Universidad del País Vasco / Euskal Herriko Unibertsitatea
eman ta zabal zazu
INFORMATIKA FAKULTATEA
FACULTAD DE INFORMÁTICA

Konputazio Ingeniaritza eta Sistema Adimentsuak Unibertsitate Masterra
Máster Universitario en Ingeniería Computacional y Sistemas Inteligentes

INGENIERÍA COMPUTACIONAL Y SISTEMAS INTELIGENTES

Introduction to Time Series Data Analisis

Time Series Classification

Estudiante: Eneko Perez

Estudiante: Alan García

Curso: 2024-2025

Fecha: January 13, 2025

Description of the time series

We are working with sensor data originally collected by Carlos Guestrin (CMU) and later used in the paper *Online Latent Variable Detection in Sensor Networks*. These data were reformatted for classification tasks by Eamonn Keogh.

The main task is to distinguish between two specific sensors: **q8calibHumid**, which measures humidity, and **q8calibHumTemp**, which measures temperature. A key challenge in this dataset is the presence of missing values ("dropouts"), which were represented as 0 in the original data. This adds complexity, as these values are difficult to detect when the data is normalized.

The dataset consists of 1272 time series, of which 685 belong to class 1 (**q8calibHumid**) and 587 to class 2 (**q8calibHumTemp**). Each time series is characterized by 84 attributes.

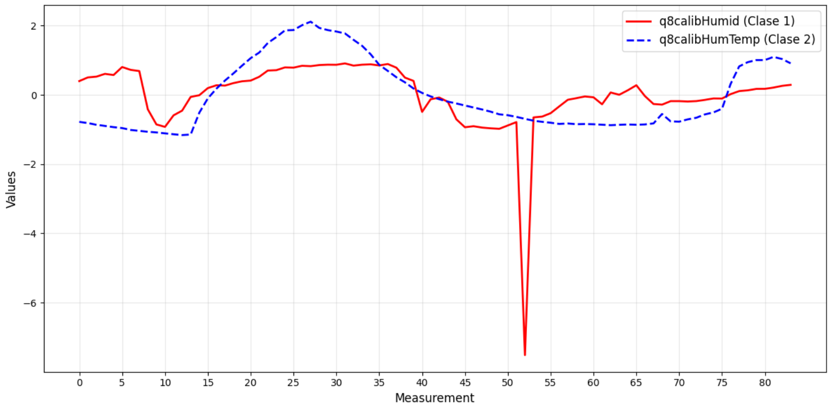

Figure 1: Time series of each class.

Description of the chosen model

1-NN with DTW

Whole series TSC (Time Series Classification) algorithms usually employ a similarity measure between series that quantifies the distance between two series after compensation for localized distortions. The 1-NN method with DTW (Dynamic Time Warping) is a technique used for time series classification. It combines the nearest neighbor approach (1-Nearest Neighbor) with a similarity metric based on the flexible temporal alignment provided by DTW.

Unlike the Euclidean method, which relies on directly comparing each point in one series with its equivalent in another series, DTW seeks the "closest point" between each point of the two series. This allows it to identify similar shapes that may be warped or misaligned. To achieve this, a similarity matrix is generated, comparing the two series ($a = (a_1, a_2, ..., a_m)$ and $b = (b_1, b_2, ..., b_m)$) and calculating the distances between all points of one series and all points of the other, where $M_{i,j} = (a_i - b_j)^2$. The " most optimal path" is then selected from this matrix, representing the shortest distance between each pair of points.

Although it is a simple method, it stands out for its excellent performance in classifying time series. However, one of its drawbacks is its high computational cost, $O(N^2)$, which can make it challenging to apply to very large data sequences.

Shapelets

Shapelets are defined as "subsequences that are in some sense maximally representative of a class" [1]. In other words, a shapelet signifies a subsequence of time series data that is maximally discriminative, effectively capturing patterns that distinguish one class from another.

To train a classifier using shapelets, the original time series data is transformed into the shapelet space. In this space, each time series is represented as a vector of similarity scores, where each score quantifies how closely the series matches a specific shapelet. A logistic regression model is then trained on these transformed vectors. During the training process, both the shapelet coefficients (which define the shapelets) and the logistic regression weights are adjusted. The shapelet coefficients are optimized to ensure that the transformed space becomes linearly separable, enabling the regressor to effectively classify the data.

Additionally, a search heuristic [2] was employed to determine the appropriate number and length of shapelets for the training data, resulting in 4 distinct shapelets with a length of 8.Finally, the time series transformed into the shapelet space have been represented using t-SNE:

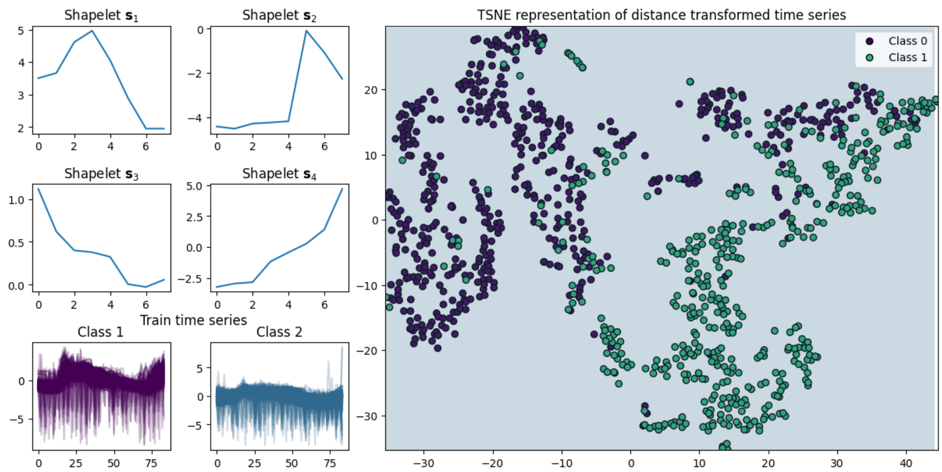

Figure 2: Train time series transformed onto the shapelet space

Results and comparison

The Table 1 presents the performance of two classifiers, the nearest neighbor with Dynamic Time Warping (1-NN with DTW) and Shapelets, evaluated in terms of precision, recall, and F1-score for two classes (0 and 1). The 1-NN with DTW classifier shows superior performance, achieving an average of 0.96 across all metrics. On the other hand, Shapelets demonstrates a more balanced but lower performance, with F1-scores of 0.89 for class 0 and 0.87 for class 1. These differences highlight the effectiveness of 1-NN with DTW for this dataset and these specific classes.

	Precision	**Recall**	**F1-score**	**Support**
1-NN with DTW				
Class 0	0.95	0.97	0.96	137
Class 1	0.97	0.94	0.95	118
(Avg)	**0.96**	**0.96**	**0.96**	255
Shapelets				
Class 0	0.88	0.91	0.89	137
Class 1	0.89	0.86	0.87	118
(Avg)	0.88	0.88	0.88	255

Table 1: Classifier performance.

Conclusions

In this study on time series classification, two main approaches were analyzed: the 1-NN classifier with Dynamic Time Warping (DTW) and the Shapelets-based classifier. The results show that 1-NN with DTW offers significantly superior performance in terms of precision, recall, and F1-score. Although the Shapelets-based classifier achieves slightly lower metrics, its results are still satisfactory and demonstrate competitive performance.

In terms of computational cost, the Shapelets-based classifier has a significant advantage. While 1-NN with DTW requires calculating the distance between the test instance and all training instances, a process that becomes intensive for datasets with many instances and attributes, the Shapelets approach benefits from faster classification once the discriminative Shapelets have been identified during training. This makes Shapelets a more efficient option when working with large datasets, where classification time is a critical factor.

In addition, the 1-NN algorithm with DTW generalizes poorly as it classifies based on the most similar instance, making it vulnerable in small or unrepresentative datasets. In contrast, Shapelet-based methods extract key patterns for each class, enabling more accurate classification, even with variations not present in the training data. This makes them more robust in scenarios with high variability or limited data.

Finally, Shapelets contribute to making time series classification methods more explainable because they provide interpretable features that represent localized, discriminative patterns within the data. These patterns highlight which specific segments of the time series are most influential in distinguishing between different classes, as shown in Figure 3, where a true positive case is aligned with one of the identified Shapelets.

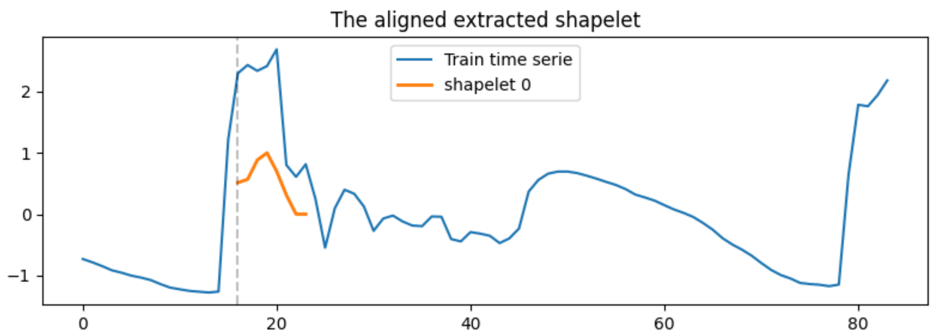

Figure 3: Aligned Shapelet with a true positive time series of class 0

References

[1] Lexiang Ye and Eamonn Keogh. Time series shapelets: a new primitive for data mining. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 947–956, 2009.

[2] 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, 2014.