Multivariate Time Series Classification Using Inter-leaved Shapelets

Om Prasad Patri
Department of Computer Science
University of Southern California
Los Angeles, CA 90089
patri@usc.edu

Rajgopal Kannan
School of EECS
Louisiana State University
Baton Rouge, LA 70803
rkannan@csc.lsu.edu

Anand V. Panangadan
Department of Computer Science
California State University, Fullerton
Fullerton, CA 92834
apanangadan@fullerton.edu

Viktor K. Prasanna
Ming-Hsieh Department of Electrical Engineering
University of Southern California
Los Angeles, CA 90089
prasanna@usc.edu

Abstract

We propose an approach for classifying multivariate time series based on extracting *shapelets*, which are discriminative subsequences within time series capable of differentiating between different classes. While univariate shapelet extraction methods are well established, there has been little effort towards extending them for the multivariate scenario. The few approaches that do exist for multivariate shapelet-based classification make the assumption that the data from individual sensors is independent of each other. In industrial processes such as manufacturing, we can observe the effect of a change in one sensor affect the data from another sensor, as all the sensors record data synchronized in time. We attempt to incorporate these temporal dependencies across sensors through a simple interleaving heuristic, and evaluate our approach on a real silicon wafer manufacturing dataset.

1 Introduction

The number of sensors used in real-world applications and the amount of measurements from these sensors has been steadily increasing over the years, and has accelerated with the rise of the Internet-of-Things. Sensor data from such real-world processes are typically not independent and identically distributed (i.i.d.) and not drawn from a fixed distribution. This has led to the development of methods such as Shapelets [1] which rely on matching discriminative subsequences that are learned from a labeled training dataset and hence do not place any assumptions or restrictions on the structure of data (unlike autoregressive and ARIMA time series models [2]). Most recent work on Shapelet-based methods have focused on univariate time-series classification. In this work, we present a novel extension of the Shapelets method to address the problem of multivariate time-series classification.

Time series shapelets [1] focus on extracting discriminative subsequences within the training set that are most relevant for distinguishing between the positive and negative classes. A shapelet is a subsequence or local temporal pattern in a time series that is a representative feature of the class to which this time series belongs.

The advantage of using a shapelet-based approach is its fast classification time because most of the computational complexity is in extracting shapelets from the training data; this training dataset is not needed during classification. Shapelets are visually interpretable and domain experts can use

them for deeper investigation and root cause analysis. Shapelets have been shown to be effective and highly accurate for a variety of time series data mining tasks including classification, clustering, summarization, and visualization [1, 3, 4, 5, 6, 7, 8].

However, most algorithms for shapelet mining only accept univariate time series data. There are a few approaches (such as [5, 7, 9]) which propose extensions of the univariate shapelet extraction method for multivariate use cases but all of them make the assumption that shapelets can be extracted from different sensors independently of each other. This assumption does not hold for complex, real-world datasets. For example, in the quality monitoring step of a manufacturing process, an anomalous measurement in one of the sensors is also likely to be reflected in other sensors.

In this work, we explore interleaving measurements from multiple sensors as a means of applying univariate shapelet-based classification algorithms to multivariate time-series. Following this step, shapelets can be extracted by jointly considering the discrimination ability *across all sensors*. Interleaving is treated as a generalization of the approach of concatenating time-series from different sensors. While naive concatenation also enables measurements from multiple time-series to be considered together, shapelets cannot be extracted since the size of the shapelet search window depends on the length of each time-series itself. We demonstrate that the interleaving-based approach solves this issue for shapelets and thus leads to an effective multivariate time-series classification algorithm. We evaluate this approach on publicly available datasets that are widely used in time-series classification research and show that the proposed method equals or outperforms state-of-the-art shapelets methods in terms of classification accuracy.

2 Related Work

There are several approaches possible for extending the univariate shapelet extraction method (the state-of-the-art being Fast Shapelets [4]) to multivariate data. A simple approach is to concatenate all the dimensions sequentially to convert multi-dimensional to one-dimensional data as proposed by Mueen et al. [3]. Ghalwash et al. [5] propose shapelet extraction through solving a convex-concave optimization problem which has a restriction of extracting only one subsequence per dimension.

The two previous approaches most closely related to this work are Shapelet Forests [6] and Shapelet Ensembles [9]. The Shapelet Forests approach combines shapelet extraction with feature selection. Univariate shapelets are extracted from each dimension, shapelet decision trees are formed and the final prediction is the weighted decision of trees from all dimensions, the weights for sensors being learned from the data. The Shapelet Ensembles approach uses a majority-based voting from the decision trees to set up an ensembling system - giving rise to an ensemble of decision trees. This ensembling approach is contrasted against a concatenation-based approach.

Executing shapelet-based approaches can be slow, especially when using long time series instances, since the complexity is quadratic in the length of time series (m) but linear in the number of time series (n), i.e. $O(nm^2)$. Therefore, efforts have focused on efficient pruning techniques including Logical Shapelets [3] and Fast Shapelets methods [4]. An alternative approach for shapelet extraction using stochastic gradient learning was proposed by Grabocka et al. [10].

3 Approach

We now describe how aspects from both concatenation and ensemble-based methods can be combined to yield a novel multivariate shapelet-based time-series classification algorithm that is more effective than existing methods for cases where the relative position of discriminative subsequences in different sensors varies.

We first illustrate the intuition behind our approach. Consider the case with two sensors where a time-series is to be classified to be in the positive class if a distinct shape (e.g., a peak) appears in Sensor 1's data stream followed by its appearance in Sensor 2. The appearance of these shapes in any other order indicates that the time-series is to be classified as the negative class. Two instances of this type are shown in Figure 1. In this case, most existing multivariate shapelet-based classification algorithms that extract shapelets from each sensor *independently* of the other would fail since the mere presence or absence of the shapelets is not sufficiently discriminative. On the other hand, if the two sensor streams are concatenated, a (single) discriminative shapelet composed of the distinct

shapes from each sensor can be identified, as shown in Figure 2. However, the shapelets extracted based on concatenating fixed length segments is *not* local – the length of the shapelet is dependent on the length of the window used for concatenation. In addition, the discriminative capacity of such multi-variate shapelets depends on the relevance of each sensor stream to the class – the inclusion of time-series segments from irrelevant sensors in the concatenated shapelet will increase the error rate in shape pattern matching during classification. We have therefore developed an interleaving and sensor ranking-based approach to make the extracted multi-variate shapelets invariant to the length of segments and number of sensors.

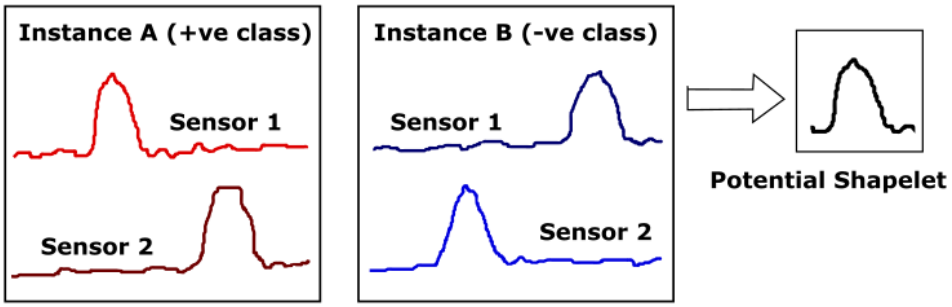

Figure 1: Two multivariate instances each consisting of two time series (two sensors). The peak in the data is a potential shapelet candidate but it is NOT discriminative enough to differentiate between the two instances (which belong to opposite classes), because a similar pattern occurs in all four time series.

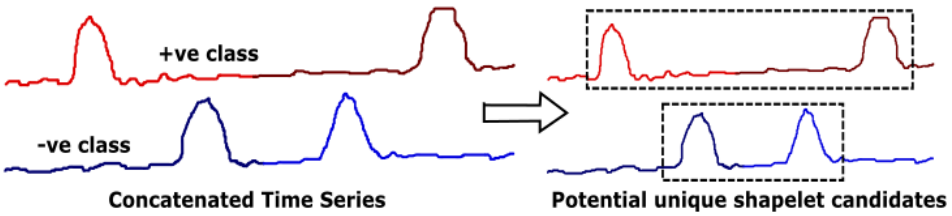

Figure 2: Time series formed by concatenating data from all sensors for each instance in the previous example. Now the potential shapelet candidate from the concatenated time series can discriminate between the two classes.

Our proposed approach is called *Inter-leaved Shapelets (ILS)*, and the overall approach is depicted in Figure 3. The idea is to inter-leave time series segments across sensors from multiple dimensions to form the final concatenated one-dimensional time series for each instance. The simplest way to do this is to consider a fixed interval inter-leaving - segment size k at which we regularly cut the time series and inter-leave the segments, i.e. first k elements of the first sensor, then first k elements of the second sensor and so on until the first k elements of the last sensor, after which we concatenate the $(k+1)^{th}$ to $2k^{th}$ element of the first sensor again and so on. In the end, we will have the same number of univariate time series as the number of instances.

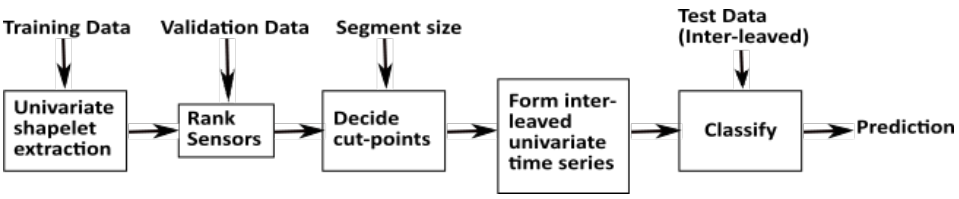

Figure 3: Inter-leaved Shapelets Approach

To improvise upon our naive approach, we rank the sensors before performing the inter-leaving and order the data according to the sensor ranks - from most important to the least. By doing this, we ensure that (i) the shapelet extraction algorithm encounters data from the highly ranked sensors first where a shapelet is more likely to be found, and (i) we can eliminate data from the lower ranked sensors completely leaving them out of the inter-leaving process (they will be at the end of each segment concatenation round). To implement this ranking scheme, we divide the training data into a training and validation set. The ranking of the sensors is based upon using the shapelets extracted from the training set to perform classification on the validation set.

Note that ILS can be extended by determining determining the "cut-points" for inter-leaving automatically depending on detection of change-point events [11, 12] in the data. This can ensure that we do *not* segment the time series when a change in structure is detected (such as the middle of a data peak) and retain the shape of the change structure. It also reduces the number of cuts if nothing anomalous is detected (in contrast with the current approach of segmenting at fixed intervals).

4 Evaluation

We use a real-world multivariate time series dataset, Wafer, which is publicly available[1], to evaluate the performance of our proposed ILS approach and compare it to the state-of-the-art. The Wafer data is from a process used for manufacturing silicon wafers for semiconductor chips. It consists of 1194 multivariate instances with 6 time series in each instance. The data is heavily imbalanced – only 127 of the 1194 instances belong to the anomaly class, the rest being normal.

The results are shown in Table 4. We experiment on two datasets, both created with 75%-25% train-test splits - (i) the full Wafer dataset, as provided with the training and testing partitions, and (ii) a balanced subset of the Wafer dataset with a 1:1 ratio of positive and negative class instances (positive instances being randomly chosen, and all negative instances being used). In either case, we use one-third of the provided training dataset for our validation set.

We experiment with a range of segment sizes for fixed interval ILS, denoted by ILS-k where k is the segment size. We also compare our ILS approach to two state-of-the-art approaches – Shapelet Forests (SF) [7] (majority voting feature selection considered) as well as Shapelet Ensembles (SE) [9] (step ratio parameter set to the default 0.95). We can observe that our proposed approach equals or outperforms both the state-of-the-art methods on the full as well as balanced datasets for appropriate segment sizes.

Dataset	SF	SE	ILS-2	ILS-10	ILS-20	ILS-25	ILS-30	ILS-50
Full	98.0	92.8	91.0	96.7	**98.3**	97.7	92.6	90.0
Balanced	96.9	82.8	90.6	95.3	87.5	87.5	**96.9**	84.4

Table 1: Classification accuracy on the Wafer dataset. ILS-k is the proposed approach where k is the segment size.

5 Conclusion

This work presented our initial attempts towards a multivariate shapelet-based time series classification approach based on a inter-leaving approach while performing concatenation of time series. Even with a simple fixed interval cut-point heuristic, we were able to outperform baseline and state-of-the-art approaches. In the future, we will perform more rigorous evaluation of this approach and work on methods to determine the cut-points automatically through change-point detection, as well as eliminate or filter data from lower ranked sensors as required.

Acknowledgments

This work is supported by Chevron U.S.A. Inc. under the joint project, Center for Interactive Smart Oilfield Technologies (CiSoft), at the University of Southern California.

[1] At `http://www.cs.cmu.edu/~bobski/data/`

References

[1] Lexiang Ye and Eamonn Keogh. Time series shapelets: a novel technique that allows accurate, interpretable and fast classification. *Data Mining and Knowledge Discovery*, 22(1-2):149–182, 2011.

[2] Tak-chung Fu. A review on time series data mining. *Engineering Applications of Artificial Intelligence*, 24(1):164–181, 2011.

[3] Abdullah Mueen, Eamonn Keogh, and Neal Young. Logical shapelets: an expressive primitive for time series classification. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1154–1162. ACM, 2011.

[4] T. Rakthanmanon and E. Keogh. Fast shapelets: A scalable algorithm for discovering time series shapelets. *Proceedings of the 13th SIAM International Conference on Data Mining*, 2013.

[5] Mohamed F Ghalwash, Vladan Radosavljevic, and Zoran Obradovic. Extraction of interpretable multivariate patterns for early diagnostics. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, pages 201–210. IEEE, 2013.

[6] Om P. Patri, A. Panangadan, C. Chelmis, and V. K. Prasanna. Extracting discriminative features for event-based electricity disaggregation. In *Proceedings of the 2nd Annual IEEE Conference on Technologies for Sustainability*. IEEE, 2014.

[7] Om P. Patri, Abhishek B. Sharma, Haifeng Chen, Guofei Jiang, Anand V. Panangadan, and Viktor K. Prasanna. Extracting discriminative shapelets from heterogeneous sensor data. In *Big Data (Big Data), 2014 IEEE International Conference on*, pages 1095–1104. IEEE, 2014.

[8] Om P. Patri, Nabor Reyna, Anand Panangadan, Viktor Prasanna, et al. Predicting compressor valve failures from multi-sensor data. In *SPE Western Regional Meeting*. Society of Petroleum Engineers, 2015.

[9] Mustafa S Cetin, Abdullah Mueen, and Vince D Calhoun. Shapelet ensemble for multi-dimensional time series. In *Proceedings of the 15th SIAM International Conference on Data Mining*. SIAM, 2015.

[10] Josif Grabocka, Nicolas Schilling, Martin Wistuba, and Lars Schmidt-Thieme. Learning time-series shapelets. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 392–401. ACM, 2014.

[11] Marc Lavielle and Gilles Teyssiere. Detection of multiple change-points in multivariate time series. *Lithuanian Mathematical Journal*, 46(3):287–306, 2006.

[12] Rebecca Killick and Idris Eckley. changepoint: An r package for changepoint analysis. *Journal of Statistical Software*, 58(3):1–19, 2014.