EAST Representation: Fast Discriminant Temporal Patterns Discovery From Time Series

Xavier Renard[1,3], Maria Rifqi[2], Gabriel Fricout[3] and Marcin Detyniecki[1,4]

[1]Sorbonne Universités, UPMC Univ Paris 06, CNRS, LIP6, Paris, France.
[2]Université Panthéon Assas, Univ Paris 02, LEMMA, Paris, France.
[3]Arcelormittal Research, Maizières-lès-Metz, France.
[4]Polish Academy of Sciences, IBS PAN, Warsaw, Poland.

Abstract. Mining discriminant temporal patterns is one problem for the time series classification currently led by the shapelet. We expose this issue from the perspective of a standard feature-space classification task. This approach is enabled by the recent observation that most enumerable subsequences from a time series are redundant and can be discarded. In addition to its simplicity the approach has state-of-the-art classification performances with extremely fast computations. It also provides a flexible framework with interesting perspectives.

1 Introduction & related work

Two main approaches exist to discover localized and phase independent discriminant temporal patterns from time series: the shapelets and the derivation of features from intervals of the series [2]. We focus here on the shapelet principle. We invite the reader to refer to the seminal article [12] for details, to summarize the shapelet discovery relies on three steps.

Step 1 Exhaustive enumeration of the subsequences from a set of time series.
Step 2 Evaluation of the subsequences. The minimal euclidean distances (MED) between subsequences and time series of the dataset are computed to get the *discriminatory power* (usually the information gain) of each subsequence.
Step 3 The subsequences that most consistently separate the time series with respect to their classes are conserved.

Several variations have been proposed after the original shapelet tree that learns a tree of successive discriminant shapelets at a price of a large time complexity. The shapelet transform [7] has become a classical instantiation of the shapelet principle: the shapelet discovery is performed in one pass and the MED between time series and shapelets feed a classifier. Among the other approaches we can cite the logical shapelets to discover conjunction or disjunction of shapelets [8], the fast shapelets [9] and the learnt shapelets [4].

The shapelet has two limitations: the time complexity and, from our point of view, the independent evaluation of the *discriminatory power* of each subsequence. The first limitation results from the discovery complexity in $O(L^3N^2)$

with L the time series length and N the number of time series in the dataset [12]. The second limitation lies in the *Step 2* of subsequence evaluation (based on the information gain or a similar metrics): each shapelet has to be sufficient to discriminate a set of time series. Even if a set of characteristic subsequences is very discriminant, the shapelet discovery would fail to find it. To reduce the time complexity, several improvements have been proposed [12,9] but it remains large. It has been recently observed that most subsequences extracted from a time series are redundant. A drastic random sub-sampling among the subsequences at *Step 1* instead of an exhaustive enumeration has been shown effective to reduce the time complexity while preserving the classification performances [3,10,5].

In this work, we rely on this observation to propose a flexible representation called EAST (Enumerate And Select discriminant Temporal patterns). It aims at improving the two aforementioned limitations (time complexity and independent evaluation). We postulate that each subsequence extracted from a set of time series should be considered as a *feature* and the exhaustive set of subsequences forms a *feature vector*. The redundancy is eliminated by the random sub-sampling of the subsequences. A relevant subset of features (*i.e.* subsequences) for the classification is conserved after a feature selection stage. The originality of the approach resides in the problem formulation that enables the use of well-established feature selection techniques to perform a powerful discovery of discriminant temporal patterns. We show, with instances of our proposition, that the classification performances already reach the state-of-the-art while being extremely fast (evaluation of a few thousands subsequences at most). We also demonstrate its scalability: the number of subsequences to evaluate is independent of the number of time series in the dataset.

2 Proposition: discriminant temporal pattern discovery

We have a training set D of time series T_n with $n \in [1, 2, \ldots, N]$ where $T_n = [t_n(1), \ldots, t_n(i), \ldots, t_n(|T_n|)]$. T_n has a length $|T_n| \in [L_{min}, \ldots, L, \ldots, L_{max}]$ with $L_{min}, L, L_{max} \in \mathbb{N}^*$, where L_{min} is the smallest time series of D and L_{max} is the longest one. A subsequence s of length l at a starting position j in a time series T_n is noted as $s_j^{j+l}(T_n) = [t_n(j), \ldots, t_n(j+l-1)]$. S is the set of all the subsequences s that it is possible to extract from D, whatever their lengths and starting positions are.

The problem is framed in the field of time series classification: each time series T_n has exactly one class label $y(T_n) \in Y$. In this work, our concern is the discovery of meaningful temporal patterns to perform time series classification. We make the assumption that there exists a strongly-discriminant set of patterns $Z = \{z_1, \ldots, z_p, \ldots, z_P\}$ with $p, P \in \mathbb{N}^*$ and $|z_1|, |z_p|, |z_P| \in [1; L_{max}]$ where z_p is discriminant of one class or a subset of classes of Y. Z is strongly-discriminant in that it contains all the possible subsequences, which taken independently or not, are discriminant enough to solve the classification problem. The transformation of T_n using Z produces a feature vector X_n, such that a classifier f is able to learn a mapping $f(X_n) \rightarrow y(T_n)$. The transformation is based on a distance.

However Z is unknown. Our objective is to discover from D a set $\hat{Z} = \{\hat{z}_1, \ldots, \hat{z}_j, \ldots, \hat{z}_J\}$ of patterns that produces a feature vector $X^{\hat{Z}}$ such that the classification performance of $f(X^{\hat{Z}})$ is as close as possible of the one of $f(X)$ with X obtained with a transformation based on Z.

2.1 Proposition

To determine $\hat{Z}$ we propose to combine a random enumeration of subsequences from D with a feature selection stage to retain a relevant set of patterns with respect to the classification task.

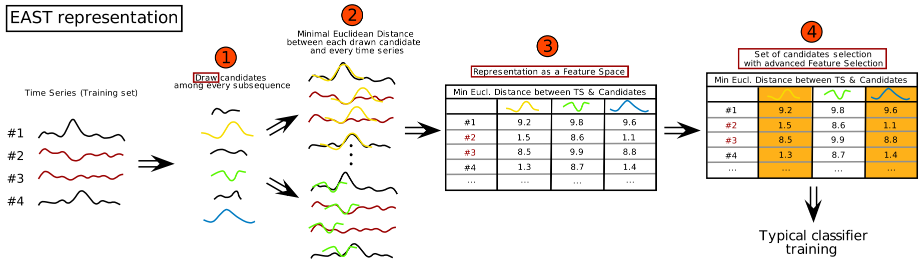
Step 1: random sub-sampling to handle subsequence redundancy The first step of our proposition relies on a random sampling $\hat{S}$, among all the subsequences S, because of the efficiency of this principle mentioned in introduction. Each subsequence $s_j^{j+l}(T_n)$ is given the same probability to be picked, whatever its time series, position and length. We demonstrate later that the number of subsequences $q = |\hat{S}|$ to draw to obtain a given classification performance is not related to the size of the dataset D (*i.e.* the number of time series).

Step 2: learning the representation by selecting a set of discriminant subsequences Once $\hat{S}$ is drawn we need to discover the set $\hat{Z} \subset \hat{S}$ that maximizes the classification performance. We propose to formalize the problem from the perspective of a standard feature-space classification task. The minimal euclidean distance d_{min} between a subsequence s with $|s| = l$ from T_1 and a time series T_2 such that:

$$d_{min}(s, T_2) = min([d(s, s_1^{1+l}(T_2)), \ldots, d(s, s_{|T_2|-l+1}^{|T_2|}(T_2))])$$

d_{min} is calculated between each subsequence of $\hat{S}$ and every time series of D. The result is a feature space X (Fig. 1) where the distances to the subsequences of $\hat{S}$ are the attributes (columns) and the time series of D are the instances (rows). The number of columns of X (*i.e.* number of attributes) may still be large and it is very likely that it contains numerous meaningless features: no selection has been performed yet with respect to the classification problem. In other terms, irrelevant subsequences $s \in \hat{S}$ are still present in the feature space X.

To reduce $\hat{S}$ to $\hat{Z}$ and derive a feature space $X^{\hat{Z}}$ relevant to train a classifier $f(X^{\hat{Z}}(T_n)) \rightarrow y(T_n)$ we use the *feature vector* formalization of the problem to exploit classical *feature selection* approaches. They allow to efficiently identify relevant attributes in a feature space with respect to a classification task. Advanced feature selection techniques offer the possibility to discover both single discriminant subsequence and sets of subsequences where each subsequence is characteristic of a class or a subclass, while the whole set is discriminant. Numerous feature selection techniques exist, the approaches used in this work are presented in the experimentation section (we use them as black boxes).


EAST representation
Time Series (Training set)
#1
#2
#3
#4
1
Draw candidates among every subsequence
2
Minimal Euclidean Distance between each drawn candidate and every time series
3
Representation as a Feature Space
4
Set of candidates selection with advanced Feature Selection
Typical classifier training

Min Eucl. Distance between TS & Candidates

#1	9.2	9.8	9.6
#2	1.5	8.6	1.1
#3	8.5	9.9	8.8
#4	1.3	8.7	1.4
...	...	...	...

Min Eucl. Distance between TS & Candidates

#1	9.2	9.8	9.6
#2	1.5	8.6	1.1
#3	8.5	9.9	8.8
#4	1.3	8.7	1.4
...	...	...	...

Fig. 1: EAST principle workflow. After a drastic subsequences sub-sampling (1), the distances between subsequences and time series (2) form a feature space of reasonable size (3) on which advanced feature selection techniques can be applied to discover discriminant set of subsequences (4).

The overall principle of the proposed approach to discover discriminant temporal patterns is summarized Fig. 2. A classifier can be trained with $X^{\hat{Z}}$. The result of the training is both a set $\hat{Z}$ of patterns and a classifier f. To perform the classification of new instances, time series are transformed into a feature vector according to $\hat{Z}$ and the classification is performed with f.

```
$\hat{S} \leftarrow$ Draw $q$ subsequences from time series from $D$
$X \leftarrow$ Calculate $d_{min}$ between time series from $D$ and subsequences from $\hat{S}$
$X^{\hat{Z}}, \hat{Z} \leftarrow$ Perform feature selection on $X$ with respect to labels of $Y$
```

Fig. 2: Learning of the EAST representation in 3 key steps: random sub-sampling, minimal euclidean distance and feature selection.

3 Experimentation

The objective of the experimentation is to evaluate the relevance of advanced feature selection in a standard feature space for the temporal pattern discovery over the classical selection scheme used by the shapelet (usually the information gain). The classification performances are observed together with the time complexity required by the pattern discovery with several configurations. For this purpose, EAST is instantiated with several feature selection approaches and classifiers for various values q of subsequences drawn. The experimentation performed for this work is framed into the classical UCR univariate time series classification framework (45 datasets from this repository are used).

With EAST, the feature selection stage is open to any approach. To perform the experimentation we use some of them. Feature selection is an established field: we do not contribute but instead we rely on it. Also, we don't advocate one approach is better than another. Feature selection methods are usually classified into three groups: *filters*, *wrappers* and *embedded methods* [6]. For the experimentation we select one *wrapper*, the Recursive Feature Elimination with cross-validation associated with a linear SVM (named RFE+SVM), and two *embedded methods*, the Randomized Logistic Regression (RLR) and the Random Forest (RF). For the RLR we test two classifiers: a SVM with a RBF kernel and a random forest (respectively named RLR+SVM and RLR+RF). These approaches are able to learn combinations or sets of features (*i.e.* subsequences). On the contrary, the shapelet approach, which usually makes use of the information gain that is part of the *filters*, is unable to learn such sets or combinations. For the random draw of $\hat{S}$ several values $q = |\hat{S}|$ are tested: $q \in [50, 100, 500, 1000, 2000, 5000]$.

The results are compared with the current leading shapelet approach, the shapelet ensemble (SHPT) [2]. The authors state that shapelet ensemble performs identically or better than other shapelet approaches. We reproduce here their results. We also compare the results with the random-shapelet (RSHPT) [10] that has the same selection stage than the classical shapelet but on a small fraction of the exhaustive set of subsequences. The same number of subsequences is picked for the random-shapelets and for the EAST instantiations.

A strict evaluation protocol is required to assess the EAST representation and the random-shapelets because they contain a random generation step. We rely on the evaluation protocol proposed in [1] for a proper way to analyze the performances of randomized algorithms. Each single test of the the EAST representation and the random-shapelets is reproduced 10 times to evaluate the variability. The complete description of the evaluation protocol of the provided results is described in additional material [11].

3.1 Results

Classification performances

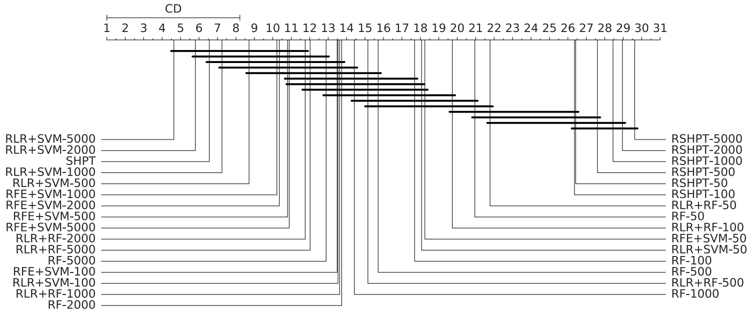


Fig. 3: Comparison of the approaches with the Nemenyi test. Groups of approaches not significantly different ($\alpha = 0.05$) are connected. CD is the critical difference.

The classification performances of the proposed approach are significantly similar to the ones obtained by the shapelet ensemble (Figure 3 & 4). These performances are obtained with only 2000 subsequences drawn for the best implementation of the proposition that is an infinitesimal fraction of subsequences evaluated by the exhaustive shapelet ensemble: the largest tested UCR dataset reaches 5.10^8 subsequences. The relevance of advanced feature selection approaches over independent discriminant subsequences is also shown by the experimentation. With the same number of subsequences drawn, our proposition systematically outperforms a random sampling associated with the classical shapelet evaluation procedure based upon the information gain. The parameter q is obviously critical, but until a certain point: with this experimentation we observe no statistical difference in the classification performances between the best performing configuration with $q = 2000$, $q = 5000$ and those of the shapelet ensemble, the state-of-the-art (Figure 3). It is also worth noting the low standard deviation in the performances, in particular for the RLR+SVM approach, with a maximum of 2.7% (for one single dataset) with most standard deviation below or around 1% (for $q = 2000$). Raw results of the experimentation are available in additional material [11].

	RLR+SVM-5000	**SHPT**	RLR+SVM-2000	RLR+SVM-1000	RLR+SVM-500	RFE+SVM-500	RFE+SVM-1000	RLR+RF-2000	RLR+RF-5000	RFE+SVM-5000	RFE+SVM-2000	RF-5000	RLR+RF-1000	RF-2000	RFE+SVM-100	RLR+SVM-100	RF-1000	RLR+RF-500	RF-500	RLR+SVM-50	RFE+SVM-50	RF-100	RLR+RF-100	RF-50	RLR+RF-50	**RSHPT-50**	**RSHPT-100**	**RSHPT-500**	**RSHPT-1000**	**RSHPT-2000**	**RSHPT-5000**
Score	29	**27**	27	25	18	16	15	12	12	11	11	8	6	5	3	3	1	1	-4	-10	-11	-11	-11	-15	-18	**-25**	**-25**	**-25**	**-25**	**-25**	**-25**

Fig. 4: Scores of all the approaches based on Wilcoxon tests. Higher is better.

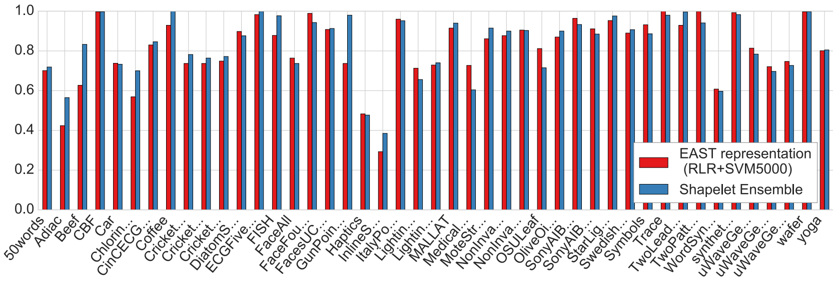


Fig. 5: Average classification performances of the EAST representation (RLR+SVM 5000) vs. the exhaustive shapelet ensemble. On the UCR benchmark the two approaches have similar performances with a drastic time complexity reduction for the EAST approach.

Time complexity For the approaches used for this experimentation, most of the time is spent in the distance calculations between each subsequence enumerated and the time series. This fact is illustrated figure 6 and is especially true for large datasets and with values of q increasing: the time spent in the feature selection becomes insignificant. We use this specificity to compare the time complexity of the approaches and avoid implementation or hardware bias.

The EAST representation enumerates a fixed number of subsequences, in this work the maximal value is $q = 5000$. The typical shapelet approach performs an exhaustive enumeration. Figure 7 shows that the exhaustive shapelet discovery (SHPT) evaluates subsequences sets several order of magnitude larger than the EAST approach while having comparable classification performances. For the datasets used in the experimentation, the exhaustive number of subsequences to extract varies from 20,100 (Italy Power Demand) to 524,800,000 (Star Light Curves). For all the datasets and for significantly similar classification performances our proposition uses $q = 2000$ subsequences. On the bigger dataset this is less than 0.0002% of the exhaustive number of subsequences. The exhaustive number of subsequences depends on L (time series length) and N (number of time series in D). We demonstrate that the number of subsequences to draw from S to determine $\hat{Z}$ is not dependent of N. This allows a considerable gain for the training phase on datasets with numerous time series. The following lemma is demonstrated in additional material [11].

Lemma The probability of drawing a relevant subsequence $\hat{z}_j$ for the classification task is independent of the number of time series N in D.

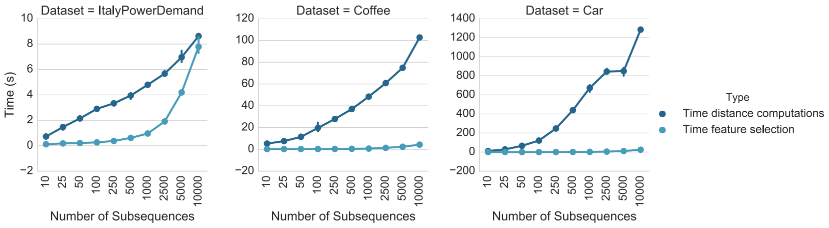


Fig. 6: Time spent in the distance calculations vs. Time spent in the feature selection for EAST. For small datasets (ItalyPowerDemand), the feature selection requires a similar amount of time than the distance calculations. For larger datasets (Coffee, Car) feature selection becomes insignificant in front of distance computations. We use this specificity to compare the time complexity of the approaches based on the number of distance computations and avoid implementation or hardware bias.

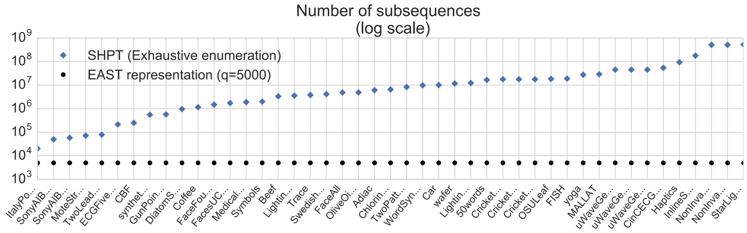


Fig. 7: Number of subsequences evaluated by EAST and the exhaustive shapelet discovery (log scale). EAST enumerates a constant number of subsequences over the datasets with comparable classification accuracies than the shapelet ensemble that generates subsequences sets several orders of magnitude larger.

4 Conclusion

This work evaluates advanced feature selection relevance to discover discriminant temporal patterns for time series classification. This approach is enabled by the previous observation that most subsequences in a time series are redundant and can be discarded. We state that each subsequence represented by its distance to the time series is a feature in a feature vector on which classical feature selection can be applied. The experimentation on 45 datasets of the UCR shows significantly similar classification performances to the state-of-the-art with a time complexity drastically reduced. Moreover the scalability of the approach is demonstrated. The proposed approach may allow the discovery of sophisticated patterns, such as multivariate patterns, thanks to the use of advanced feature selection: this study is our next step.

References

1. A. Arcuri and L. Briand. A Hitchhiker's guide to statistical tests for assessing randomized algorithms in software engineering. *Software Testing Verification and Reliability*, 24(3):219–250, 2014.
2. A. Bagnall, J. Lines, J. Hills, and A. Bostrom. Time-Series Classification with COTE: The Collective of Transformation-Based Ensembles. *IEEE Transactions on Knowledge and Data Engineering*, 27(9):1–10, 2014.
3. D. Gordon, D. Hendler, and L. Rokach. Fast Randomized Model Generation for Shapelet-Based Time Series Classification. *arXiv*, sep 2012.
4. J. Grabocka, N. Schilling, M. Wistuba, and L. Schmidt-Thieme. Learning Time-series Shapelets. *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 392–401, 2014.
5. J. Grabocka, M. Wistuba, and L. Schmidt-Thieme. Fast classification of univariate and multivariate time series through shapelet discovery. *Knowledge and Information Systems*, pages 1–26, 2015.
6. I. Guyon and A. Elisseeff. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.
7. J. Hills, J. Lines, E. Baranauskas, J. Mapp, and A. Bagnall. Classification of time series by shapelet transformation. *Data Mining and Knowledge Discovery*, 28(4):851–881, may 2014.
8. A. Mueen, E. Keogh, and N. Young. Logical-shapelets: an expressive primitive for time series classification. *The 17th ACM SIGKDD international conference*, pages 1154–1162, 2011.
9. T. Rakthanmanon and E. Keogh. Fast shapelets: A scalable algorithm for discovering time series shapelets. *Proceedings of the thirteenth SIAM conference on data mining (SDM)*, pages 668–676, 2013.
10. X. Renard, M. Rifqi, W. Erray, and M. Detyniecki. Random-shapelet : an algorithm for fast shapelet discovery. *IEEE International Conference on Data Science and Advanced Analytics*, pages 1–10, 2015.
11. X. Renard, M. Rifqi, G. Fricout, and M. Detyniecki. https://github.com/xrenard/EAST-Representation, 2016.
12. L. Ye and E. Keogh. Time series shapelets: A New Primitive for Data Mining. *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '09*, page 947, 2009.