Time Series Discretization Based on the Approximation of the Local Slope Information

Willian Zalewski[1,2], Fabiano Silva[1], Huei Diana Lee[2], Andre Gustavo Maletzke[2], and Feng Chung Wu[2]

[1] Federal University of Parana – UFPR, Curitiba, Brazil
Formal Methods and Artificial Intelligence Laboratory – LIAMF
{wzalewski,fabiano}@inf.ufpr.br
http://www.inf.ufpr.br

[2] State University of West Parana – UNIOESTE, Foz do Iguassu, Brazil
Bioinformatics Laboratory – LABI
{andregustavom,hueidianalee,wufengchung}@gmail.com
http://www.foz.unioeste.br/labi

Abstract. In the last decade symbolic representations approaches have shown effectiveness for knowledge discovery in time series, such as the Symbolic Aggregate Approximation (SAX). However, SAX doesn't preserve the local slope information of the time series because it uses only the mean values of the segments. The modification Extended SAX (ESAX) proposed to treat this problem by the dimensionality increase. In this paper, we present a symbolic representation method that preserves the behavior of local slope characteristics in the symbolic representations of the time series. The proposed method was evaluated with three different discretization approaches and compared with the SAX and the ESAX algorithms. The experimental evaluation, using artificial and real datasets with 1-nearest-neighbor classification, demonstrate the method effectiveness to reduce the error rates of time series classification and to keep the local slope information in the symbolic representations.

Keywords: Time Series, Knowledge Discovery, Symbolic Representation, Classification, Dimensionality Reduction

1 Introduction

The traditional data mining algorithms were developed to analyze data without temporal relation. However, the storage increase of continuous data with temporal interdependencies, such as time series, has motivated the development of new data mining approaches [1, 2]. The time series are collections of observations made chronologically and this type of data is present in almost all domains such as business, industry, medicine, science and entertainment. Time Series Data Mining (TSDM) is a relatively new area that uses data mining methods adjusted to take into consideration the temporal nature of data [3, 4].

Over the last decade many interesting TSDM techniques were proposed and have shown to be useful in many applications [5]. Specifically, symbolic representations have demonstrated to be a very effective tool to reduce the dimensionality

of the time series [2, 6–8] and to preserve the underlying information and produce interpretable symbols within the domain [1, 9].

The most common symbolic representation is the Symbolic Aggregate Approximation (SAX) [2]. The variation Extended SAX (ESAX) was proposed to keep the slope information into symbolic representation. However, the ESAX algorithm causes the increase of the dimensionality and uses additional values of the raw data that can be affected by the noise presence [10].

In previous work [11] we proposed a initial symbolic representation method to preserve the approximated local slope information between the time series observations. In this work, we extend the previous work by presenting a sliding window function to transform the time series data. Furthermore, we propose and evaluate the application of different discretization approaches.

The rest of this paper is organized as follows. Section 2 presents background on time series data mining and related works. Section 3 introduces our symbolic method. Section 4 contains an experimental evaluation of the symbolic method on a variety of the time series datasets. In Section 5 the effectiveness of the symbolic method is also analyzed. Finally, Section 6 presents the conclusions and directions for future works.

2 Background and Related Works

In the context of TSDM, the time series representation is a fundamental problem because direct manipulation of high dimensional data in an efficient way is extremely difficult in traditional data mining techniques. A common approach is to use a time series representation based on some dimensionality reduction technique, while preserving the relevant characteristics of a particular dataset [1, 8]. Many numerical time series representation approaches have been proposed in the literature to reduce the high dimensionality [5, 2].

There are domains, such as medicine and finances, where symbolic representation rather than numerical analysis is needed to produce more comprehensive knowledge of the time series [12]. Many works have also considered symbolic representations of time series, such as Shape Description Alphabet (SDA); Interactive Matching of Patterns with Advanced Constraints in Time Series Databases (IMPACTS); *Clipping*; *Persist*; and Piecewise Vector Quantized Approximation (PVQA) [5, 2, 3].

Most of the symbolic representations cited are affected by two main aspects. Firstly, the intrinsic dimensionality of the symbolic representation is the same as the raw data, thus the data mining algorithms scale poorly with high dimensionality. Second, the unavoidable noise presence in time series can produce meaningless symbols. The SAX is the first symbolic approach that applies dimensionality reduction technique as a preprocessing step, in this case the PAA algorithm [2]. The smoothing property of the PAA contributes to minimize the noise effect.

Piecewise Aggregate Approximation: To transform m-dimensional vector space X to an w-dimensional vector space Y, the data is divided into w equal-size segments, and the mean value of each segment is used to represent original time series with lower w-dimension. The time series $T = \{t_1, \ldots, t_m\}$ of length m can be represented in w-dimensional space by a vector $\overline{T} = \{\bar{t}_1, \ldots, \bar{t}_w\}$ and the ith element of $\overline{T}$ is calculated by the Equation 1 [2]:

$$\bar{t}_i = \frac{w}{m} \sum_{j=\frac{m}{w}(i-1)+1}^{\frac{m}{w}i} t_j \tag{1}$$

Symbolic Aggregate Approximation: The SAX symbolic representation is performed in two steps. First the PAA algorithm is applied to the raw time series (Figure 1(a)). Second, the distribution space (y-axis) is divided into equiprobable regions under a Gaussian curve and the mean segment values from PAA are converted into symbols corresponding to each segment [2]. The SAX symbolic representation can be defined by the function $SAX(\overline{T}, w, a) = \hat{T} = \{\hat{t}_1, \ldots, \hat{t}_w\}$ where $\hat{t}_i$ represent the ith symbol, w is the number of segments and a is the alphabet size. In the Figure 1(b) is presented a SAX example of a symbolic sequence **baabccbc** with the alphabet {a, b, c}.

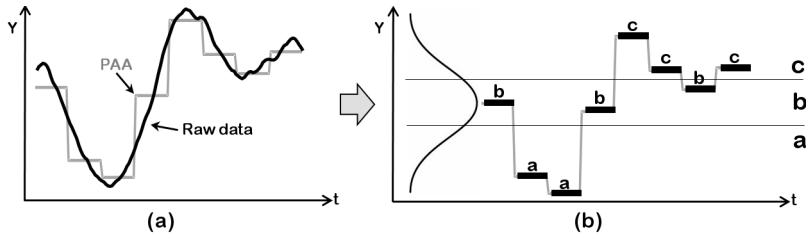

Fig. 1. (a) PAA application example and (b) SAX application example.

In the last decade, SAX has been widely applied to many fields and obtained good results [5, 2]. However, the smoothing effect by only using the PAA algorithm may lose useful information, especially the segment slope information. Furthermore, the equiprobable feature of SAX symbols produces low performance for non-uniform time series [7].

The ESAX approach proposed in [10] is based in addition to the mean value two new symbols for each segment representation, the maximum value and the minimum value of the interval. The ESAX symbolic representation can be defined by the function $ESAX(\overline{T}, w, a) = \hat{T} = \{\hat{t}_1, \ldots, \hat{t}_w\}$ where $\hat{t}_i = \{p_{min}, p_{mid}, p_{max}\}$ is the ith symbolic segment, w is the number of segments and a is the alphabet size. The position of the symbols p_{min} (minimum value), p_{mid} (mean value) and p_{max} (maximum value) in each symbolic segment is determined by the increasing order. However, the ESAX approach has some problems, such as the dimensionality increase by three times the dimensionality of the SAX ap-

proach. Furthermore, the selection of the maximum value and minimum value in each segment can be affected by the noise presence in these points.

In [11] was proposed a symbolic representation method to preserve the approximated local slope information between the time series observations based on the first order differences calculus to the PAA representation and the k-means algorithm to create the symbols. But, this approach also presents some problems such as computing cost to define the initial centers in k-means application and the need to use some training set to create the symbols.

3 Symbolic Representation Method

In this section we present a new symbolic representation method for time series. The method is performed in three sequential steps: Dimensionality Reduction; Data Transformation; and Symbol Creation. The first step is performed by the application of PAA algorithm (Equation 1).

In this work, we proposed an intermediate step between dimensionality reduction and symbolic representation. The data transformation step is used to keep approximated information about the local slope of the time series. In this step, we calculated the first order differences between the adjacent values $\overline{T}$ produced by PAA algorithm.

For each pair of adjacent elements $(i, i+1)$ in the reduced dimension $\overline{T}$, where $1 \leq i \leq w-1$, the new first order difference value is $\delta(i) = \bar{t}_{i+1} - \bar{t}_i$. After, a sliding window function θ of size three is applied for each $\delta(i)$ value where $1 \leq i \leq w-3$. The function θ is defined by the Equation 2. The transformed time series is given by the values $\Theta = \{\theta(1), \ldots, \theta(w-3)\}$.

$$\theta(i) = \frac{\delta(i) + 2 \times \delta(i+1) + \delta(i+2)}{2} \tag{2}$$

The sliding window function θ is used to emphasize continuous adjacent segments in the same direction and to minimize the transitions between adjacent segments with different directions.

Symbol creation is performed based on time series produced by the data transformation step. A discretization algorithm is used to divide in k groups the values $\{\theta(1), \ldots, \theta(w-3)\}$ and to calculate k centroids $C = \{c_1, .., c_k\}$. The values in C are used to associate the values in Θ to symbols. The k value represents the alphabet size for symbolic representation.

The symbol is defined by a function called $Symb$ (Equation 3) that receives a $\theta(i)$ value and the centroids $\{c_1, .., c_k\}$ as input to compute the correspondent symbol.

$$Symb(\theta(i), C) = which.min(\{|c_1 - \theta(i)|, \ldots, |c_k - \theta(i)|\}) \tag{3}$$

where the function $which.min$ finds the c_j value that has the minimum difference to the $\theta(i)$ value, where $1 \leq j \leq w-3$.

The function $Symb$ is applied for each value in Θ and the result set is the symbolic representation $\hat{T}$ of the time series T. The values of the symbols in $\hat{T}$ is

the approximated difference value between the points in $\overline{T}$ representation. Thus, it is possible to associate to the symbols one meaningful information, such as the slope angle which is given by $\alpha = tan^{-1}(c_j)$. By example, suppose the symbolic sequence CBBDACD considering the alphabet size $a = 4$ and the respective centroids $C = \{+2.5, +1.3, -1.2, -2.8\}$ we can represent the sequence by the approximated angles values $(-50^o, +52^o, +52^o, -70^o, +68^o, -50^o, +68^o)$.

In the symbol creation to compute the centroids we proposed three different approaches:

Equal Fixed-Values Discretization (EFVD): In this approach the values from the data transformation step are divided into equal-sized regions between the predefined values $min = tan(-90^o \times \pi/180)$ and $max = tan(+90^o \times \pi/180)$. In the Figure 2(a) is presented a figurative example considering the alphabet size $a = 4$ where the regions can be viewed as a distribution of angles and the set of symbols $\{A, B, C, D\}$ is associated to the mean value of each region. In this example the centroids are $\{A = +67.5^o, B = +22.5^o, C = -22.5^o, D = -67.5^o\}$.

Equal Width Discretization (EWD): This discretization method is performed by dividing the range value, provided by the data transformation step, into equal width regions. In this approach should be used a set of time series to build the intervals. For each region, the mean of the values are calculated and associated to one symbol. The figurative example presented in the Figure 2(b) uses a alphabet size $a = 4$ and the set of symbols $\{A, B, C, D\}$. In this approach the symbol values depend of the contained values in the time series used to the discretization.

Equal Frequency Discretization (EFD): This discretization method is similar to the EWD algorithm, but in this approach the range value is divided into equal frequency of values in each region (Figure 2(c)).

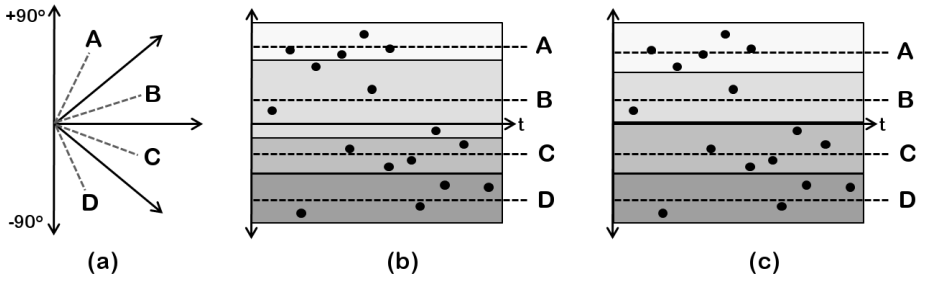

Fig. 2. Discretization algorithms: (a) EFVD; (b) EWD and (c) EFD.

4 Experimental Evaluation and Results

In this section we present an extensive empirical comparison between the symbolic representations SAX, ESAX and our method with the proposed discretiza-

tions approaches EFVD, EWD and EFD. As suggested in [13], we performed an initial experimental classification to the symbolic representations using one nearest neighbor classifier and Euclidean Distance as similarity measure between two symbolic sequences. The method codification and the experimental tests was built using R Language.

In our experiments we used 20 time series datasets provided by the UCR Time Series Data Mining Archive [13] that contains artificial and real-world data. The dataset features are presented in Table 2, such as dataset name, the number of classes (NC), the size of training set (STr), the size of testing set (STe) and the time series length (LS).

Table 1. Summary of datasets.

Dataset Name	NC	STr	STe	LS	Dataset Name	NC	STr	STe	LS
Synthetic Control	6	300	300	60	FaceFour	4	24	88	350
Gun-Point	2	50	150	150	Lightning2	2	60	61	637
CBF	3	30	900	128	Lightning7	7	70	73	319
FaceAll	14	560	1690	131	ECG	2	100	100	96
OSU Leaf	6	200	242	427	Adiac	37	390	391	176
Swedish Leaf	15	500	625	128	Yoga	2	300	3000	426
50words	50	450	455	270	Fish	7	175	175	463
Trace	4	100	100	275	Beef	5	30	30	470
Two Patterns	4	1000	4000	128	Coffee	2	28	28	286
Wafer	2	1000	6174	152	Olive Oil	4	30	30	570

We performed experiments on different combinations of dimensionality reduction and alphabet size for each dataset and for each symbolic representation method. The alphabet size a was evaluated in the interval from 2 until 20 and the dimensionality w in the interval from 2 until 50% of time series length. Each time we increase by two the value of w.

In order for select the parameters w and a for the testing set classification we evaluated the accuracy for each symbolic representation method on training data using leave-one-out cross validation. Sometimes, the correct selection of the optimal values of parameters can be affected in situations where the learning set cannot fully reflect the structure of the test set [14]. Therefore, we have chosen the parameters with the ten best accuracy results for the testing data evaluation. After, the best accuracy among these ten results on each dataset is used for comparison with the other symbolic representations approaches.

The experimental results are shown in Table 2. Accuracy performance for the methods SAX, ESAX, EFVD, EWD and EFD are presented in the 2nt, 3rd, 4th, 5th and 6th columns, respectively (the best accuracy results are bolded). Also the parameters w and a are presented for the methods ESAX, EFVD, EWD and EFD in the 7th, 8th, 9th and 10th columns, respectively (the w value for SAX is a third of ESAX).

As recommended in [15], in order to show that an algorithm is useful, it is necessary predict ahead of time when the method will have superior accuracy.

Table 2. Experimental 1-NN classification results.

Name Dataset	Accur. SAX	Accur. ESAX	Accur. EFVD	Accur. EWD	Accur. EFD	w/a ESAX	w/a EFVD	w/a EWD	w/a EFD
ECG200	0.9000	0.8700	**0.9500**	0.9400	0.9400	96/7	38/19	38/19	42/19
Synthetic	**0.9867**	0.9833	0.9633	0.9567	0.9500	36/13	14/17	12/10	12/10
Coffee	0.8929	**0.9643**	**0.9643**	0.9286	**0.9643**	396/19	20/3	46/13	52/20
CBF	0.9170	0.8989	0.9456	**0.9856**	0.9478	30/14	22/11	8/5	20/9
Beef	0.5667	0.5400	**0.8000**	0.6333	0.6000	84/16	190/2	50/20	202/10
Trace	0.7300	0.7100	0.8100	0.8200	**0.8700**	132/16	82/7	106/8	48/6
SwedishLeaf	0.7648	0.7984	0.8272	0.8144	**0.8432**	126/18	58/20	58/20	48/18
OliveOil	0.1667	0.1667	0.8333	0.8333	**0.9000**	12/2	220/20	200/18	212/19
OSULeaf	0.5290	0.5248	0.5579	0.5620	**0.5827**	156/11	58/17	82/8	170/4
Lightining2	0.7705	0.7869	**0.8197**	0.7869	0.7705	18/19	76/8	52/16	24/18
Lightining7	**0.6576**	0.5480	0.5617	0.6165	0.5891	18/11	24/19	8/6	4/3
Gun Point	0.8200	0.8267	0.9067	**0.9333**	0.9267	18/19	44/19	50/19	46/7
FaceFour	0.7387	0.8523	0.7728	0.8296	**0.8750**	36/18	16/4	28/2	26/7
FaceAll	0.7006	0.7172	**0.7385**	0.7379	0.7299	108/14	36/19	40/19	38/14
Adiac	0.1637	0.1586	0.5729	0.6599	**0.7238**	240/19	52/20	82/16	80/18
50words	0.6022	**0.6726**	0.6374	0.6506	0.6286	48/7	20/14	24/12	32/7
Fish	0.6915	0.6972	0.8343	0.8458	**0.8629**	312/15	120/19	68/17	202/8
Two Patterns	**0.9370**	0.7340	0.9085	0.8890	0.8980	54/5	24/5	18/15	18/9
Yoga	0.8220	0.8230	0.8180	**0.8320**	0.8220	576/16	78/10	106/16	104/6
Wafer	0.9924	0.9926	0.9889	**0.9940**	0.9932	78/4	12/3	48/2	42/2

They proposed the calculus of the function $gain = A/B$ to measure the *expected gain* (on training data) and the *actual gain* (on testing data). The values A and B represents the accuracy performance for a method A and for a method B, respectively. In the Figure 3 is presented the comparison gain for $EFVD/ESAX$, EFD/EFD and $EFD/ESAX$. We remove the datasets *OliveOil* ($gain > 5$) and *Adiac* ($gain > 2.5$) to the best visualization of the charts. The region TP (True Positive) indicates: the method A is more accurate than B for training and testing; the region TN (True Negative): the method B is more accurate than A for training and testing; the region FN (False Negative): the method A is more accurate than B for testing but not for training; the region FP (False Positive): the method B than A is more accurate for testing but not for training.

In the Figure 4 we summarize some results by plotting the dimensionality reduction performance for each dataset into pairwise scatter plots. The points above the diagonal line indicate that the method in the horizontal-axis has a greater dimensionality reduction, and the points below the diagonal line indicate that the method in the vertical-axis has a greater dimensionality reduction.

In the statistical evaluation of the symbolic representations performance, we use the approach applied in [14]. The Iman and Davenport version of the F-test is used to test the null-hypothesis that all symbolic representations have the same performance and the observed differences are merely random. As post hoc test we used the Nemenyi test to compare all methods to each other.

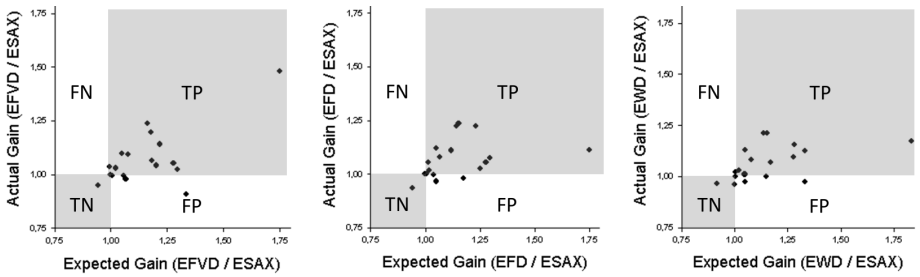

Fig. 3. Gain accuracy comparisons.

In our accuracy analysis the corresponding critical value is equal to 3.92 for $\alpha = 0.05$ (mean ranks: SAX=3.88, ESAX=3.70, EFVD=2.73, EWD=2.28, EFD=2.33). The null-hypothesis that all methods has the same accuracy is rejected (p-value is 0.025). In the post hoc test the rejected comparisons are: EWD*vs*SAX, EFD*vs*ESAX, EWD*vs*ESAX and SAX*vs*EFD.

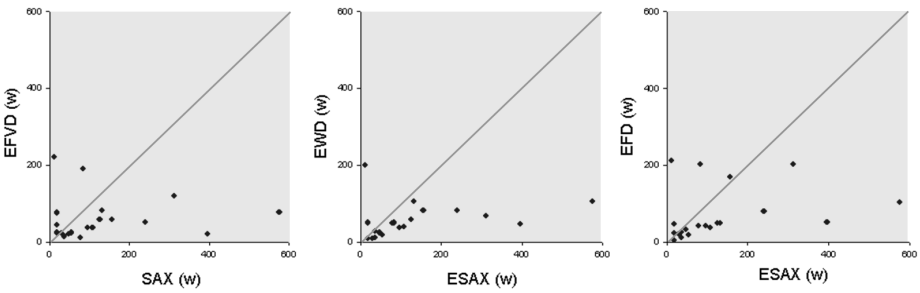

Fig. 4. Dimensionality reduction comparisons.

5 Discussion

Time series data mining techniques have become an important tool to discover novel relevant patterns that can help in decision making process. The human decision making in time series analysis is commonly based on domain expert perceptions [9]. In this cases, a symbolic representation is preferred instead a numerical representation [12] and the symbols should preserve the underlying information [6].

The SAX symbolic representation has been widely used in the literature [5, 2, 7] due to fast processing and smoothing the noise. However, this approach causes a high possibility to miss important patterns in time series data, such as the local trend of the time series [10]. Furthermore, the Gaussian assumption of the symbols distribution has effects on the SAX performance for non-uniform or correlated time series [7]. The ESAX representation was proposed to minimize the missing of the local trend information of the symbols, but the ESAX has

a poor dimensionality reduction and presents the same SAX problems for non-uniform time series.

In this context, we proposed a new symbolic representation method to exclude the existing problems in SAX and ESAX. Our method introduces one intermediate step between dimensionality reduction and symbol creation to preserve the approximated local slope information into the symbols. We also introduce three new discretization algorithms: EFVD, EWD and EFD. The last two are our modifications of existing methods.

The classification accuracy and the dimensionality reduction are the parameters of interest evaluated in this work. In particular, to the approaches based on the preservation of the slope information is desirable to maintain a similar performance or better than SAX to these parameters.

According the charts in the Figure 4 we can see that our method outperforms the dimensionality reduction of the ESAX for most datasets. Furthermore, the results presented in the Figure 3 demonstrate the ability of our method to predict ahead of time when it will have superior accuracy or not. Note in the charts that the most points are into region TP.

Comparing our method EFVD/EWD/EFD and the SAX/ESAX approaches for each dataset accuracy (results in the Table 2) we can observe some very good improvement cases, such as for the datasets *Beef*, *Olive Oil*, *Adiac* and *Fish*. By the other hand, SAX/ESAX approaches do not present expressive improvement for any dataset. For the dimensionality reduction, only the *Olive Oil* dataset presented a poor result to our method.

The statistical evaluation indicates that our method using the EFD and EWD discretization approaches, are more accurate than SAX and ESAX for one nearest neighbor classification. For EFVD, no statistical significant difference in comparison to the other approaches, therefore we can consider that the EFVD have equivalent accuracy performance them. Furthermore, the EFVD discretization do not need to use a training set to calculate the centroids in a previous step, such as need EWD and EFD.

The experimental evaluation presented in this work has demonstrated the competitivity of our method in comparison to SAX and ESAX. In particular, our method is a good symbolic representation alternative to preserve the local slope information, instead ESAX, since has better performance on dimensionality reduction and classification accuracy.

6 Conclusions and Future Works

In this paper we have presented a symbolic representation method to preserve the slope information between the time series segments. We have performed a evaluation on 20 widely used datasets including artificial and real-world time series. The experimental results analysis demonstrate the effectiveness of our representation method in time series classification for low error rates and for dimensionality reduction in comparison with SAX and ESAX approaches.

Future works include the application of the other techniques on our method to improve the dimensionality reduction, such as Adaptive Piecewise Constant

Approximation; evaluate other distance measures, such as Dynamic Time Warping and others Lp-norms; and also test different classification algorithms.

Acknowledgments. We would like to acknowledge Dr. Eamonn Keogh for his experimental datasets.

References

1. Antunes, C.M., Oliveira, A.L.: Temporal Data Mining: an overview. In: KDD Workshop on Temporal Data Mining, pp. 1–13 (2001)
2. Lin, J., Keogh, E., Lonardi, S., Chiu, B.: A symbolic representation of time series, with implications for streaming algorithms. Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery DMKD 03, pp. 2–11 (2003)
3. Karamitopoulos, L., Evagelidis, G.: Current Trends in Time Series Representation. In: Proceedings of 11th Panhellenic Conference on Informatics, pp. 217–226 (2007)
4. Laxman, S., Sastry, P.S.: A survey of temporal data mining. Sadhana. 31, 173–198 (2006)
5. Fu, T.c.: A review on time series data mining. Engineering Applications of Artificial Intelligence. 24, 164–181 (2010)
6. Hugueney, B.: Adaptive segmentation-based symbolic representations of time series for better modeling and lower bounding distance measures. In: Frnkranz, J., Scheffer, T., Spiliopoulou, M. (eds.) PKDD 2006. LNCS, vol 4213, pp. 545–552. Springer, Heidelberg (2008)
7. Pham, N.D., Le, Q.L., Dang, T.K.: Two Novel Adaptive Symbolic Representations for Similarity Search in Time Series Databases. In: Proceedings of the Conference International AsiaPacific Web, pp. 181–187 (2010)
8. Ding, H., Trajcevski, G., Scheuermann, P., Wang, X., Keogh, E.: Querying and mining of time series data: experimental comparison of representations and distance measures. In: Proceedings of the VLDB Endowment, vol. 1, pp. 1542–1552 (2008)
9. Batyrshin, I., Sheremetov, L.: Perception-based approach to time series data mining. Applied Soft Computing. 8, 1211–1221 (2008)
10. Lkhagva, B., Suzuki, Y., Kawagoe, K.: Time Series Representation ESAX for Financial Applications. In: Proceedings of the 22nd International Conference on Data Engineering Workshops, pp. 115 (2006)
11. Zalewski, W., Silva, F., Lee, H. D., Maletzke, A. G., Wu, F. C.: A Symbolic Representation Method to Preserve the Characteristic Slope of Time Series. In: Proceedings of the 21st Brazilian Symposium on Artificial Intelligence (2012)
12. Alonso, F., Martinez, L., Perez-Perez, A., Santamaria, A., Valente, J.P.: Modelling Medical Time Series Using Grammar-Guided Genetic Programming. In: Perner, P. (ed.) ICDM 2008. LNCS, vol 5077, pp. 32–46. Springer, Heidelberg (2008)
13. Keogh, E., Zhu, Q., Hu, B., Hao. Y., Xi, X., Wei, L., Ratanamahatana, C. A.: The UCR Time Series Classification/Clustering Homepage (2011): `www.cs.ucr.edu/~eamonn/time_series_data/`
14. Gorecki, T., Luczak, M.: Using derivatives in time series classification. Data Mining and Knowledge Discovery, pp. 1–22. Springer, Netherlands (2012)
15. Batista, G. E. A. P. A., Wang, X., Keogh, E. J.: A Complexity-Invariant Distance Measure for Time Series. In: SDM 2011, pp. 699–710 (2011)