

Activity Discovery: Sparse Motifs from Multivariate Time Series

David Minnen Thad Starner Irfan Essa Charles Isbell
College of Computing
Georgia Institute of Technology
Atlanta, GA 30332
dminn | thad | irfan | isbell @cc.gatech.edu

In a set of time series or other sequence data, a *motif* is a collection of relatively short subsequences that exhibit high self-similarity yet are distinguishable from other subsequences of the data. Typically, the occurrence of a motif corresponds to some meaningful aspect of the data such as a particular structure or binding site in biological sequences, a spoken word in speech data, or a specific robot behavior or response pattern. We address the problem of *activity discovery*, which deals with locating and modeling motifs in multivariate time series such as those captured by on-body sensors or from a video camera observing people engaged in some activity. We extend previous work in motif discovery to derive an algorithm that handles non-linear time warping, variable-length motifs, and which is efficient even when the motif occurrences are sparse relative to the full dataset.

In bioinformatics, systems such MEME [1] were developed to discover motifs in DNA and protein sequences, while Jensen *et al.* [4] recently generalized motif discovery over both categorical and continuous data and across arbitrary similarity metrics. These algorithms were developed for sequences, however, and do not account for the dynamic nature of time series data. Within the data mining community, an efficient, probabilistic algorithm for motif discovery using locality-sensitive hashing was developed [2]. This approach only discovers fixed-length motifs in univariate data, however. Tanaka and Uehara generalized the approach to work with multivariate time series and to allow variable length motifs but not time warping [6]. Their solution is simply to use a univariate algorithm with the first principal component of the time series, a transformation that will often mask many motifs. Finally, the PERUSE algorithm discovers motifs in multivariate time series and allows non-linear time warping and variable-length motifs [5]. This approach, however, assumes that the motifs are densely distributed and is not efficient for sparse data.

Our approach to activity discovery proceeds in three main phases: (1) motif seed discovery via analysis of a quantized representation, (2) seed refinement in the continuous domain, and (3) occurrence detection using probabilistic models trained from the refined seeds. Fundamentally, activity discovery is difficult because little information is known about the motifs ahead of time. Specifically, the discovery system does not know the number of motifs, the location or length of the occurrences, or the shape of each motif. In order to deal with this lack of knowledge and still discover the motifs efficiently, our approach transforms the continuous, multivariate time series into strings of discrete symbols and utilizes a generalized suffix tree for linear-time subsequence searches [3].

Each unique subsequence with a user-specified length is used as a query to retrieve all of the occurrences in the dataset while allowing for dynamic time warping. The motif representing the most information, accounting for both the motif complexity and the number of occurrences, is selected and removed. This process repeats until the amount of information represented by the best motif is too small.

Our algorithm then refines this set of initial seed motifs. Refinement consists of merging, splitting, and temporal extension. In the splitting phase, the occurrences of each seed motif are analyzed using agglomerative clustering to determine if the motif is actually a combination of two different motifs. The merging phase

Topic: data mining, learning algorithms

Preference: oral/poster

then checks to see if a single motif is represented by multiple seeds using a similar agglomerative clustering method. Finally, the occurrences of each motif are temporally extended if the variance of the preceding or trailing frames is comparable to the variance already within the seed motif.

The final phase of our approach uses the refined seed motifs to train a set of probabilistic models, which are then matched to the time series data to find all of the motif occurrences. We use a model based on that of the PERUSE algorithm, which is similar to a left-right hidden Markov model [5]. Matching is kept efficient despite the unknown occurrence boundaries and dynamic time warping by appropriately initializing the alignment trellis to require only a single pass along each time series. This leads to a linear rather than a quadratic matching algorithm.

We have begun to experimentally evaluate our approach using data captured from a mock exercise regime composed of six different dumbbell exercises. A sensor was attached to the subject's wrist that contained a three-axis accelerometer and gyroscope. In total, 32 time series were captured, averaging 52s each and consisting of 864 repetitions (roughly 144 of each of the six exercises). Our algorithm correctly determined that the data consists of six motifs and successfully located 96.3% of the occurrences. The system correctly identified 832 occurrences but incorrectly located 51 occurrences (insertion errors) and failed to detect 32 real occurrences (deletion errors), leading to an overall accuracy of 86.7% with a precision of 88.4%.

Although these results are quite good considering the difficulty of the activity discovery problem, further validation is required. This validation includes testing the system on other datasets taken in different domains and with different sensors, ensuring the system scales with larger datasets, and exploring its sensitivity to different parameters settings. To this end, we are currently evaluating the system on an American Sign Language dataset consisting of 500 sentences composed from a 40 word vocabulary and captured at 10Hz by a video camera. We are also planning to evaluate our approach using English speech data, Kung Fu forms captured by an on-body accelerometer, and with another ASL dataset captured with a glove that measures finger posture.

We have presented a new approach for activity discovery, the problem of locating and modeling variable-length motifs in multivariate time series representing human activity data. Our method accounts for possible non-linear time warping of each motif occurrence and is efficient even when the motifs are sparsely distributed relative to the total dataset. In light of surprisingly accurate results on a real dataset, we are currently evaluating the approach on additional datasets from a variety of domains and sensors. We are also working on reducing the number of user-specified parameters as well as the system's sensitivity to these parameters. Finally, we are investigating methods for learning higher-level structure from the discovered activities and utilizing such higher-level information to improve discovery performance.

References

- [1] T. Bailey and C. Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In *Second International Conference on Intelligent Systems for Molecular Biology*, pages 28–36. AAAI Press, 1994.
- [2] B. Chiu, E. Keogh, and S. Lonardi. Probabilistic discovery of time series motifs. In *Conf. on Knowledge Discovery in Data*, pages 493–498, 2003.
- [3] D. Gusfield. *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge University Press, 1997.
- [4] K. Jensen, M. P. Styczynski, I. Rigoutsos, and G. Stephanopoulos. A generic motif discovery algorithm for sequential data. *Bioinformatics*, 22(1):21–28, 2006.
- [5] T. Oates. PERUSE: An unsupervised algorithm for finding recurring patterns in time series. In *Int. Conf. on Data Mining*, pages 330–337, 2002.
- [6] Y. Tanaka and K. Uehara. Discover motifs in multi-dimensional time-series using the principal component analysis and the mdl principle. In *International Conference on Machine Learning and Data Mining*, pages 252–265, 2003.