Improving the Discovery of Shapelets in Time Series for Thunderstorm Classification

M. Geuzaine[1,*], M. Arul[2] and A. Kareem[1]

[1] NatHaz Modeling Lab, Dept. of Civil and Environmental Engineering and Earth Sciences, University of Notre Dame,
156 Fitzpatrick Hall of Engineering, Notre Dame, IN 46556, (USA)
mgeuzai4@nd.edu; kareem@nd.edu

[2] The Charles E. Via, Jr. Department of Civil and Environmental Engineering, Virginia Tech,
102-B Patton Hall, VA 24061, (USA)
marul@vt.edu

Abstract A large volume of data related to wind has been recorded over the last decade in Mediterranean ports in order to better understand the effects of thunderstorm winds on civil engineering structures. Automated classification techniques have thus been developed to detect these events of interest in such large databases. To ensure the autonomy and interpretability of the process, it is convenient to use a machine learning classifier trained on shapelet transforms. Techniques such as randomized sampling method, rotation forest classifier and ensemble voting rule are utilized in this paper to accelerate the discovery of shapelets while continuing to increase the accuracy and the stability of the procedure. These improvements in terms of both computational and operational efficiency regarding the identification of thunderstorms are assessed in this paper. Overall, the time needed to discover shapelets is divided by 10 to 100 when the same ratio of candidates is selected from the entire pool. Doing so does not harm the classification process afterwards, even though the best shapelets are slightly less discriminative.

Keywords: shapelet transform, thunderstorm, classification, machine learning

1 Introduction

To gain a better understanding of the devastating effects that thunderstorm winds can have on civil engineering structures, more particularly in Mediterranean ports, extensive measurement campaigns have been carried out during the last decade (Solari et al. 2012; Burlando et al. 2018). To do so, numerous monitoring systems and stations have been installed to record high-dimensional wind field measurements in a continuous way and with a high sampling rate. Given the tremendous amount of data that they generated over the years, it has become necessary to develop automated methods dedicated to detecting different events of interest, like thunderstorms and downbursts, from the analysis of time series.

Several techniques existed before (De Gaetano et al. 2014), but they often required the intervention of expert judgement through a detailed visual inspection of the time series, to compensate for the absence of some statistical descriptors. From a big data perspective, the use of machine learning has therefore appeared as a potential solution to ensure the autonomy of the process. Two main techniques have emerged since then. Chen and Lombardo (2020) used a one-dimensional convolutional neural network classifier trained on segmented records while Arul and Kareem (2021) used a random forest classifier trained on Shapelet Transforms (ST).

Shapelets are highly discriminative sequences which are detected in the time series and somehow represent the respective signature of any wind state. In short, these shapelets have first to be identified in a training dataset and can then be compared to a testing dataset to classify the recordings into two or more groups. By detecting local or global similarities in time series, this method reproduces what humans would naturally do when visualizing the measurements and has the advantage of being easily interpretable. In Arul et al. (2022), this procedure also recognized more thunderstorms than the above-mentioned statistical approaches.

In general, though, the discovery of shapelets is extremely consuming in terms of computational power when performed using brute force, as explained in Section 2. To solve this problem, a new technique based on the randomized sampling of the shapelet candidates is implemented here. It randomly selects a subset of shapelet candidates from a larger pool. The smaller the number of candidates, the faster the evaluation of their discriminative power, but the poorer the stability of the classification. To counter this negative effect, a new type of classifier is also introduced below. These two features are presented in Section 3.1 and Section 3.2, respectively.

An assessment of their performance for the identification of thunderstorms is finally conducted in Section 4, along with a parametric study. First, it focuses on the effect that the sampling method has on the discovery of shapelets, depending on the proportional number of candidates that are considered with respect to the total number lying in the pool. Once that done, two types of forest classifiers are compared. The influence of an additional parameter, the number of trees, is thus evaluated as well. To do so, the wind speed records collected in the Mediterranean port of Livorno (Italy) by the Giovanni Solari WinDyn research group are used.

2 Shapelet-based Classification of Time Series

In the first stage, the algorithm starts by determining which subsequences of a few labeled signals are discriminative for the events of interest. This process is discussed in Section 2.1 and is called the discovery of shapelets. In the second stage, the minimum distance between the best shapelets and each time series is calculated. This operation is presented in Section 2.2, and yields the transforms, which are then gathered inside a single matrix. In the third stage, these shapelet transforms are used to train a machine learning classifier in recognizing the events of interest. This object is introduced in Section 2.3, and then serves to assign classes to unlabeled time series.

2.1 Shapelet Discovery

Listed below are the steps required to conduct the discovery of shapelets. They are described in a concise way, while the underlined concepts are explained in further details in the following paragraphs. For even more information, interested readers can finally refer to Arul et al.'s explanations (2022).

1. Create a learning set with signals of both classes.
2. Separate the time series into shapelet candidates.
3. Run each sub-sequence through each time series.
4. Loop on the complete pool of shapelet candidates.
    a. Loop on the time series of the learning dataset.
        i. Compute the Euclidean distance at every place.
        ii. Get the minimum distance for each time series.
    b. Sort the series by increasing minimal distance.
    c. Get the IG scores for all of the splitting points.
    d. Keep the maximum IG score of each candidate.
5. Use non self-similar shapelets with the highest IG.

Generation of shapelet candidates – The pool of prospective shapelets is formed by the sub-sequences of the time series that are found in the learning set. Overall, $(m - n) + 1$ candidates of length n are generated per signal of length m, with n ranging from 3 to m. This yields a total number of $(2 - 3m + m^2)/2$ candidates per time series. They are then normalized independently of one another to ensure that the discovery process becomes invariant to changes of scale or offset.

Calculation of Euclidean distances – In the present context, the Euclidean distance is computed for all starting indices s, ranging from 1 to $(m - n)$, as the square root of the sum of the squared differences between the data points x_i of the shapelet candidate and the data points y_{s+i} of the time series. This metric evaluates the similarity between any of these pairs of sub-sequences of equal length. If a time series exhibits a pattern similar to the shapelet candidate, the minimal distance is thus low, and vice-versa.

Evaluation of the information gain – Once the minimal distances are sorted in ascending order for a given shapelet candidate, the list is splitted into two parts. This separation results in a reduction of the entropy which can be measured by the information gain (IG). The highest information gains are therefore obtained when each part of the list contains values that are similar to one another, but distinct from the values that are left in the other part. This is indicative of a strong discriminative power.

Rejection of self-similar shapelets – In the end of the process, it is important to discard any shapelet having data points in common with another one whose information gain is higher. Failure to do so can otherwise create a lack of diversity in the shapelets to be used after and therefore compromise the reliability of the subsequent classification. For similar reasons, if the information gain associated to a specific shapelet does not exceed the limit value of 0.05, this candidate is abandoned as well.

2.2 Shapelet Transform

After their discovery, the shapelets are used to convert the time series, also called the instances, into a new matrix of features, a.k.a. the transforms. In this local-shape space, these attributes correspond to the minimal distances between the signals and the shapelets. Along with the class labels, previously attached to each instance, this matrix can finally be treated as a generic input to train any machine learning classifier. It is also depicted in Fig. 1, with a lightning and a sunshine for storm and clear weather events.

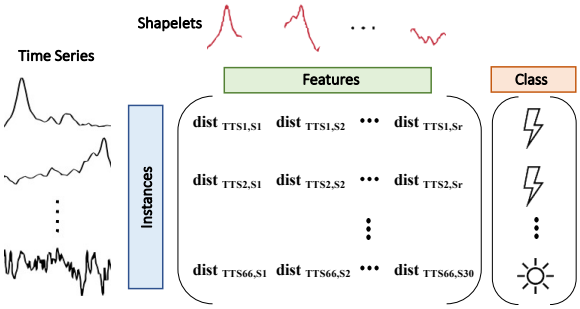

Fig. 1 Shapelet Transform

2.3 Shapelet Classifier

As is typical in machine learning contexts, a random forest classifier is currently implemented in the algorithm. It is composed of 500 trees and the default probability threshold is set to 0.5, meaning that classes are predicted based on a majority vote in a binary setting, as shown in Fig. 2. These predictions vary indeed because each tree receives a different selection of features at start. Their nodes are then allowed to expand until all leaves are pure, without restriction on depth.

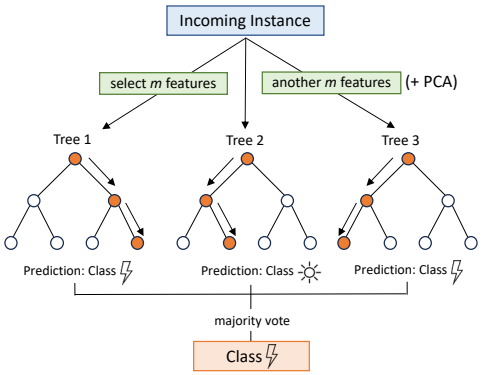

Fig. 2 Random Forest Classifier

3 Improvements of the Existing Algorithm

3.1 Randomized Sampling

Instead of considering all sub-sequences of all time series in the pool of shapelet candidates, a random number of them can be skipped at each iteration (Renard et al. 2015). As schematically represented in Fig. 4, these candidates of similar length and position would have otherwise had the same capacity of discrimination. The probability distribution of the sampling parameter is therefore chosen to create a diverse but smaller subset of candidates, in a given proportion of the whole population. The computational time needed to discover the shapelets is then expected to decrease by the same amount. The question to be answered, though, is to what extent this reduction does not harm the classification process afterwards.

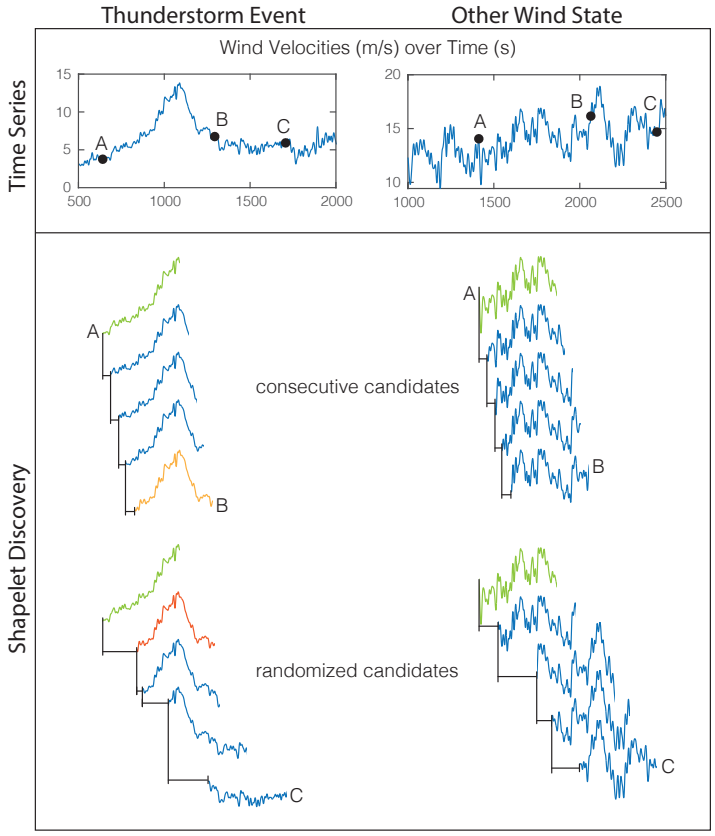

Fig. 3 Randomized Sampling

4.2 Rotation Forest

The rotation forest resembles the random forest but addresses one of its most important flaws, being that it partitions the feature space orthogonally only. In here, indeed, the feature matrix is multiplied by a rotation matrix, which comes from a principal component analysis. As it changes for each tree, diversity is created in the forest, and the robustness of the classification increases.

4.3 Ensemble Rule

Due to the random nature of the sampling, different transforms can now be created based on the same dataset. If several of them are used to train various classifiers, it is necessary to perform an additional task which consists in reunifying their results, as shown in Fig. 4. Through this combination, the resulting algorithm should be more stable and provide a more precise decision.

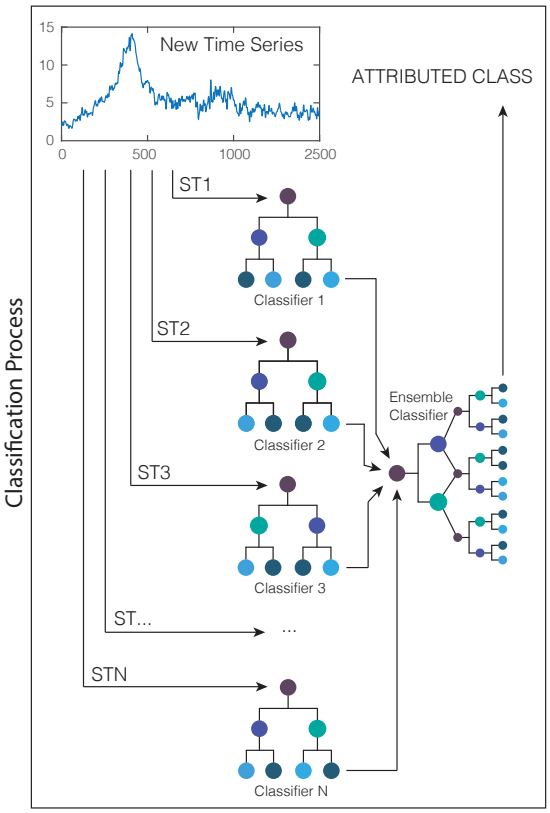

Fig. 4 Ensemble Rule

4 Application to the Detection of Thunderstorms

The techniques presented hereabove are applied to the wind speed signals recorded by ultrasonic anemometers in the Mediterranean port of Livorno, in Italy, during the "Wind and Ports" project (see the characteristics of this complete database in Table 1). The results obtained with different sampling parameters and classifiers will be compared in the sequel, to evaluate the improvements due to the addition of these new features in using less computational power and detecting more true positives.

Table 1 Description of the sensing system at the port of Livorno

Port	Sensor	Height (m)	Period	Type	Rate
Livorno	1-4 5	20 75	2010-2014	3-axial	10 Hz

As explained earlier, this parametric study starts with the creation of a training and a testing dataset. Overall, these two datasets respectively contain 77 and 33 of the 110 catalogued thunderstorms described in Burlando et al. (2018), plus the same number of non-thunderstorm events selected from the rest of the database. The training dataset is then analyzed to find representative shapelets for both the thunderstorm and non-thunderstorm events. To do so, a random sample of candidates is drawn, in varying proportions of the entire population.

Given the non-deterministic nature of the process, the discovery of shapelets is repeated nine times for each proportion (0.1, 0.2, 0.3, 1, 2, 3,10, 20 and 30 percent). In Fig. 5, the mean durations for each of these nine runs are represented by a solid blue line. Meanwhile, their individual durations are illustrated by dashed grey lines. As it was expected, the relationship between the number of candidates, and the duration of is linear. Hence, the fewer the faster.

As regards to the quality of the shapelets, then, IG scores are displayed in Fig. 6. The mean values of the five best results obtained for all runs are shown in red, with plus and minus the standard deviations indicated in orange. Globally, it demonstrates that the scores increase and vary less when more candidates are evaluated. This is really significant between the first and the second trios of proportions, but not that much anymore with respect to the last one.

The 30 shapelets identified at each run are subsequently used to transform both the training and the testing dataset. Each (i,j)-th element of the resulting matrices correspond to the minimal distances between the i-th signal and the j-th shapelet, with $i = 1$ to 154 for training or 66 for testing, and $j = 1$ to 30. Two types of classifiers, the random and the rotation forests, with different numbers of trees are then trained and tested based on these shapelet transforms.

A baseline is first established by using a proportion of 30% for the shapelet candidates and a number of 500 for the decision trees. The expected and the predicted classes attributed to the signals of the testing dataset are displayed in Fig. 7, by means of confusion tables. The ensemble classification rule applied to build them consists in keeping the most frequent results.

A few misclassified signals are then shown in Fig. 8 and Fig. 9, together with the number of times it happened. Looking at these false positives and false negatives helps to understand what went wrong. On the right-hand side of Fig. 8, for instance, the time series exhibits peaks that are usually associated with a thunderstorm. On the left-hand side of Fig. 9, a ramp-up phase, typical of thunderstorms, is observed. Yet, no shapelet covered this specific feature.

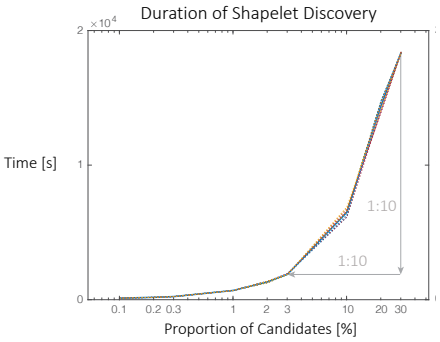

Fig. 5 Computational Times

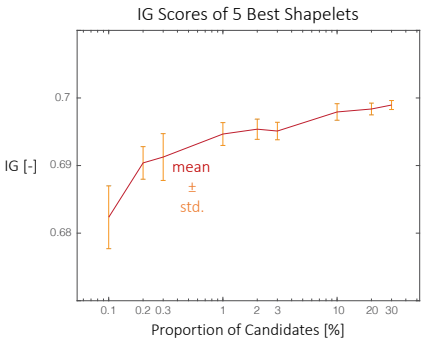

Fig. 6 Information Gains

Last, but not least, the accuracy of the classification process is computed as the sum of all true positives and all true negatives divided by the total number of instances. The mean values and the standard deviations obtained for this performance metric on nine runs are depicted in Fig. 10 and Fig. 11, respectively, depending on the proportion of candidates and the number of trees. These figures globally show that the rotation forest classifier performs better as it generally gives more precise and more stable results with respect to the baseline.

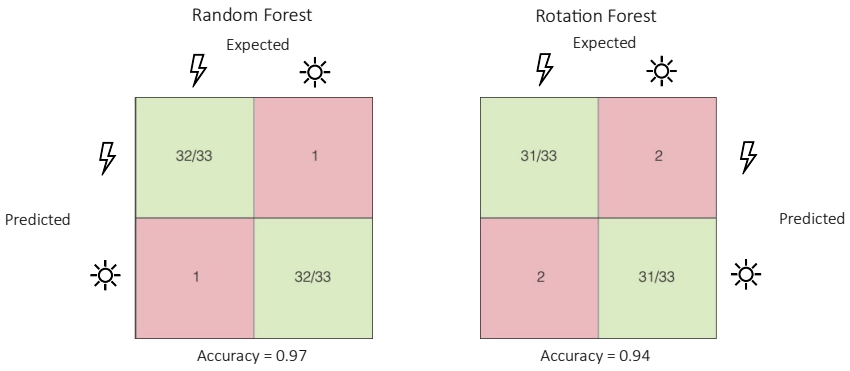

Fig. 7 Confusion Tables

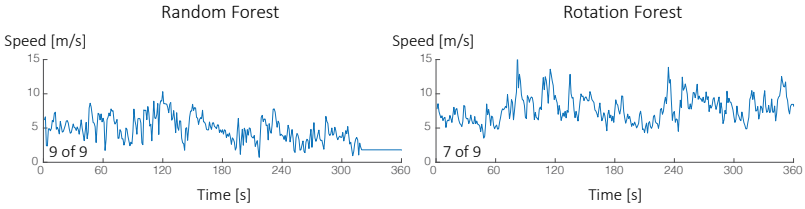

Fig. 8 Examples of False Positives

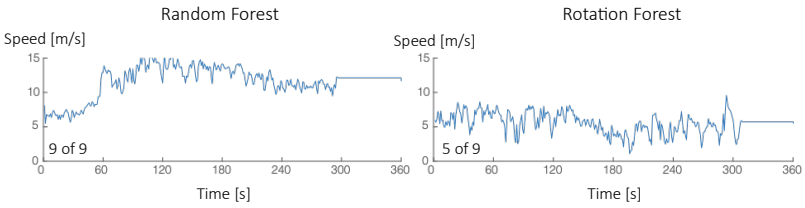

Fig. 9 Examples of False Negatives

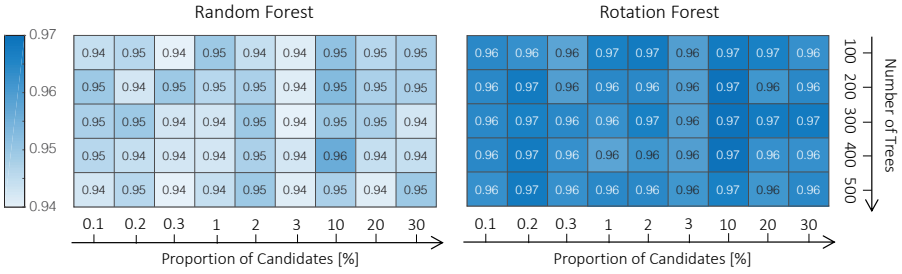

Fig. 10 Accuracy – Mean Values

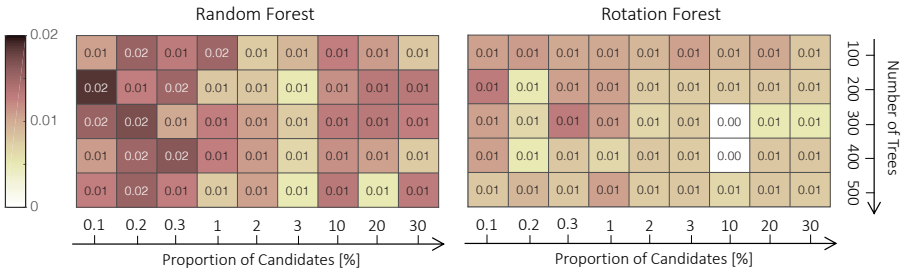

Fig. 11 Accuracy – Standard Deviations

5 Conclusion

Innovative techniques have been introduced in this paper to improve the ability to identify thunderstorms. Their efficiency in reducing the computation burden while increasing the detection accuracy is demonstrated on the wind speed signals that have been recorded by five ultrasonic anemometers in the Mediterranean port of Livorno, throughout the completion of the "Wind and Ports" project. In the meantime, a parametric study has also been conducted to evaluate the influence that the proportion of shapelet candidates, the type of forest classifier, and the number of decision trees have on the classification process. Overall, it revealed the superior performance of the rotation forest classifier in terms of precision and stability.

Acknowledgements This paper is dedicated to the memory of Prof. Giovanni Solari and his dedication to the area of thunderstorm winds under the project THUNDERR. The first author acknowledges the financial support of a postdoctoral grant awarded by the Francqui Foundation, as part of the Belgian American Educational Foundation.

References

Arul, Monica, and Ahsan Kareem. 2021. "Applications of Shapelet Transform to Time Series Classification of Earthquake, Wind and Wave Data." *Engineering Structures* 228: 111564.

Arul, Monica, Ahsan Kareem, Massimiliano Burlando, and Giovanni Solari. 2022. "Machine Learning Based Automated Identification of Thunderstorms from Anemometric Records Using Shapelet Transform." *Journal of Wind Engineering and Industrial Aerodynamics* 220: 104856.

Burlando, Massimiliano, Shi Zhang, and Giovanni Solari. 2018. "Monitoring, Cataloguing, and Weather Scenarios of Thunderstorm Outflows in the Northern Mediterranean." *Natural Hazards and Earth System Sciences* 18 (9): 2309–30.

Chen, Guangzhao, and Franklin T. Lombardo. 2020. "An Automated Classification Method of Thunderstorm and Non-Thunderstorm Wind Data Based on a Convolutional Neural Network." *Journal of Wind Engineering and Industrial Aerodynamics* 207: 104407.

Gaetano, Patrizia De, Maria Pia Repetto, Teresa Repetto, and Giovanni Solari. 2014. "Separation and Classification of Extreme Wind Events from Anemometric Records." *Journal of Wind Engineering and Industrial Aerodynamics* 126: 132–43.

Renard, Xavier, Maria Rifqi, Walid Erray, and Marcin Detyniecki. 2015. "Random-Shapelet: An Algorithm for Fast Shapelet Discovery." *Proceedings of the 2015 IEEE International Conference on Data Science and Advanced Analytics, DSAA 2015*, 1–10.

Solari, Giovanni, Maria Pia Repetto, Massimiliano Burlando, Patrizia De Gaetano, Marina Pizzo, Marco Tizzi, and Mattia Parodi. 2012. "The Wind Forecast for Safety Management of Port Areas." *Journal of Wind Engineering and Industrial Aerodynamics* 104–106: 266–77. https://doi.org/10.1016/j.jweia.2012.03.029.