Parameter Free Piecewise Dynamic Time Warping for time series classification

Vanel Steve Siyou Fotso [1 2] **Engelbert Mephu Nguifo** [1 2] **Philippe Vaslin** [1 2]

Abstract

The Piecewise Aggregate Approximation (PAA) is widely used in time series data mining because it allows to discretize, to reduce the length of time series and is used as a subroutine by algorithms for patterns discovery, indexing, and classification of time series. However, it requires setting one parameter: the number of segments to consider during the discretization. The optimal parameter value is highly data dependent in particular on large time series data. In this work, we propose an approximation to automatically estimate PAA's parameter. The approach - FDTW - is based on simple and intuitive ideas from time series classification. Several properties of our approximation are studied, and an extensive experimental comparison with several time-series classification algorithms demonstrates its efficiency and effectiveness in terms of compression ratio and accuracy. We analyze its impact in compression interpretability and classification accuracy.

1. Introduction

Time series are ubiquitous in sciences as for example in economics, in medicine, in finance or in computer science. Time series databases are often large and several transformations have been introduced in order to represent them in a more compact way. One of these transformations is Piecewise Aggregate Approximation (PAA) (Keogh et al., 2001), which consists in dividing a time series into several segments of fixed length and replacing the data points of each segment with their averages. Due to its simplicity and low computational time, PAA has been widely used as a basic primitive by other temporal data mining algorithms such as (Lin et al., 2003; Sun et al., 2014; Lkhagva et al., 2006), in order to construct symbolic representations of time series; (Camerra et al., 2010; Ulanova et al., 2015), to index time series; (Zhao & Itti, 2016; Keogh & Pazzani, 2000; Kate, 2016), and to classify time series. One major challenge with PAA is the choice of the number of segments to consider. If the number of segments is too small, the resulted representation is compact, but it contains less information. On the other side, if the number of segments is too large, the obtained representation is less compact and more prone to the noise contained in the original time series (Fig. 1). Our idea is that, a number of segments for PAA will be considered as good if it allows to obtain a compact representation of the time series, and also if it preserves the quality of the alignment of time series. So when considering classification task for example, one of the best classification algorithm to use for evaluating the quality of time series alignment is one nearest neighbor (1NN). Indeed, its classification error directly depends on time series alignment, since 1NN has no other parameters (Wang et al., 2013) and the comparison method to consider is Dynamic Time Warping alignment algorithm (DTW)(Sakoe & Chiba, 1978; Zhang et al., 2015) because it considers that some small time distortions and DTW often has better results than the Euclidean Distance (Chen et al., 2015).

In this paper, we propose a parameter free heuristic for aligning piecewise aggregate time series with DTW, which approximates the optimal value of the number of segments to be considered with PAA. In this heuristic, the number of segments is chosen based on the quality of the alignment, which is evaluated by the classification error. The number of segments thus obtained should improves the quality of symbolic representations of time series used for the motifs discovery, their indexation and their classification of these. The accuracy of FDTW has been compared to that of 35 other classification algorithms on 84 datasets.

2. Heuristic search of the number of segments

The idea of our heuristic is the following:

1. We choose N_c candidates distributed in the space of possible values to ensure that we are going to have small, medium and large values as candidates. The candidates values are: n, $n - \left\lfloor \frac{n}{N_c} \right\rfloor$, $n - 2 \times \left\lfloor \frac{n}{N_c} \right\rfloor$, ..., $n - N_c \times \left\lfloor \frac{n}{N_c} \right\rfloor$. For instance, if the length of time series is $n = 12$ and the

*Equal contribution [1]Clermont Auvergne University, LIMOS, BP 10448, F-63000 CLERMONT-FERRAND, FRANCE [2]CNRS, UMR6158, LIMOS, F-63178 AUBIERE, FRANCE. Correspondence to: Vanel Steve Siyou Fotso <vanel_steve.siyou_fotso@uca.fr>.

ICML 2017 Time Series Workshop, Sydney, Australia, 2017.

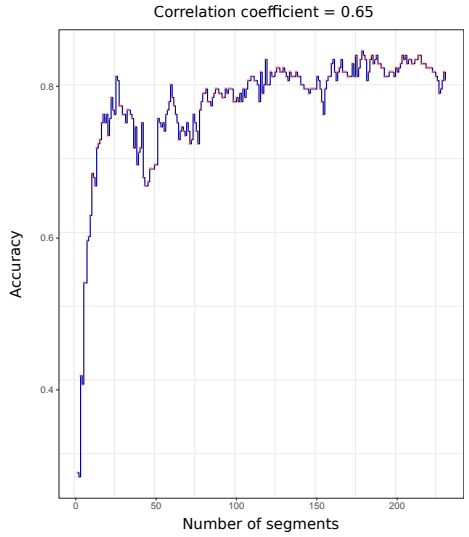

Figure 1. Relation between Accuracy and the number of segment on FISH dataset. The accuracy is computed from the algorithm one nearest neighbor associated with PDTW. When the number of segments considered is very small, there is a loss of information and the accuracy is reduced. However, considering all the points in the time series, we also do not obtain maximum accuracy due to the presence of noise or singularities (Keogh & Pazzani, 2001) in the data.

number of candidates is $N_c = 4$, we are going to select the candidates 12, 9, 6, 3.

$$1, 2, [3], 4, 5, [6], 7, 8, [9], 10, 11, [12]$$

2. We evaluate the classification error with $1NNPDTW$ for each chosen candidate, and we select the candidate that has the minimal classification error: it is the best candidate. In our example, we may suppose that we get the minimal value with the candidate 6 : it is thus the best candidate at this step.

$$1, 2, 3, 4, 5, [6], 7, 8, 9, 10, 11, 12$$

3. We respectively look between the predecessor (i.e., 3 here) and successor (i.e., 9 here) of the best candidate for a number of segments with a lower classification error : this number of segments corresponds to a local minimum. In our example, we are going to test values 4, 5, 7 and 8 to see if there is a local minimum.

4. We restart at step one, while choosing different candidates during each iteration to ensure that we return a good local minimum. We fix the number of iterations to $k \leq \lfloor log(n) \rfloor$. At each iteration the first candidate is $n - (number\_of\_iteration - 1)$.

In short, in the worst case, we test the first N_c candidates to find the best one. Then, we test $\frac{2n}{N_c}$ other candidates to find the local minimum. We finally perform $nb(N_c) = N_c + \frac{2n}{N_c}$ tests. The number of tests to be performed is a function of the number of candidates. Hence, how many candidates should we consider to reduce the number of tests? The first derivative of nb function vanishes when $N_c = \sqrt{2n}$ and its second derivative is positive; so the minimal number of tests is obtained when the number of candidates is : $N_c = \sqrt{2n}$.

Lemma 1. : For a given a dataset d_i, $FDTW(d_i) \leq 1NNDTW(d_i)$. The quality of the alignment of our heuristic is better than that of DTW.

Proof : $1NNDTW(d_i) = 1NNPDTW(d_i, n)$. Then, $1NNDTW(d_i)$ is one of the candidates considered by the heurisitic $FDTW$. Since $FDTW$ returns the minimal classification error from all candidates, the classification error of $1NNDTW$ is always greater than or equal to $FDTW$.

A heuristic does not always give the optimal value. To ensure that it gives a result not far from the optimal value, one approach is to guarantee that the result of the heuristic always lies in an interval with respect to the optimal value (Ibarra & Kim, 1975).

In our case, we are looking for the number of segments that allows a good alignment of time series. The alignment is good when the classification error with 1NN is minimal or when the accuracy is maximal.

Let d_i be a dataset:

$acc_{max(d_i)} = 1 - \min_{1 \leq \alpha \leq n} \{1NNPDTW(di, \alpha)\}$ is the maximal accuracy for the dataset d_i,

$acc_{DTW} = 1 - 1NNDTW(d_i)$ is the accuracy with d_i and 1NNDTW and

$acc_{FDTW} = 1 - FDTW(d_i)$ is the accuracy of our heuristic.

To ensure the quality of our heuristic FDTW, wee hypothesized tat $1NNDTW$ is better than Zero Rule classifier. Zero Rule classifier is a simple classifier that predicts the majority class of test data (if nominal) or average value (if numeric). Zero Rule is often used as baseline classifier (Cuřín et al., 2007). The minimal value of the accuracy of Zero Rule is $\frac{1}{c}$ where c is the number of classes of the dataset.

Proposition 1. : For a given dataset d_i that has c_i classes, $c_i \in \mathbb{N}^*$,

$if\ acc_{DTW} \geq \frac{1}{c_i}\ then\ \frac{1}{c_i} \times acc_{max} \leq acc_{FDTW} \leq$

acc_{max}

Proposition 1 shows that when 1NN associated with DTW has a better accuracy than the baseline classifier Zero Rule, the FDTW heuristic is an approximation.

: By definition, $acc_{FDTW} \leq acc_{max}$ We look for $\beta \in \mathbb{N}$ such that

$$\frac{1}{\beta} \times acc_{max} \leq acc_{FDTW} \quad (1)$$

$$\frac{1}{\beta} \times acc_{max} \leq acc_{FDTW} \Leftrightarrow \frac{acc_{max}}{acc_{FDTW}} \leq \beta \quad (2)$$

$$However, \quad \frac{acc_{max}}{acc_{FDTW}} \leq \frac{1}{acc_{FDTW}} \quad (3)$$

$$because \quad acc_{max} \leq 1 \quad (4)$$

$$And, \quad \frac{1}{acc_{FDTW}} \leq \frac{1}{acc_{DTW}} \quad (5)$$

$$because \quad acc_{DTW} \leq acc_{FDTW} \quad (6)$$

$$So, \quad \frac{1}{acc_{DTW}} \leq c_i \quad (7)$$

$$because \quad \frac{1}{c_i} \leq acc_{DTW} \quad by \quad hypothesis \quad (8)$$

3. Experiments and discussion

3.1. Datasets

The performance of FDTW has been evaluated on 84 datasets of the UCR time series datamining archive (Chen et al., 2015), which provides a large collection of datasets that cover various domains. Each dataset is divided into a training set and a testing set.

3.2. Compression

When it is used with a suitable segments number determined with FDTW, PAA allows compression of the time series of the **Coffee** without loss of information. Although they are more compact, the obtained time series capture the main variations of the original time series (Fig. 2).

3.3. Classification

To evaluate the quality of FDTW, we compared its classification errors with that of 35 other classification algorithms (Bagnall et al., 2016b) of the literature on 84 datasets of UCR archive. The classification error was calculated based on the holdout model evaluation. FDTW used the training set to find the number of segments N using 3-fold cross validation. If two numbers of segments N_1 and N_2 are associated with the same classification error, we retain the largest. The performances of the algorithms are compared using the Nemenyi test that compares all the algorithms pairwise and provides an intuitive way to visualize the results (Fig. 3). The Nemenyi test allows to rank the classification algorithms according to their average accuracy on 84 datasets.

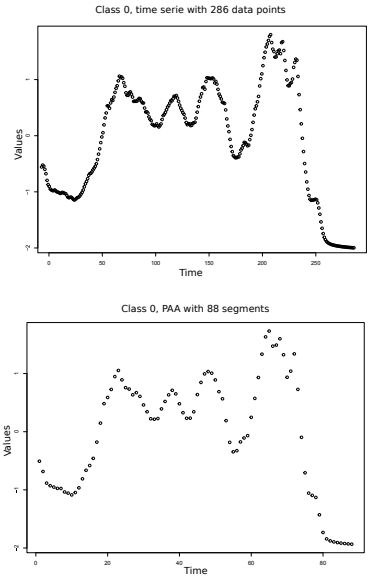

Figure 2. Coffee dataset time series compression with PAA: original time series (top) versus PAA represetion using 88 segments (down). The number of segments is found by FDTW and allow to reduce the length of the time series while retaining the information that it contains.

The value of the segment number N found on the trainingset may in some cases not be appropriate for the testingset. We speak of an error of generalization which is due to the representativeness of the trainingset. Thus, FDTW obtains good results on the simulated data sets 3rd / 37 algorithms in terms of average accuracy (Fig. 3) because the data of the trainingset and the testingset are generated by the same models.

However, to evaluate the significance of the difference be-

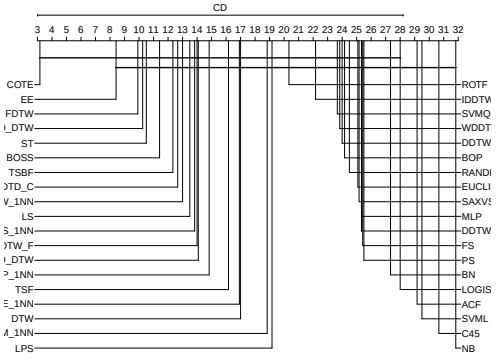

Figure 3. Critical difference diagram for FDTW and 36 other classification algorithms on 6 simulated datasets.

tween the classification algorithms on 84 datasets, we use the Wilcoxon signed rank test with continuity correction which has more statistical power.The results of these experiments are summarized below.

The experiments show that despite data compression :

- FDTW have better performance than Naive Bayes (NB), C45, logistic regression (Logistic), BN;
- FDTW has similar performance to that of 26 other algorithms in the literature, namely : SVMQ, RANDF, ROTF, MLP, EUCLIDEAN_1_NN, DDTW_R1_1NN, DDTW_RN_1NN, ERP_1NN, LCSS_1NN, MSM_1NN, TWE_1NN, WDDTW_1NN, WDTW_1NN, DD_DTW, DTD_C, LS, BOP, SAXVSM, TSF, TSBF, LPS, PS, CID_DTW, SVML, FS, ACF;
- DTW_F, Shapelet Transform (ST), BOSS, Elastic Ensemble (EE) and COTE perform better overall than FDTW.

This demonstrates its competitiveness, details are available here (Siyou Fotso et al., 2016). Moreover, FDTW outperforms the best result reported in the literature on UWaveGestureLibraryAll dataset (Fig. 4). The challenge with the UWaveGestureLibraryAll dataset is to recognize the gesture made by a user from measurements made by accelerometers. As reported here (Bagnall et al., 2016a) the best accuracy obtained on this dataset is 83.44% with TSBF algorithm. FDTW out performs this result and allow to obtain **91.87%** of accuracy.

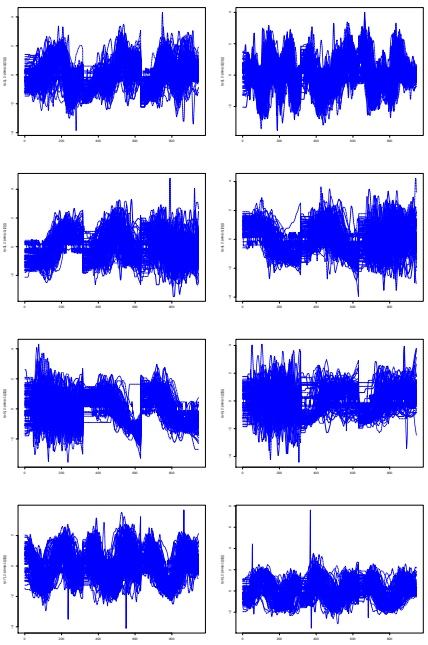

Figure 4. Eight types of time series corresponding to the vocabulary of 8 gestures.

4. Conclusion

FDTW allows to reduce the storage space and the processing time of time series classification without decreasing the alignment quality. As a perspective, we plan to use piecewise aggregate to find the number of segments to be considered for symbolic representations of time series like SAX, ESAX, SAX-TD.

References

Bagnall, Anthony, Keogh, Eamonn, Lines, Jason, Bostrom, Aaron, and Large, James. Time Series Classification Website. http://timeseriesclassification.com, October 2016a.

Bagnall, Anthony, Lines, Jason, Bostrom, Aaron, Large, James, and Keogh, Eamonn. The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, pp. 1–55, 2016b.

Camerra, Alessandro, Palpanas, Themis, Shieh, Jin, and Keogh, Eamonn. isax 2.0: Indexing and mining one billion time series. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pp. 58–67. IEEE, 2010.

Chen, Yanping, Keogh, Eamonn, Hu, Bing, Begum, Nurjahan, Bagnall, Anthony, Mueen, Abdullah, and Batista, Gustavo. The ucr time series classification archive. http://www.cs.ucr.edu/~eamonn/time_series_data/, July 2015.

Cuřín, Jan, Fleury, Pascal, Kleindienst, Jan, and Kessl, Robert. Meeting state recognition from visual and aural labels. In *International Workshop on Machine Learning for Multimodal Interaction*, pp. 24–25. Springer, 2007.

Ibarra, Oscar H and Kim, Chul E. Fast approximation algorithms for the knapsack and sum of subset problems. *Journal of the ACM (JACM)*, 22(4):463–468, 1975.

Kate, Rohit J. Using dynamic time warping distances as features for improved time series classification. *Data Mining and Knowledge Discovery*, 30(2): 283–312, 2016. ISSN 1573-756X. doi: 10.1007/s10618-015-0418-x.

Keogh, E J and Pazzani, M J. Derivative dynamic time warping. *1st SIAM International Conference on Data Mining*, pp. 1–11, 2001.

Keogh, Eamonn, Chakrabarti, Kaushik, Pazzani, Michael, and Mehrotra, Sharad. Dimensionality reduction for fast similarity search in large time series databases. *Knowledge and information Systems*, 3(3):263–286, 2001.

Keogh, Eamonn J and Pazzani, Michael J. Scaling up dynamic time warping for datamining applications. In *sixth ACM SIGKDD*, pp. 285–289. ACM, 2000.

Lin, Jessica, Keogh, Eamonn, Lonardi, Stefano, and Chiu, Bill. A symbolic representation of time series, with implications for streaming algorithms. In *8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, pp. 2–11. ACM, 2003.

Lkhagva, Battuguldur, Suzuki, Yu, and Kawagoe, Kyoji. Extended sax: Extension of symbolic aggregate approximation for financial time series data representation. *DEWS2006 4A-i8*, 7, 2006.

Sakoe, Hiroaki and Chiba, Seibi. Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing*, 26(1):43–49, 1978.

Siyou Fotso, Vanel Steve, Mephu Nguifo, Engelbert, and Vaslin, Philippe. Comparison of classification algorithms to fdtw. `http://fc.isima.fr/~siyou/fdtw`, October 2016.

Sun, Youqiang, Li, Jiuyong, Liu, Jixue, Sun, Bingyu, and Chow, Christopher. An improvement of symbolic aggregate approximation distance measure for time series. *Neurocomputing*, 138:189–198, 2014.

Ulanova, Liudmila, Begum, Nurjahan, and Keogh, Eamonn. Scalable clustering of time series with u-shapelets. In *2015 SIAM International Conference on Data Mining*, pp. 900–908. SIAM, 2015.

Wang, Xiaoyue, Mueen, Abdullah, Ding, Hui, Trajcevski, Goce, Scheuermann, Peter, and Keogh, Eamonn. Experimental comparison of representation methods and distance measures for time series data. *Data Mining and Knowledge Discovery*, 26(2):275–309, 2013.

Zhang, Zheng, Tang, Ping, and Duan, Rubing. Dynamic time warping under pointwise shape context. *Information Sciences*, 315:88–101, 2015.

Zhao, Jiaping and Itti, Laurent. shapedtw: shape dynamic time warping. *arXiv preprint arXiv:1606.01601*, 2016.