *k*-segmentation of a real-valued time series (Bellman 1961). The solution, and more efficient approximations, fit a linear model to each segment and the number of segments is known. In real-world domains like speech data, knowing the number of words or sentences is an unreasonable assumption. However, even with that knowledge, the quantity of segments makes the approaches inefficient. Methods for segmenting speech data are successful, but begin with too much language information and gloss over how they arrive at their representation of basic elements like phonemes (Hammerton 2002).

Categorical data offers more leverage in real-world domains with small alphabets (e.g., bioinformatics, written language). In (Wolff 1975), compression is advocated as a way to discover meaningful segments by replacing frequently occurring bigrams with a unique symbol. More recently, (Nevill-Manning & Witten 1997) makes more judicious merges to compress a single string. The resulting hierarchical representation finds macro segments, but is limited to the single input string.

UNDERTOW, in part, extends the Voting-Experts (Cohen, Heeringa, & Adams 2002) algorithm to real-valued data through the use of computing boundary entropy at nodes in a trie populated by sliding a fixed-length window over a time series. The philosophy of our approach differs from the related work in that a segmentation algorithm should be robust with respect to alphabet size changes; the output of a segmentation algorithm, resulting in macro segments of the input alphabet, should also suffice as an input to the same algorithm.

Approach and Results

UNDERTOW (see figure 1) takes as input a real-valued time series, sliding window size, and an entropy change bound. First, it converts a real-valued time series into a fixed alphabet-sized time series with the Symbolic Aggregate Approximation (SAX) representation (Lin *et al.* 2003)[1]. Second, in the training phase (this phase is nearly identical to that of Voting-Experts), a sliding window passes over the

[1] Consult the original paper for the virtues of the SAX representation of time series.

```
UNDERTOW($timeSeries, w, h$)
  $timeSeries \leftarrow$ SAX($timeSeries$, $|\Sigma|$)
  repeat
        for $i < |timeSeries| - w$
           do
              INSERTTRIE($timeSeries[i, i + w]$, $w$)
        NODEENTROPY()
        NODE$\Delta$ENTROPY()
        for $i < |timeSeries| - w$
           do
              $window \leftarrow timeSeries[i, i + w]$
              $max \leftarrow$ FINDMAX$\Delta$ENTROPY($window$)
              if $max > h$
                then
                     INSERTSEGMENTPOINT
   until $output \neq previousOutput$
```

Figure 1: UNDERTOW Pseudocode

data and stores the values in a trie. Each node in the trie computes its entropy and change in entropy from its parent. Third, novel data windows find the maximum change in entropy in the trie and if that value is over the threshold, a segmentation point is inserted and the prior data are reified as a segment. Finally, the output becomes the new input to the training phase where the alphabet consists of output segments.

UNDERTOW gains all of the benefits of categorical approaches by using a small SAX alphabet (a value from 2 to 20 which can be varied based on the domain of inquiry) and a representation that lower bounds distance measures. UNDERTOW lacks the crippling requirements of segment counts and linear segment models found in real-valued approaches. In figure 2, the top plot displays the SAX representation of the first 100 elements of a random walk data set (Keogh 2006) and the bottom is the true time series. The segmentation points and hierarchy discovered by UNDERTOW in the first 95 data points are overlaid on the top plot as discovered and the bottom plot for comparison[2].

Discussion and Future Work

UNDERTOW finds a multi-level segmentation of real-valued times series by expanding existing techniques for categorical data. Additional results on categorical data like natural language text indicate that phrases and sentences can be discovered. Evaluating the output of UNDERTOW is a challenging task given the nature of the input time series. Methods for categorical data that perform on natural-language text begin with the gold standard. Future work will proceed in three directions. First, we will use robot sensor data which provides another rich source of data that can be evaluated without an expert. Second, the real-valued analogue of text is speech and with sufficient corpora we plan to evaluate UNDERTOW on finding phonemes, word, sentence, and speaker change boundaries. Finally, we will explore using other metric-based approaches (e.g., Kullback-Leibler divergence, Gish likelihood ratio) to evaluate the quality of segmenting at points of high entropy change.

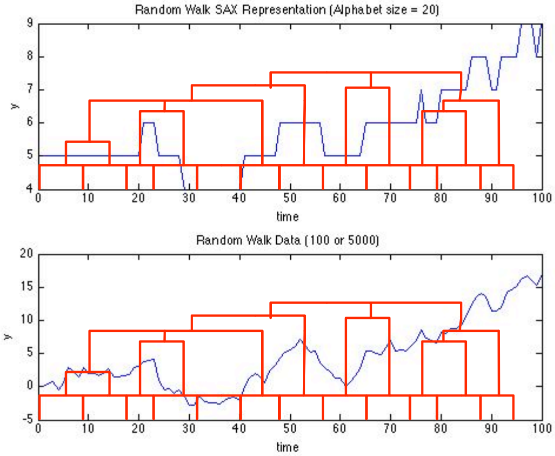

Figure 2: Random walk data (Keogh 2006) and segmentation points found by UNDERTOW

References

Bellman, R. 1961. On the approximation of curves by line segments using dynamic programming. *Commun. ACM* 4(6):284.

Christophe, A.; Gout, A.; Peperkamp, S.; and Morgan, J. 2003. Discovering words in the continuous speech stream: the role of prosody. *Journal of Phonetics* 31:585–598.

Cohen, P. R.; Heeringa, B.; and Adams, N. M. 2002. Unsupervised segmentation of categorical time series into episodes. In *ICDM*, 99–106.

Hammerton, J. 2002. Learning to segment speech with self-organising maps. *Language and Computers* Computational Linguistics in the Netherlands(14):51–64.

Keogh, E. 2006. The ucr time series data mining archive. http://www.cs.ucr.edu/ eamonn/TSDMA/index.html.

Lin, J.; Keogh, E.; Lonardi, S.; and Chiu, B. 2003. A symbolic representation of time series, with implications for streaming algorithms. In *DMKD '03: Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, 2–11. New York, NY, USA: ACM Press.

Nevill-Manning, C. G., and Witten, I. H. 1997. Identifying hierarchical structure in sequences: A linear-time algorithm. *Journal of Artificial Intelligence Research* 7:67.

Shi, J., and Malik, J. 2000. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*.

Wolff, J. G. 1975. An algorithm for the segmentation of an artificial language analogue. *British Journal of Psychology* 66:79–90.

[2] Additional information and results are available at `http://www.coral-lab.org/~arm1/undertow/`