Unsupervised Discovery of Motifs Under Uniform Amplitude Scaling and Shifting in Time Series Databases

Eric Drewniak
Advisor: Tom Armstrong

Mathematics and Computer Science Department
Wheaton College
26 East Main Street
Norton, MA 02766 USA

Abstract

We introduce an algorithm for unsupervised discovery of frequently occurring patterns in time series databases. Unlike prior approaches that can handle pattern distortion in the time dimension only, our algorithm is robust at finding pattern instances with amplitude shifting and with amplitude scaling. Using an established discretization method, SAX, we augment the existing real-valued time series representation with additional features to capture shifting and scaling. We point toward experimental results on synthetic and real-world data sets that demonstrate our improvement over the state-of-the-art algorithms.

Introduction

A time series motif is a subsequence that reoccurs in a data source with relatively high frequency. Finding previously unknown motifs in data is the problem of unsupervised motif discovery. This paper provides a novel representation of time series data and extends a motif-discovery algorithm to locate previously undiscoverable motifs. Our representation draws inspiration from speech processing and enables current algorithms to handle motifs that vary in amplitude either by a uniform shift or a scale in value. Previous approaches considered motif discovery under uniform scaling in the time dimension only (Yankov et al. 2007) or were not robust to these transformations (Catalanto, Armstrong, and Oates 2006).

The problems of mining time series for patterns, indexing signal data for interesting occurrences, and efficiently querying large datasets for particular subsequence have seen a significant amount of contributions recently. The formulation of the motif discovery problem comes from the genomics research community interested in finding exact or nearly exact genomic subsequences. The extension to real-valued data in multiple dimensions has expanded the potential applications of the work.

Traditionally, experts or trained analysts scoured over these data to find patterns and interesting subsequences for closer inspection. Relying on human expertise or intervention is, unfortunately, slow and increasingly error prone. For example, an obstetrician may be called away only to return to the fetal heart rate monitor after several minutes. Humans necessarily cannot provide real-time monitoring of massive amounts of real-valued data. Additionally, the proliferation and volume of information make it impossible for humans to solve these problems efficiently for all data. Therefore, we must turn to

computational approaches.

In the following sections, we review background on the motif-discovery problem and perspectives from other data mining approaches. Next, we discuss the related approaches and point to a gap in the literature. Then, we detail our algorithmic contribution and derived features. Finally, we conclude with some empirical results and point toward future work.

Background

(l, d)-Motif Discovery Problem

The motif discovery problem comes from the challenges of handling large volumes of biological data and mining those data for information. Originally termed the **Planted (l, d)-Motif Discovery Problem,** which (Buhler and Tompa 2002) defines as a generalization of the work of (Sagot 1998) and (Pevzner and Sze 2000):

"**Planted (l, d)-Motif Problem**: Suppose there is a fixed but unknown nucleotide sequence *M* (the *motif*) of length *l*. The problem is to determine *M*, given *t* nucleotide sequences each of length *n*, and each containing a planted variant of *M*. More precisely, each such planted variant is a substring that is *M* with exactly *d* point substitutions."

Buhler et al. provided an algorithm using random projection to find motifs in biological sequences (Buhler and Tompa 2002).

Time Series Representations

There are countless representation choices for time series discretization (see (Lin et al. 2003) for an extensive taxonomy). Recently, SAX, the symbolic aggregate approximation, has gained traction as the representation of choice for time series data mining (Lin et al. 2003). SAX has a number of advantageous properties: 1) the size of the discretization alphabet is variable; 2) places an implicit order on alphabet elements; and 3) SAX has a lower-bounding distance measure. See Figure 1 for a time series overlaid with SAX alphabet symbols and alphabet divisions where $\Sigma = 8$.

In our approach, we use the SAX string representations of time series as the basis for our augmented features.

Figure 1: (Top) The original time series. (Bottom) A time series and SAX alphabet symbols where the

size of the alphabet is 8.

Speech Processing

Speech processing is a domain in which finding recurring episodes in signals, segmenting those signals, and querying the data for particular subsequences are common challenges. In the standard speech processing pipeline (Jurafsky and Martin 2000), speech waveforms are preprocessed and sequences of spectral feature vectors are extracted to represent the original sound.

In addition to the spectral features (e.g., mel-frequency cepstral coefficients), the signals are augmented with derived features. The first and second derivates, the *deltas* and *delta-deltas*, are added as additional features to capture changes over time in the signal. Motif discovery algorithms have not exploited the potential of this additional information. In biological domains, where the alphabets of the datasets have no ordering, this information is unavailable. But, the SAX string representation of real-valued time series is able to consider delta and delta-deltas as additional features for time series data.

Related Work

Unsupervised discovery of novel, recurring subsequences in time series databases is an open problem. An exhaustive indexing of the entire time series solves the motif discovery problem, however space limitations or having a streaming time series preclude this as a possibility in many cases.

One approach considers a randomized algorithm approach to the problem leveraging hashing collisions to indicate likelihood of motifs (Chiu, Keogh, and Lonardi 2003) continuing work done using the piecewise aggregate approximation (Lin et al. 2002). An extension to this approach handles uniform scaling of the motifs in the time dimension (Yankov et al. 2007). Both approaches use the SAX approximation. However, the size of the discovered motifs are limited to the length of user-selected window length. Not all motifs occur in every dimension nor over the same time scale in each dimension -- the problem of subdimensional motif discovery (Minnen et al. 2007).

Other approaches that do not explicitly discretize the inputs and do not limit the size of the motifs. Most use Dynamic Time Warping (DTW) to handle candidate motifs that differ in time. The SAND algorithm takes a sampling approach to discover portions of motifs in real-value time series (Catalano, Armstrong, Oates 2006). The small motif chunks are stitched together to form larger motifs. The probabilistic sampling approach can be improved if constraints on the motifs in the time series are known (Yasser Mohammad, and Toyoaki Nishida 2009).

These algorithms have been applied to a variety of novel data. For example, motifs are discovered in exercise data (Minnen et al. 2007b), episodes are found in activity data (Vahdatpour, Amini, and Sarrafzadeh 2009), and experiences of mobile robots are clustered (Oates 1999; Oates, Schmill, and Cohen 2000).

Algorithm

We propose a randomized algorithm that extends the standard random projection algorithm (Chu, Keogh, and Lonardi 2003), which we call SRP. Unlike other extensions that find motifs that vary uniformly in the time dimension (Yankov et. al 2007), we consider finding motifs that vary by an amplitude shift, amplitude scaling, or some combination of both.

The algorithm operates analogously to SRP, but expands on the time series representation using derived features. The motivation for the derived features comes from speech processing. With speech waveforms, frequently occurring subsequences need not always have the same waveform and may vary significantly

in amplitude. Exclusively using the SAX alphabet string representations for time series has potential pitfalls in some domains. For example, depending on the size of the alphabet and the composition of the time series, motfis that differ by slight amplitude shifts may have completely different representations -- the motif representations will differ in most positions.

SRP begins with a SAX representation of the time series. The alphabet size and window size for the SAX representation are user-selected parameters. The user also selects the window size to consider when processing the SAX string. SRP extracts a list of substrings from the sequence by passing a fixed-length window over the SAX string. Our approach differs at this point, and we discuss the difference in the next section. Following the windowing step, a fixed series of rounds occurs. In each round, SRP masks a random selection of positions in the string and uses the resulting string as a key. Each string is hashed using that key. Strings that hash to the same location are considered similar because their SAX representations match at those fixed positions.

Over the subsequent rounds, strings that are similar in most positions tend to hash to the same location more often than those that do not. A collision matrix stores the hashing collisions between all string pairs. Candidate motif string pairs are read off of the collision matrix in sorted order with a minimum threshold statistic. The augmented version of the SRP algorithm is presented in pseudocode in Figure 2 where *timeseries* is the time series data, w is the size of the window, Σ is the size of the SAX alphabet, and n is the number of random projection rounds. The additional features overall do not change the computational complexity of the algorithm in practice. The only increase comes on the order of the window size, but the window size is typically a small constant.

Derived Features

Unlike SRP, we do not use the SAX string alone as input to the random projection portion of the algorithm. Instead, we developed several derived features to use in place of the SAX string. We selected these features to capture additional motifs for domains that do not always contain perfect motif matches.

One underutilized feature of the SAX representation is its implicit ordering of alphabet symbols. There is a strict ordering on alphabet elements, and we exploit this ordering to extend the representation with derived features. Motivated by the speech processing derived features, we augment the SAX strings with three additional features. First, we use the change in value between the SAX string letters as a proxy for

the first time derivative. We call this the *delta* feature, and for example, the SAX letters *abaa* has a [1, -1, 0] delta feature.

Second, we use the change in value between the delta features as the second time derivative. We call this the *delta-delta* feature, and for example, the delta feature [1, -1, 0] has a [-1, 1] delta-delta feature. Third, we approximate the concavity or convexity of the subsequence with the sign of the delta feature. We call this the *shape* feature, and for example, a [1, -1, 0] delta feature has a [+, -, +] shape feature. Consider the SAX string *bbdddaaacacccd* and the derived features for a window of length four in Table 1.

The delta, delta-delta, and shape features are intended to capture amplitude shifting and scaling that would normally result in a non-match with SRP. The delta and delta-delta features primarily serve as a mechanism to account for amplitude shifted motifs. The shape feature captures amplitude scaling where the delta and delta-delta features would not match.

Consider the following three SAX strings: (1) *aaabbbaaa*, (2) *cccdddccc*, and (3) *aaadddaaa*. Given the ordering of the discretization alphabet, these two strings are similar, but (1) and (2) are shifted by two SAX letters. SRP using SAX strings alone to discover motifs disallows the possibility for strings (1) and (2) to be instances of the same motif. The delta and delta-delta features match for (1) and (2). Alternatively, (1) and (3) match on some delta and delta-delta features, but do not in the middle of the strings. The shape features, however, do match allowing these two as possible motifs.

Experimental Results

We evaluated our algorithm and compared the results to other motif-discovery algorithms using three kinds of data: synthetic datasets with randomly inserted motifs, real-world datasets with known motif locations, and real-world datasets with unknown motif locations.

First, we discuss the evaluation of our algorithm and SRP on random walk data containing inserted peaks, valleys, constants, and sine waves. Second, we consider a baseline industrial dataset with known motifs. Third, we evaluate our algorithm on finding motifs in stock closing prices and on-body sensor data.

Synthetic Amplitude Scaling Data

We inserted fixed-length peaks into random positions in the dataset. Each peak differed in magnitude, but begins and ends at the same amplitude. In this particular experiment, our approach discovered the inserted motif in 12 out of the top 30 candidates; SRP discovered 0 out of the top 30 candidates. Figure 3 shows one of the discovered motifs.

Figure 3. (Top) Motif subsequences overlaid from SAX representation of original data set; and (Bottom) Complete random walk time series and highlighted motif subsections.

Synthetic Amplitude Shifting Data

We inserted fixed-length sine waves and constant values into random positions in the dataset. Each sine wave has the same magnitude, but varies in initial amplitude. In this particular experiment, our approach discovers the inserted motif in 12 out of the top 30 candidates; SRP discovered 0 out of the top 30 candidates. Figure 4 shows one of the discovered motifs.

Figure 4. (Top) Motif subsequences overlaid from SAX representation of original data set; and (Bottom) Complete random walk time series and highlighted motif subsections.

Synthetic Amplitude Scaling and Shifting Data

We inserted fixed-length peaks into random positions in the dataset. Each peak differs in magnitude and begins and ends at different magnitudes. In this particular experiment, our approach discovered the inserted motif in 12 out of the top 30 candidates; SRP discovered 0 out of the top 30 candidates. Figure 5 shows one of the discovered motifs.

Figure 5. (Top) Motif subsequences overlaid from SAX representation of original data set; and (Bottom) Complete random walk time series and highlighted motif subsections.

Winding Data

The *Winding* dataset is an industrial time series and we consider the sensor readings recording the angular velocity of a single reel sampled at 10 Hz. This dataset is a baseline for identifying known motifs (see Figure 6 for one example). In a comparison between our algorithm and SRP, both successful identify known motifs in the dataset. In the top 30 results, our algorithm discovers more motifs than SRP.

Figure 6. (Top) Motif subsequences overlaid from original data set; (Middle) Motif subsequences

overlaid from SAX representation of original data set; and (Bottom) Industrial winding data time series of the angular velocity of reel two complete time series and highlighted motif subsections.

Stock Data

We considered the closing price of Exxon Mobile (XOM) between September 1, 2006 and November 1, 2008. In stock prices, analysts have identified a phenomenon called *head and shoulders* -- a motif -- which is a bearish indicator. The pattern is exemplified by a small peak followed by a large peak and then another small peak. In Figure 7, two instances of this motif are highlighted. Our augmented representation discovered several examples of this motif.

Figure 7. Motif subsequences overlaid from original data set; (Middle) Motif subsequences overlaid from SAX representation of original data set; and (Bottom) *Head-and-shoulders* motif discovered in Exxon Mobile closing price stock data from September 1, 2006 through November 1, 2008.

Exercise Data

As a low-cost proxy for more extensive on-body sensors, we recorded a collection of outdoor exercise routines using a Garmin Forerunner 305. The Forerunner 305 is a GPS-receiver training watch with a heart rate monitoring strap that samples data at 1 Hz. The device enabled us to record multi-variate time series data of a variety of activities. For example, in the time series in Figure 8, a user runs around the campus quad and slows to a walk at three random points in the dataset. The data points are speed over time and extremely noisy compared to other datasets. Our augmented representation discovered the variable walking motifs.

Figure 8. (Top) Motif subsequences overlaid from original data set; (Middle) Motif subsequences overlaid from SAX representation of original data set; and (Bottom) Complete Garmin Forerunner speed time series and highlighted motif subsections.

Discussion

The results from our experiments indicate that augmented features allows for discovery of previously ruled out motif pair possibilities. We make two critical observations about our feature selection. First, there exists a many-to-one mapping between SAX string representations and additional feature vectors. That is, two equivalent SAX strings map to the same derived features, and additional unequal SAX strings also map to the same derived features.

This is useful for two reasons: 1) our augmented representation hashes to the same location the same number of times as the original representation; and 2) the multiple strings that map to the same additional features are those that we want to consider as motifs. Second, after empirical analysis, we determined that including the SAX string and the additional features together as the representation decreased the performance of the algorithm. Instead, we exclusively use the derived features and throw out the original SAX strings.

Spurrious Motifs

Not all of the top results from the collision matrix contain actual motifs. In the random walk data, we discovered several spurious motifs. For example, in Figure 9, we see two sections of the time series which are random data. The motifs are clearly visible and the algorithm discovers them, but also claims that this pair is a motif.

Figure 9. (Top) Motif subsequences overlaid from original data set; (Middle) Motif subsequences overlaid from SAX representation of original data set; and (Bottom) Complete random walk time series with inserted peaks and sine waves and highlighted spurious motif subsections

Conclusion and Future Work

In this paper, we presented motivation to augment traditional discretization of time series data with derived features used in other domains (e.g., speech processing). These additional features, combined with the strict ordering of the SAX alphabet capture useful time series dynamics. Our representation is robust with respect to uniform shifts in amplitude and amplitude scaling.

We presented a survey of empirical results that provided: 1) a sanity check that our representation indeed discovers the same motifs as the prior approach and representation; 2) an application of the representation to a domain (e.g., finding head-and-shoulder motifs that vary in amplitude) where the motifs would be downrated as non-matches; and 3) an application to exercise data.

Future work will proceed in several directions. First, we evaluated our approach using univariate time series. We will focus on multivariate time series and addressing issues of dimensionality. Our representation choice is applicable for use in the related work on subdimensional motif discovery (Minnen et al. 2007) and uniform scaling in the time dimension (Yankov et al. 2007) -- both multivariate approaches. We will explore the impacts that our representation choices have on these multivariate time series motif-discovery algorithms. Second, the inspiration for the augmented features comes from speech waveforms -- a potentially massive data source filled with interesting motifs. We will investigate how appropriate the SAX representation may be for speech and applications of motif-discovery algorithms to speech.

References

Buhler, J., and Tompa, M 2002. Finding motifs using random projections. Journal of computational Biology 9(2):225-242.

Catalano, J.; Armstrong, T.; and Oates, T. 2006. Discovering patterns in real-valued time series. In

Proceedings of the 10th European Conference on Principles and Practice of Knowledege Discovery in Databases (PKDD).

Catalano, J.; Armstrong, T.; and Oates, T. 2006. Discovering patterns in real-valued time series. In Proceedings of the 10th European Conference on Principles and Practice of Knowledege Discovery in Databases (PKDD).

Chiu, B.; Keogh, E.; and Lonardi, S. 2003. Probabilistic discovery of time series motifs. In 9th International Conference on Knowledge Discovery and Data Mining (SIGKDD'03), 493–498.

Jurafsky, D., and Martin, J. 2000. Speech and language processing. Prentice Hall New York.

Lin, J.; Keogh, E.; Lonardi, S.; and Patel, P. 2002. Finding motifs in time series. In Proceedings of the Second Workshop on Temporal Data Mining.

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.

Minnen, D.; Isbell, C.; Essa, I.; and Starner, T. 2007a. Detecting subdimensional motifs: An efficient algorithm for generalized multivariate pattern discovery. In IEEE Int. Conf. on Data Mining (ICDM), volume 1.

Minnen, D.; Starner, T.; Essa, I.; and Isbell, C. 2007b. Improving activity discovery with automatic neighborhood estimation. In International Joint Conference on Artificial Intelligence, 6–12.

Oates, T.; Schmill, M. D.; and Cohen, P. R. 2000. A method for clustering the experiences of a mobile robot that accords with human judgments. In AAAI/IAAI, 846–851.

Oates, T. 1999. Identifying distinctive subsequences in multivariate time series by clustering. In Chaudhuri, S., and Madigan, D., eds., Fifth International Conference on Knowledge Discovery and Data Mining, 322–326. San Diego, CA, USA: ACM Press.

Pevzner, P., and Sze, S. 2000. Combinatorial approaches to finding subtle signals in DNA sequences. In Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology, volume 8, 269–278. Citeseer.

Sagot, M. 1998. Spelling approximate repeated or common motifs using a suffix tree. Lecture Notes in Computer Science 1380:374–390.

Vahdatpour, A.; Amini, N.; and Sarrafzadeh, M. 2009. Toward unsupervised activity discovery using multidimensional motif detection in time series. In IJCAI, 1261–1266.

Yankov, D.; Keogh, E.; Medina, J.; Chiu, B.; and Zordan, V. 2007. Detecting time series motifs under uniform scaling. In KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, 844–853. New York, NY, USA: ACM.

Yasser Mohammad and Toyoaki Nishida. 2009. Constrained Motif Discovery in Time Series. New Generation Computing 27(4):319–346