A Dispersion-based PAA Representation for Time Series

Leonidas Karamitopoulos
Dept. of Applied Informatics,
University of Macedonia, Thessaloniki, Greece
Tel: +302310-891844,
Email: lkaramit@uom.gr

Georgios Evangelidis
Dept. of Applied Informatics,
University of Macedonia, Thessaloniki, Greece
Tel: +302310-891844
Email: gevan@uom.gr

Abstract

Time series data generation has exploded in almost every domain such as in business, industry, or medicine. The demand for analyzing efficiently the huge amount of this information necessitates the application of a representation on the purpose of reducing the intrinsically high dimensionality of time series. In this paper we introduce D-PAA, a new representation that can be considered as a variation of Piecewise Aggregate Approximation (PAA). D-PAA segments a time series into a series of equal length sections and the corresponding mean and standard deviation are recorded for each one of them. The difference with PAA is that D-PAA takes into consideration not only the central tendency but also the dispersion present in each section. We evaluate our representation by applying 1-NN classification on 20 widely utilized datasets in the literature. Experimental results indicate that the proposed representation performs better than other commonly applied representations in the majority of the datasets.

1. Introduction

Several procedures generate huge amounts of data in the form of time series in almost every domain, such as in business, industry, medicine or science. In addition to that, image or video data can be also considered as time series. The increasing need of analyzing efficiently the huge amount of this information led to the development of Data Mining techniques adjusted in a way that takes into consideration the temporal nature of data.

The temporal aspect of data arises some special issues to be considered and/or imposes some restrictions in the corresponding applications. First, it is necessary to define a similarity measure between two time series and this issue is very important, since it involves a degree of subjectivity that might affect the final result. Second, it is necessary to apply a representation scheme on the time series data. Since the amount of data may range from a few megabytes to terabytes, an appropriate representation is necessary in order to reduce the intrinsically high dimensionality of time series.

There is a large number of representations proposed in the literature that aim at reducing the dimensionality of a time series dataset, while retaining as much information as possible. As it is expected, there is no representation that can be considered as the most appropriate for all cases. The application under consideration, the data mining method and the specific characteristics of data play an important role in the selection of a representation scheme. In this paper, we introduce a new representation, named Dispersion-based PAA (D-PAA), which can be considered as a variation of Piecewise Aggregate Approximation (PAA) [1]. A time series is segmented into a series of equal length sections and the corresponding mean and standard deviation are recorded for each one of them. The difference with PAA is that D-PAA takes into consideration not only the central tendency but also the dispersion present in each section. This additional information could improve the performance of dimensionality reduction in the context of time series similarity search. Extensive experiments on 20 widely utilized datasets are conducted in order to evaluate D-PAA and compare it to PAA, as well as with other commonly applied representations in Time Series Data Mining applications.

In Section 2 we provide a brief review on Time Series Data Mining and related work with respect to representations. In Section 3 we introduce the novel representation along with a distance measure. The experimental methodology is described in Section 4, whereas the corresponding results are presented and discussed in Section 5. Finally, conclusions and future work is provided in Section 6.

2. Background

2.1 Time Series Data Mining

Time Series Data Mining (TSDM) is comprised by Data Mining methods adjusted in a way that they take into consideration the temporal nature of data. According to the research in this field, the main tasks of TSDM methods are: query by content, clustering, classification, novelty detection, motif discovery and rule discovery [2]. At the core of all these tasks lies the notion of similarity. Two time series can be considered similar when they exhibit similar shape or pattern. The presence of noise demands allowing imprecise matches among sequences. Consequently, it is necessary to define an appropriate similarity measure, since the notion of similarity involves a degree of subjectivity. A thorough discussion on similarity measures is provided in [3]. Another important issue that arises from the temporal nature of data is the intrinsic high dimensionality, which affects substantially the efficiency of data mining techniques. High dimensionality affects the calculation speed of similarity measure among series and, moreover, prohibits the construction of an efficient indexing structure. The idea is to reduce the dimensionality of the original data by representing it in a lower dimension, analyze it in this dimension and, finally, tune the results in order to obtain the same solution with the one that would have been derived, if the original data had been used in the analysis. A representation scheme is applied in order to reduce the dimensionality. However, this representation should guarantee that there will not be any false dismissals. This property, known as Lower Bounding Lemma, can be described as follows. Suppose that t_1 and t_2 are two time series that need to be investigated for similarity and R denotes a representation scheme. Given a distance function D between two time series, R should satisfy the following property in order to guarantee no false dismissals:

$$D(R(t_1) , \mathrm{R}(\mathrm{t}_2)) \leq D(t_1 , \mathrm{t}_2)$$

This property states that the distance measure in the k-dimension feature space should lower bound the corresponding distance measure in the original space. Besides the Lower Bounding Lemma, it is important for the representation to lower bound the true distance as tightly as possible. That is, the distance measure in the k-dimension feature space should be as close as possible to the corresponding distance measure in the original space in order to reduce the number of false hits and consequently the post-processing time.

Apparently, similarity measures and representation schemes are interrelated to each other and play an important role in efficiently applying any time series data mining task.

2.2 Related Work

There is a wealth of representation schemes proposed in the literature, the detailed description of which is beyond the scope of this paper. A hierarchy of various time series representations is presented in a tree diagram in [4]. We will briefly refer to the most popular representations within the Data Mining context.

One of the first representations proposed was the Discrete Fourier Transform (DFT) [5]. DFT transforms a time series from the time domain into the frequency domain by expressing the time series as a linear combination of trigonometric functions. Another suggested representation is Discrete Wavelet Transform (DWT), which transforms a time series into the time/frequency domain by decomposing it into a series of wavelet basis functions [6]. Singular Value Decomposition (SVD) [7] performs a global transformation by rotating the axes of the entire dataset and represents the time series as a linear combination of the most important “principal components”. Another approach that is called Piecewise Linear Approximation (PLA), approximates a time series by a sequence of linear segments [8].

The most closely related representation to the one proposed in this paper is Piecewise Aggregate Approximation (PAA). PAA is a dimensionality reduction technique that was proposed independently by Keogh et al. [1] and, Yi and Faloutsos [9]. This technique segments a time series of length n into k consecutive sections of equal-width and calculates the corresponding mean for each one. The series of these means is the new representation of the original data. PAA is simple, fast to calculate and it has been shown empirically that it is as efficient as other approaches. Moreover, it can handle time series of different lengths. Another similar approach [10] uses two additional values, besides the mean value, for representing a segment of a time series, namely, the corresponding minimum and maximum values. These values are mapped to an alphabet to produce a symbolic representation. This approach is an extension to SAX representation [4] and aims at improving the representation of financial time series data.

A different approach in representation that utilizes the dispersion of a time series is introduced by Nanopoulos et al. [11]. The authors proposed the extraction of global statistical features from a time series, which, in conjunction with a multi-layer perceptron neural network, can be utilized for time series classification. In their work, a time series $X = x_1, x_2, \cdots, x_n$ of length n is represented by a feature vector of length 8. The first 4 features are the mean value, the standard deviation, the skewness and the kurtosis. The next 4 features are extracted by calculating the same statistical measures on the transformed time series $X' = x'_1, x'_2, \cdots, x'_{n-D}$, where:

$$x'_t = (x_{t+D} - x_t), \quad 1 \le t \le n - D \qquad (1)$$

and D is a user-defined parameter. Hereafter, this feature-based representation will be denoted as FB.

3. The D-PAA Representation

According to D-PAA, a time series of length n is segmented into k consecutive sections of equal-width and the corresponding mean and standard deviation are recorded for each one of them. The difference with PAA is that D-PAA takes into consideration not only the central tendency but also the dispersion present in each section.

More formally, a time series $X = x_1, x_2, \cdots, x_n$ can be represented by a series

$$\tilde{X} = (\bar{x}_1, s_{x,1}), (\bar{x}_2, s_{x,2}), \cdots, (\bar{x}_k, s_{x,k})$$

where:

$$\bar{x}_i = \sum_{j=(i-1)\frac{n}{k}+1}^{i\frac{n}{k}} x_j / (\frac{n}{k}) \qquad (2)$$

$$s_{x,i} = \sqrt{ \sum_{j=(i-1)\frac{n}{k}+1}^{i\frac{n}{k}} (x_j - \bar{x}_i)^2 / (\frac{n}{k} - 1) } \qquad (3)$$

These statistical measures may also be referred as features. The quantity n/k in (2) and (3) expresses the number of points that each section consists of and it is assumed that it is an integer. Otherwise, an appropriate number of zeros can be added at the end of the original series prior to representation. As it is stated in [9], this modification does not affect query results. The dimensionality of the transformed time series is equal to 2·k, since for each of the k segments two values are recorded.

In order to determine the distance between two time series in the feature space, a distance measure is needed. In this work, a weighted Euclidean distance is proposed, as follows. Suppose that there are two time series $X = x_1, x_2, \cdots, x_n$ and $Y = y_1, y_2, \cdots, y_n$ along with their representations

$$\tilde{X} = (\bar{x}_1, s_{x,1}), (\bar{x}_2, s_{x,2}), \cdots, (\bar{x}_k, s_{x,k})$$
$$\tilde{Y} = (\bar{y}_1, s_{y,1}), (\bar{y}_2, s_{y,2}), \cdots, (\bar{y}_k, s_{y,k}).$$

The distance between $\tilde{X}$ and $\tilde{Y}$ is defined in the following equation.

$$D_{STD}(\tilde{X}, \tilde{Y}) = \sqrt{ w \cdot \sum_{i=1}^{k} (\bar{x}_i - \bar{y}_i)^2 + (1 - w) \cdot \sum_{i=1}^{k} (s_{x,i} - s_{y,i})^2 }, \quad 0 \le w \le 1 \qquad (4)$$

where w is a user-specified parameter that assigns different weights to the differences of means and standard deviations. This parameter enables the user to assign relative importance to mean values and standard deviations, according to the specific application and/or data type under consideration. In the extreme case of w = 1, the D-PAA representation is equivalent to PAA, since only the mean values are taken into account. As w decreases, more weight is assigned to standard deviations. In the other extreme of w = 0, the D-PAA representation ignores the mean values. Note that the case of w = 0.5 does not imply that each feature has equal absolute influence on the distance function value [12]. One way of giving all features equal influence in characterizing overall dissimilarity between time series is to normalize the values of each feature prior to calculating the weighted distance and set w equal to 0.5. However, this normalization incurs an extra computation cost, since it must be realized each time a new query arrives.

It can be proved that the proposed distance lower-bounds the Euclidean distance between X and Y, that is,

$$D_{STD}(\tilde{X}, \tilde{Y}) \le D(X, Y)$$

In this paper, the proof is omitted due to limited space.

By contrasting PAA with D-PAA, which is the most similar representation, an important observation that arises is that the first utilizes twice as many segments than the latter. D-PAA aims at describing each segment more thoroughly than PAA does, at the expense of describing fewer segments. Moreover, this fact implies that D-PAA representation can be of length that is a multiple of 2, since two measures are calculated for each segment, whereas there is no such restriction for PAA.

On the other hand, the FB representation, which also takes into consideration the dispersion in data, has a constant length 8, since there are four statistical measures calculated for the values of the original time series and the transformed one.

4. Framework of Experimentation

4.1 Datasets

The experiments were conducted on 20 real world and synthetic datasets, which are available upon request from [13]. Most of them have been used extensively in the TSDM literature, for the purpose of testing the performance of novel representation schemes and similarity measures with respect to specific Data Mining tasks and/or indexing. Since they have been utilized extensively as benchmark datasets for testing classification algorithms, they are separated in training and testing sets. All series are labeled according to the class they belong to. The number of classes ranges from 2 to 37, whereas the length of time series ranges from 60 to 637. All time series are normalized, that is, the mean is equal to zero and the standard deviation is equal to one.

4.2 Method & Rival Representations

In order to evaluate the performance of the proposed representation, we perform one-nearest neighbor classification and evaluate it by means of the classification error rate.

First, we compare the performance of D-PAA and Piecewise Aggregate Approximation (PAA), since these two representations are closely related to each other. We also provide results of the feature-based approach (FB), which utilizes the standard deviation, as D-PAA does. In addition to these, we present results for two widely utilized representations, namely the Discrete Fourier Transform (DFT) and Singular Value Decomposition (SVD) along with Euclidean Distance (ED) and Dynamic Time Warping (DTW), which are applied on raw data.

The above representations require determining specific parameters. First, the FB representation involves the parameter D (1) that defines the transformed series for which the statistical measures are computed (besides the original series). We conducted extensive experiments on the datasets described in the previous section for varying values of D. In Figure 1, a summarization of results is presented as the average error rate of 1-NN classification of all datasets across the varying values of D. It is clear that the best average performance is achieved when D equals to 1, whereas for larger values there is a relatively stable performance.

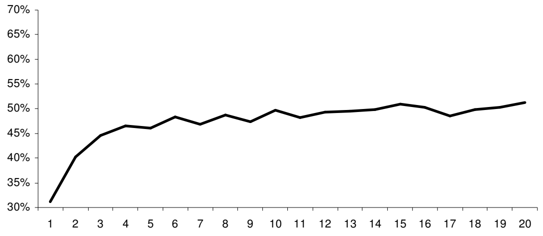

Figure 1. Average error rate for varying values of D

Regarding D-PAA, there are two parameters that need to be determined. The first one is the weight w in the proposed distance function (4). The values tested in the experiments range from 0 to 1 in increments of 0.1.

The second parameter is the number of sections (k) into which the original time series is segmented. We experimented with values of k from 2 to 8. Since the resulting dimensionalities range from 4 to 16 in increments of 2, we appropriately set the required parameters of the other rival representations (PAA, DFT and SVD) so that we achieve the same dimensionalities.

Finally, DTW is realized as originally proposed by Berndt & Clifford [14].

5. Results

Table 1 presents the minimum error rates and the corresponding dimensionalities (dim) of 1-NN classification, when PAA and D-PAA representations were applied on data.

The main observation is that the D-PAA representation produces better results than PAA in the majority of datasets. In 15 out of 20 datasets, the improvement in error rate ranges from approximately 0.1 to 20 units. In 3 out of 20 datasets PAA performs exactly the same with D-PAA, whereas in 2 datasets, PAA has an error rate that is lower by 0.4 and 2.7 units.

Regarding the dimensionalities, in which D-PAA and PAA achieve the minimum error rates, there is not any dominant pattern across datasets (Table 1). However, it is worth noting that there are 6 out of 20 datasets, where D-PAA produced better results than PAA with fewer dimensions.

Regarding the weight (w) assigned to means and standard deviations (5), the first observation is that different values of w may result into equal classification error rates (Table 1). The second observation is that, in 10 out of 20 datasets, the minimum error rate is achieved, when more weight is assigned in standard deviations (w < 0.5). Only in 4 datasets the corresponding rate is better, when more weight is assigned to means (w > 0.5), whereas in the remaining 6 datasets, the minimum error rate is observed for values of w that lie in both areas.

Table 1. 1-NN classification minimum error rates (%)

Dataset	PAA	dim	D-PAA	dim	w
Adiac	51.7	14	40.2	14	0.1
50words	33.6	12	31.4	16	0.5
CBF	3.1	10	0.1	6	0.1, 0.2
ECG200	11.0	12	8.0	8	0.2, 0.3
FaceAll	32.2	16	25.0	14	0.7
FaceFour	18.2	16	13.6	12	0.1, 0.2, 0.3, 0.9
Fish	28.0	14	25.7	12	0.3
GunPoint	8.0	6	6.0	14	0.1, 0.2
Lighting2	18.0	8	11.5	6	0.3
Lighting7	27.4	16	27.4	14	0.2 – 0.5
OSULeaf	46.7	16	47.1	14	0.1
SwedishLeaf	20.5	14	12.5	14	0.2
Control	1.0	12	3.7	16	0.2, 0.3, 0.5, 0.6
Trace	24.0	4	4.0	12	0.1, 0.3
TwoPatterns	6.5	16	5.4	16	0.7
Wafer	0.5	8	0.4	14	0.7
Yoga	17.3	12	16.7	16	0.2
Beef	26.7	6	26.7	12	0.7 – 1.0
Coffee	3.6	14	0.0	16	0.1 – 0.7
OliveOil	13.3	6	13.3	10	0.0 - 0.2 & 1.0

* Gray areas indicate the representation, which provides the lowest error rate.

Table 2 presents the minimum error rates recorded in 1-NN classification, when several representations were applied. The main observation is that D-PAA and DTW produce the minimum error rate in the majority of the datasets.

When D-PAA is compared to other representations, it improves the error rate in 11 datasets, performes exactly the same in 3 datasets and worse in 6 datasets. When it is compared with the next "best" representation, the improvement ranges from 0.1 to 7 units, whereas, the deterioration ranges from 0.1 to 28.1 units. On the other hand, all other representations perform better only in much fewer cases. The SVD representation performs better than all the other representations in three datasets, whereas FB provides better results only in two datasets.

Compared to Euclidean distance (ED), D-PAA performs considerably better in 19 out of 20 datasets. On the other hand, the performances of D-PAA and DTW seem to be comparable to each other. D-PAA improves error rate in 9 datasets, performs exactly the same in 2 datasets and worse in 9 datasets. The improvement ranges from 0.2 to 15 units, whereas, the deterioration ranges from 0.3 to 9 units.

Table 2. 1-NN classification minimum error rates (%)

Dataset	ED	DTW	SVD	DFT	FB	PAA	D-PAA
Adiac	38.9	39.6	39.1	43.7	46.6	51.7	40.2
50words	36.9	31.0	34.3	33.9	69.2	33.6	31.4
CBF	14.8	0.3	4.9	3.3	16.6	3.1	0.1
ECG200	12.0	23.0	12.0	11.0	23.0	11.0	8.0
FaceAll	28.6	19.2	31.9	27.3	40.4	32.2	25.0
FaceFour	21.6	17.0	20.5	17.1	37.5	18.2	13.6
Fish	21.7	16.7	21.1	22.9	39.4	28.0	25.7
GunPoint	8.7	9.3	8.0	8.7	22.0	8.0	6.0
Lighting2	24.6	13.1	18.0	19.7	34.4	18.0	11.5
Lighting7	42.5	27.4	32.9	28.8	57.5	27.4	27.4
OSULeaf	47.9	40.9	46.3	48.4	19.0	46.7	47.1
SwedishLeaf	21.1	21.0	17.4	17.8	18.7	20.5	12.5
Control	12.0	0.7	1.0	1.0	13.3	1.0	3.7
Trace	24.0	0.0	25.0	27.0	11.0	24.0	4.0
TwoPatterns	9.3	0.0	5.7	5.5	29.1	6.5	5.4
Wafer	0.5	2.0	0.5	0.4	0.3	0.5	0.4
Yoga	17.0	16.4	17.2	17.4	21.5	17.3	16.7
Beef	33.3	36.7	33.3	33.3	26.7	26.7	26.7
Coffee	0.0	0.0	0.0	0.0	3.6	3.6	0.0
OliveOil	13.3	16.7	10.0	16.7	33.3	13.3	13.3

* Gray areas indicate the representation, which provides the lowest error rate.

6. Conclusion

In this paper, we introduce a novel time series representation, named D-PAA, along with a distance measure that lower bounds the Euclidean distance.

The first conclusion is that D-PAA, when compared to the most similar representation PAA, performs considerably better in the majority of the datasets without necessarily sacrificing the required dimensionality. Moreover, experiments indicate that local dispersion present in a time series possesses a further discriminating power in conjunction with the corresponding central tendency.

A second conclusion is that D-PAA performed better than other representations in the majority of the datasets, with respect to classification accuracy.

Finally, the performance of D-PAA is comparable to DTW, which constitutes a state of the art approach in measuring similarity among time series. In fact, there are datasets where D-PAA outperforms DTW.

An expected conclusion that can be drawn from the experiments of the previous section is that there is no representation scheme that outperforms all the others in every dataset. This conclusion refers mainly to their capability of capturing the most essential features of a time series, that is, to minimize the loss of important information inherent in a time series, while reducing dimensionality. Future work will include providing tighter lower bounds on the Euclidean distance and further investigating the weight of means and standard deviations on the purpose of providing a distance measure that is free of parameters.

7. References

[1] E. Keogh, K. Chakrabarti, M. Pazzani and S. Mehrotra, “Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases”, *Knowledge and Information Systems*, February 2001, vol. 3, no. 3, pp. 263-286,

[2] E. Keogh and S. Kasetty, “On the need for time series data mining benchmarks: A survey and empirical demonstration”, *Data Mining and Knowledge Discovery*, Springer Netherlands, October 2003, vol. 7, no. 4, pp. 349-371.

[3] D. Gunopoulos and G. Das, “Time Series Similarity Measures”, *Tutorial notes ACM SIGKDD 2000*, Boston, MA, USA, August 2000, pp. 243-307.

[4] J. Lin, E. Keogh, S. Lonardi and B. Chiu, “A Symbolic Representation of Time Series, with Implications for Streaming Algorithms”, In *Proc. DMKD 2003*, San Diego CA, USA, June 2003, pp. 2-11.

[5] R. Agrawal, C. Faloutsos and A. Swami, “Efficient Similarity Search In Sequence Databases”, In *Proc.4th Int. Conf. FODO*, Evanston, IL, USA, October 1993, pp. 69-84.

[6] K. Chan and A. W. Fu, “Efficient Time Series Matching by Wavelets”, In *Proc. ICDE 1999*, Sydney, Australia, March 1999, pp. 126-133.

[7] F. Korn, H. Jagadish and C. Faloutsos, “Efficiently supporting ad hoc queries in large datasets of time sequences”, In *Proc. ACM SIGMOD 1997*, Tucson, AZ, USA, May 1997, pp. 289-300.

[8] H. Shatkay, “Approximate Queries and Representations for Large Data Sequences”, *Technical Report cs-95-03,* Department of Computer Science, Brown University, 1995.

[9] B. K. Yi and C. Faloutsos, “Fast time sequence indexing for arbitrary Lp Norms”, In *Proc. VLDB-2000*, Cairo, Egypt, September 2000, pp.385-394.

[10] B. Lkhagva, Y. Suzuki and K. Kawagoe, “New Time Series Data Representation ESAX for Financial Applications”, In *Proc. of ICDEW'06*, Atlanta, GA, USA, April 2006, p. x115 -7.

[11] A. Nanopoulos, R. Alcock and Y. Manolopoulos, “Feature-based Classification of Time-series Data”, *International Journal of Computer Research*, Nova Science, 2001, pp. 49-61.

[12] T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, Canada, 2001.

[13] http://www.cs.ucr.edu/~eamonn/time_series_data/

[14] D. Berndt and J. Clifford “Using Dynamic Time Warping to Find Patterns in Time Series”, *AAAI-94 Workshop on Knowledge Discovery in Databases*, 1994, pp. 229-248.