# An Improved Piecewise Aggregate Approximation Based on Statistical Features for Time Series Mining

Chonghui Guo, Hailin Li, and Donghua Pan

Institute of Systems Engineering,
Dalian University of Technology, Dalian, 116024, China
guochonghui@tsinghua.org.cn, hailin@mail.dlut.edu.cn

**Abstract.** Piecewise Aggregate Approximation (PAA) is a very simple dimensionality reduction method for time series mining. It minimizes dimensionality by the mean values of equal sized frames, which misses some important information and sometimes causes inaccurate results in time series mining. In this paper, we propose an improved PAA, which is based on statistical features including a mean-based feature and variance-based feature. We propose two versions of the improved PAA which have the same preciseness except for the different CPU time cost. Meanwhile, we also provide theoretical analysis for their feasibility and prove that our method guarantees no false dismissals. Experimental results demonstrate that the improved PAA has better tightness of lower bound and more powerful pruning ability.

**Keywords:** Piecewise aggregate approximation, statistical feature, time series, dimensionality reduction.

## 1 Introduction

Time series is a kind of the data with time property. In most cases, efficient and accurate similarity search in time series dataset, including indexing, pattern discovery and association rule, is a very important task for knowledge and information mining. Time series is a kind of data with high dimensionality. It is difficult and inefficient to directly mine time series without dimensionality reduction. Therefore, we should use some methods to reduce the dimensionality before mining.

There has been much work in dimensionality reduction for time series. Some popular methods include discrete fourier transform (DFT) [1], discrete wavelet transform (DWT) [2], singular value decomposition (SVD) [3], symbolic aggregate approximation (SAX) [4, 5] based on piecewise aggregate approximation (PAA) [6], adaptive piecewise constant approximation (APCA) [7]. However, PAA is one of the simplest dimensionality reductions. It reduces dimensionality by the mean of equal sized segments, which also causes the missing of some important features for some kinds of time series datasets. Some people have

Y. Bi and M.-A. Williams (Eds.): KSEM 2010, LNAI 6291, pp. 234–244, 2010.

found the problem and considered the mean values with slope values to improve PAA [10]. What they have done add some valuable information to time series mining and make PAA more perfect.

In this paper we use two important statistic features including the means values and variance values as key factors to improve the original PAA.In fact, besides considering the mean values of the equal sized segments, variance values are also important for time series mining. The improved PAA methods could explain the distribution of the points of each equal sized segment, which provides more information for time series data mining. In this work, besides the theoretical analysis, we also make some experiments to show the performance of the improved PAA methods.

The rest of the paper is organized as follows. Section 2 introduces the original PAA and discusses the existing problem. In section 3 we propose two versions of the new approach. Section 4 shows the experiments and empirical comparisons of our method with completing techniques. In section 5 we offer conclusions and suggestions for the future work.

## 2 Piecewise Aggregate Approximation

Similarity search is a very important task for time series data mining. Since time series is a kind of data with high dimensionality, we always reduce the dimensionality in advance. It means that we should transform data from high space into low space. After dimensionality reduction we can derive a new distance function which is applied to the low space. As proved in reference [1], the function should be a lower bounding measure, which guarantees no false dismissals. In other words, the new function in the low space should underestimate the true distance measure. Conventionally, there are two available true distance measures, Euclidean distance and dynamic time warping (DTW) [8]. In this work, we regard Euclidean distance as the true distance measure.

Suppose we have two time series $Q$ and $C$ of dimension $m$ in original space. Euclidean measure $D(Q, C)$ is the true distance measure function. Another two time series $Q'$ and $C'$ are the new version and $LB_D(Q', C')$is the new distance function in the low space. The two functions must satisfy

$$D(Q, C) \geq LB\_D(Q', C'). \tag{1}$$

PAA is a lower bounding measure function, which is proposed by Lin et al [6]. It is also extended for the better application. Symbolic aggregate approximation (SAX) is used to condense the time series and transform them into a symbolic string, which has wide use in many field. Moreover, since the mature technique PAA is very simple, the extended SAX [9] is also very popular.

A time series $Q$ of length $m$ can be transformed into another form by a vector $\bar{Q} = \bar{q}_1, \bar{q}_2, \cdots, \bar{q}_w$ in a $w$-dimensional space. Then the $ith$ element of $\bar{Q}$ can be calculated by

$$\bar{q}_i = \frac{1}{k} \sum_{j=k(i-1)+1}^{k*i} q_j, \quad where \quad k = \frac{m}{w}. \tag{2}$$

Actually, when original time series in $m$ dimensional space is transformed into another one in $w$ dimensional space, the time series is divided into $w$ segments with equal size. The mean value of the data within a segment is obtained and a vector of these values is the final representation in the new space, as shown in Fig.1.

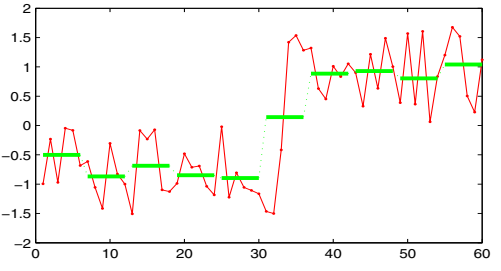


**Fig. 1.** PAA representation is obtained by transformation of space. In this case, a time series of length 60 is reduced to 10 dimensions.

Since PAA is a method based on the mean of each segment, it misses some valuable features for some kinds of time series. From Fig.1 we know that PAA only approximates the whole trend and loses the variance features in each segment. Moreover, PAA causes the inaccurate result. As shown in Fig.2, the two different subsequence of time series, $SQ$ and $SC$, have the same mean. Therefore, the original PAA may produce the inaccurate result for some time series.

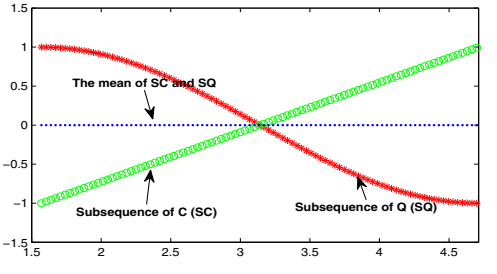


**Fig. 2.** Two different subsequences of time series have the same mean value

For the above case, although the two different subsequences have the same means, they have the different variance. We can use another statistical feature

(variance) to approximate time series because variance can provide some additional information. Therefore, our new approach is based on the mean and the variance of time series for similarity search.

## 3 An Improved Piecewise Aggregate Approximation

Although PAA is one of the best techniques to reduce the dimensionality for time series, it also has some disadvantages. We propose another two versions of new approach which is combination of two statistical features, mean and variance. One version of the new approach is the linear combination of two distance measure functions, which are mean distance measure function and variance distance measure function. We call it Linear Statistical Feature based Piecewise Aggregate Approximation (LSF_PAA). The other is the square root of the sum of the two distance measure functions. We call it Square root Statistical Feature based Piecewise Aggregate Approximation (SSF_PAA).

### 3.1 Linear Statistical Feature Based Piecewise Aggregate Approximation

In section 2, we discuss how to get the mean of each segment. All the mean values of time series are the elements of a new vector in a low space. Likewise, in the new space we can get the variance values of time series.

A time series $Q$ of length $m$ is represented in a $w$ dimensional space by a variance vector $\hat{Q} = \hat{q}_1, \hat{q}_2, \cdots, \hat{q}_w$. The $ith$ element of $Q$ is calculated by

$$\hat{q}_i = \sqrt{\frac{1}{k} \sum_{j=k(i-1)+1}^{k*i} (q_j - \bar{q}_i)^2}, \quad where \;\; k = \frac{m}{w}, \tag{3}$$

where $\bar{q}_i$ is the mean value of the $ith$ segment in $w$ dimensional space.

The two statistical features, the mean value $\bar{q}_i$ and the variance value $\hat{q}_i$, describe the time series in low dimensional space more clearly. Especially the variance can provide additional information and allow the distance measure function in low dimensional space reflect the distribution of the points of every segment. To overcome the disadvantages of PAA, we propose the Linear Statistical Feature based Piecewise Aggregate Approximation (LSF_PAA), which is linear combination of the two distance measure functions.

Given two time series $Q$ and $C$ of length $m$. We can view them as 2 vectors in original space. They can be transformed into 4 new vectors in a $w$ dimensional space, $\bar{Q}$, $\bar{C}$, $\hat{Q}$ and $\hat{C}$. Now we have some distance measure functions about the above 4 new vectors plus 2 original vectors.

The Euclidean distance measure is regarded as the true distance of time series, which is defined by

$$D(Q, C) = \sqrt{\sum_{i=1}^{m} (q_i - c_i)^2}. \tag{4}$$

The mean distance measure function of PAA is

$$\bar{D}(\bar{Q},\bar{C}) = \sqrt{k\sum_{i=1}^{w}(\bar{q}_i - \bar{c}_i)^2}. \tag{5}$$

Similarly, we have a variance distance

$$\hat{D}(\hat{Q},\hat{C}) = \sqrt{k\sum_{i=1}^{w}(\hat{q}_i - \hat{c}_i)^2}. \tag{6}$$

Finally, we obtain the LSF_PAA measure function,

$$LB\_DL(Q,C) = \bar{D}(\bar{Q},\bar{C}) + \mu * \hat{D}(\hat{Q},\hat{C}), \tag{7}$$

where $\mu \in [0,1]$. It is a lower bounding function measure.

**Proposition 1:** If there is $\mu \geq \frac{\sqrt{\bar{D}^2+\hat{D}^2}-\bar{D}}{\hat{D}}$, then we have $D(Q,C) \geq LB\_DL(Q, C)$, i.e. $D(Q,C) \geq \bar{D}(\bar{Q},\bar{C}) + \mu * \hat{D}(\hat{Q},\hat{C})$.

**Proof:** We denote $q_i = \bar{q}_i + \Delta q_i$ and $c_i = \bar{c}_i + \Delta c_i$. For the simple proof, we let $w = 1$, then $q_i = \bar{q} + \Delta q_i$, $c_i = \bar{c} + \Delta c_i$, where $\bar{q}$ and $\bar{c}$ are the mean of Q and C respectively.

$$\begin{aligned}\sum_{i=1}^{m}(q_i - c_i)^2 &= \sum_{i=1}^{m}((\bar{q} + \Delta q_i) - (\bar{c} + \Delta c_i))^2 \\ &= \sum_{i=1}^{m}((\bar{q} - \bar{c}) + (\Delta q_i - \Delta c_i))^2 \\ &= m(\bar{q} - \bar{c})^2 + 2(\bar{q} - \bar{c})\sum_{i=1}^{m}(\Delta q_i - \Delta c_i) + \sum_{i=1}^{m}(\Delta q_i - \Delta c_i)^2.\end{aligned}$$

Moreover, we have

$$\begin{aligned}(\bar{q} - \bar{c})\sum_{i=1}^{m}(\Delta q_i - \Delta c_i) &= (\bar{q} - \bar{c})\sum_{i=1}^{m}((q_i - c_i) - (\bar{q} - \bar{c})) \\ &= (\bar{q} - \bar{c})\sum_{i=1}^{m}(q_i - c_i) - m(\bar{q} - \bar{c})^2 \\ &= (\bar{q} - \bar{c})m(\frac{1}{m}\sum_{i=1}^{m}(q_i - c_i)) - m(\bar{q} - \bar{c})^2 \\ &= (\bar{q} - \bar{c})m(\bar{q} - \bar{c}) - m(\bar{q} - \bar{c})^2 \\ &= 0.\end{aligned}$$

Therefore,

$$D^2(Q,C) = m(\bar{q} - \bar{c})^2 + \sum_{i=1}^{m}(\Delta q_i - \Delta c_i)^2. \tag{8}$$

Notice that,

$$\sum_{i=1}^{m}(\Delta q_i - \Delta c_i)^2 = \sum_{i=1}^{m}\Delta q_i^2 + \Delta c_i^2 - 2\Delta q_i \Delta c_i \tag{9}$$

and

$$\begin{aligned}\hat{D}^2(\hat{Q}-\hat{C}) &= m(\hat{q}-\hat{c})^2 \\ &= m(\sqrt{\frac{1}{m}\sum_{i=1}^{m}\Delta q_i^2} - \sqrt{\frac{1}{m}\sum_{j=1}^{m}\Delta c_j^2})^2 \\ &= \sum_{i=1}^{m}\Delta q_i^2 + \sum_{j=1}^{m}\Delta c_j^2 - 2\sqrt{\sum_{i=1}^{m}\Delta q_i^2 \sum_{j=1}^{m}\Delta c_j^2}.\end{aligned} \tag{10}$$

Therefore, by Cauchy-Schwarz inequality, we have

$$\sqrt{\sum_{i=1}^{m}\Delta q_i^2 \sum_{j=1}^{m}\Delta c_j^2} \geq \sum_{i=1}^{m}\Delta q_i \Delta c_i \tag{11}$$

and

$$\sum_{i=1}^{m}(\Delta q_i - \Delta c_i)^2 \geq \hat{D}^2(\hat{Q}-\hat{C}). \tag{12}$$

Suppose we have

$$D^2(Q,C) \geq (\bar{D}(\bar{Q},\bar{C}) + \mu\hat{D}(\hat{Q},\hat{C}))^2, \quad \mu \in [0,1]. \tag{13}$$

From formula (8) and (12), the inequality (13) becomes

$$\hat{D}^2(\hat{Q},\hat{C}) \geq 2\mu\bar{D}(\bar{Q},\bar{C})\hat{D}(\hat{Q},\hat{C}) + \mu^2\hat{D}^2(\hat{Q},\hat{C}).$$

Since there is $\hat{D}(\hat{Q},\hat{C}) \geq 0$, we have $\hat{D}(\hat{Q},\hat{C}) \geq 2\mu\bar{D}(\bar{Q},\bar{C}) + \mu^2\hat{D}(\hat{Q},\hat{C})$. Through solving this one-variable quadratic inequality with respect to $\mu$, the maximum value of the variable $\mu = \frac{\sqrt{\bar{D}^2+\hat{D}^2}-\bar{D}}{\hat{D}} \in [0,1]$ is obtained.

The proof is finished.

By far we have taken a theoretical analysis which proves that LSF_PAA is a lower bounding measure. Moreover, it is a better tightness of lower bound, i.e.

$$D(Q,C) \geq LB\_DL(Q,C) \geq \bar{D}(\bar{Q},\bar{C}). \tag{14}$$

In fact, LSF_PAA is comprised of the measure function of PAA and the variance measure function. The contribution is $\mu * \hat{D}(\hat{Q},\hat{C})$. Therefore, from the mathematical analysis we also know LSF_PAA is tighter than original PAA.

### 3.2 Square Root Statistical Feature Based Piecewise Aggregate Approximation

Another version of our new approach is the square root of the sum of the mean and the variance distance measure functions. We call it Square root Statistical Feature based Piecewise Aggregate Approximation (SSF_PAA). Just as illustrated in subsection 3.1, we have the mean distance measure and the variance distance measure. The SSF_PAA distance measure function is

$$LB\_DS(Q,C) = \sqrt{\bar{D}^2(\bar{Q},\bar{C}) + \hat{D}^2(\hat{Q},\hat{C})}. \tag{15}$$

**Proposition 2:** If there is $LB\_DS(Q,C) = \sqrt{\bar{D}^2(\bar{Q},\bar{C}) + \hat{D}^2(\hat{Q},\hat{C})}$, then we have $D(Q,C) \geq LB\_DS(Q,C)$, i.e. $LB\_DS(Q,C) \geq \sqrt{\bar{D}^2(\bar{Q},\bar{C}) + \hat{D}^2(\hat{Q},\hat{C})}$

**Proof:** From formula (8), (9), (10) and (11), we derive $\sum_{i=1}^{m}(q_i - c_i)^2 \geq m(\bar{q} - \bar{c})^2 + \hat{D}^2(\hat{Q},\hat{C})$, i.e.

$$D^2(Q,C) \geq \bar{D}^2(\bar{Q},\bar{C}) + \hat{D}^2(\hat{Q},\hat{C}).$$

The proof of the SSF_PAA is finished.

Through the above mathematical analysis, SSF_PAA is also a lower bounding measure function, which can guarantee no false dismissals. Moreover, the low bounding of this function is tighter than PAA too. It is also easy to obtain

$$D(Q,C) \geq LB\_DS(Q,C) \geq \bar{D}(\bar{Q},\bar{C}). \tag{16}$$

### 3.3 Complexity Analysis of LSF_PAA and SSF_PAA

The above two versions of the new improved PAA have the same effectiveness to reduce dimensionality. Moreover, their tightness of lower bound is identical. However, their time consumption of similarity search is different.

A time series $Q$ of length $m$ is used to query its similar objects by brute-force searching method in time series dataset with $n$ time series. In LSF_PAA, it calculates the parameter value $\mu$ in advance for each pair of time series, which totally costs $O(mn)$. However, since SSF_PAA has no any parameter, it doesn't cost the time. Therefore, SSF_PAA is faster than LSF_PAA.

From the above formulas, we know that each version of the new approach has one more function than PAA, which causes more time consumption than PAA. The additional time of SSF_PAA is used to calculate the variance value of each segment. This additional time is equal to the time consumption of the mean values of each segment. Thereby, the time complexity of SSF_PAA is twice as that of PAA. In LSF_PAA, because it needs to cost $O(mn)$ to calculate the parameter and also needs to calculate the mean and variance value, its time consumption is much higher than PAA. In conclusion, the time consumption of SSF_PAA is a little higher than PAA but lower than LSF_PAA.

If we only consider the precision of the similarity search, we can choose the two versions. If we consider the productivity of search algorithm, the SSF_PAA is a better choice.

## 4 Numerical Experiments

In this section we test our approach with a comprehensive set of experiments. The experimental data is from the UCI Dataset (Synthetic Control Chart Time Series) [11]. The purpose of the experiments is to concentrate on reducing the dimensionality without false dismissals. The comparison among PAA, SSF_PAA and LSF_PAA is based on the tightness of lower bounds, pruning power and implement system as introduced in reference [12].

### 4.1 Comparison of Lower Bounding Measures

From the above discussions, we can conclude that the tightness of the two versions of improved PAA is better than the original PAA. Now we empirically test whether the conclusion is true. Since the three versions of PAA are the approximation of the true distance, we directly call them estimated distance. We let $T$ represent tightness and define it as the ratio of the estimated distance between two time series over the true distance between the same time series, i.e. $T = D'(Q, C)/D(Q, C)$, where $D(Q, C)$ is a true distance measure and $D'(Q, C)$ is the estimated distance measure including PAA, SSF_PAA and LSF_PAA. $T$ is in the range $[0, 1]$, with the larger the better.

To experiment on the tightness of lower bound of the three versions, 600 time series of length 60 were tested. Each time series was condensed into a new vector of elements. We compare each time series to the other 599 and report $T$ as average ratio from 179700 (600*599/2) comparisons we made. The tightness results of the three versions of PAA with regard to reduction ratio $k = m/w$ are shown in Fig.3.

We found that the tightness of SSF_PAA is the same to that of LSF_PAA, which means the two version of improved PAA can produce same distance measure. Although they are identical, they are larger than the original PAA for each

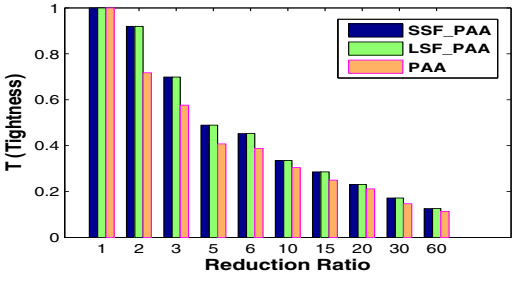


**Fig. 3.** The empirically estimated tightness of the three versions of PAA

reduction ratio. When the reduction ratio is equal to 1, it means the two versions correspond to the original PAA.

### 4.2 Comparison of Pruning Power

Pruning power indicates that the estimated measure function can decrease the number of time series which require full computation of true distance when indexing time series or similarity search. In other words, it is the fraction of the dataset that must not be examined before we can guarantee that we have found the nearest query. We define it as the ratio of the number of objects that do not require full computation of true distance over the number of object existing in the time series database, i.e.

$$P = \frac{Number\ of\ objects\ that\ are\ not\ exmined\ by\ true\ distance\ function}{Number\ of\ objects\ in\ database} \tag{17}$$

We randomly extract 100 time series from the database and regard them as query time series. We let everyone search the similar time series with regard to the reduction ratio $k = m/w$. In Fig.4, the result of pruning power shows that the two version of the improved PAA also have the same ability to reject the time series which do not require full computation by the true distance measure. Fortunately, they are larger than the original PAA. Moreover, when the reduction ratio is 1, namely, $w = 60$, it means the two version of the improved PAA degenerate to the original PAA.

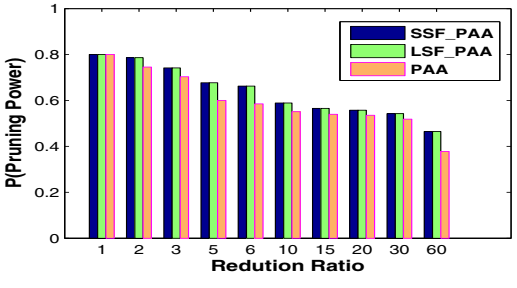


**Fig. 4.** The pruning power of the three version of PAA

### 4.3 Experiments on Implemented System

To evaluate the performance of the three versions of PAA, we test the normalized CPU cost. We define the normalized CPU cost as the ratio of average CPU time of the three version of PAA to query the time series over the average CPU time of PAA. It is easy to know that the normalized CPU time of PAA is equal to 1. We performed the experiments on Intel Core2 Duo 2.00GHZ processor with 2.00GB physical memory. We also experiment on the time series database and

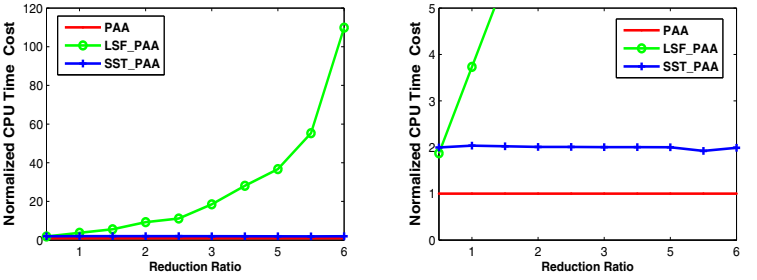


**Fig. 5.** The Normalized CPU time cost of the three versions of PAA. (a) The original view of the normalized CPU time cost; (b) The zoom view of the normalized CPU time cost.

arbitrarily choose 150 time series as the query ones. We let them query in the database according to different reduction ratio $k = m/w$. The Fig.5(a) shows the result of the experiment.

There are at least three kinds of information hiding in Fig.5(b). The first is that the time consumption of LSF_PAA is much too large. The second is that the time consumption of SSF_PAA is approximately twice as that of PAA. We also can find that the normalized CPU time of LSF_PAA is equal to that of SSF_PAA when the reduction ratio $k$ is equal to 1. In this case, it means the LSF_PAA corresponds to SSF_PAA, which is the last information. We also point out that the larger the size of the database is, the more time is cost by SSF_PAA, letting alone the LSF_PAA.

## 5 Conclusions

The main contribution of this paper is to propose two improved versions of the improved PAA in light of the precision. Meanwhile, we also provide their mathematical proof with regard to the lower bounding measure. Through considering the variance of each segment in the low space, we can let the new approach provide additional information and make new estimated distance measure have better tightness of lower bound. From the view of the accuracy of similarity search, the two versions of PAA are better than the original one. Although they cost more time, fortunately, SSF_PAA's time consumption is only twice as that of the original PAA and its effectiveness is better. LSF_PAA is not suitable for the large time series database because of its high time complexity. However, its idea is worth further researching in the future.

Considering the limitation of the improved approach, one of future work is to find a way to decrease time consumption. Since PAA has many extensions and is applied to various fields, we may extent the improved approach like the means of extended PAA.

## Acknowledgment

This work has been partly supported by the Natural Science Foundation of China under Grant No. 70871015 and the National High Technology Research and Development Program of China under Grant No. 2008AA04Z107.

## References

1. Faloutsos, C., Ranganathan, M., Manolopoulos, Y.: Fast subsequence matching in time series databases. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 419–429 (1994)
2. Popivanov, I., Miller, R.J.: Similarity search over time-series data using wavelets. In: Proceedings of the 18th International Conference on Data Engineering, pp. 212–221 (2002)
3. Theodoridis, S., Koutroumbas, K.: Feature generation I: data transformation and dimensionality reduction. In: Pattern Recognition, 4th edn., pp. 323–409 (2009)
4. Lin, J., Keogh, E.: A symbolic representation of time series with implications for streaming algorithms. In: Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, pp. 2–11 (2003)
5. Keogh, E., Lin, J., Fu, A.: Hot sax: efficiently finding the most unusual time series subsequence. In: Proceedings of the 5th IEEE International Conference on Data Mining, pp. 226–233 (2005)
6. Lin, J., Keogh, E., Wei, L., Lonardi, S.: Experiencing SAX: a novel symbolic representation of time series. Data Mining and Knowledge Discovery 15, 107–144 (2007)
7. Keogh, E., Chakrabarti, K., Mehrotra, S., Pazzani, M.: Locally adaptive dimensionality reduction for indexing large time series databases. In: Proceedings of the ACM SIGMOD Conference on Management of Data, pp. 151–162 (2001)
8. Rabiner, L., Juang, B.H.: Fundamentals of speech recognition. Prentice Hall, Englewood Cliffs (1993)
9. Lkhagva, B., Suzuki, Y., Kawagoe, K.: New time series data representation ESAX for financial applications. In: Proceedings of the 22nd International Conference on Data Engineering Workshops, pp. 115–120 (2006)
10. Hung, N.Q.V., Anh, D.T.: An improvement of PAA for dimensionality reduction in large time series databases. In: Proceedings of the 10th Pacific Rim International Conference on Artificial Intelligence: Trends in Artificial Intelligence, pp. 698–707 (2008)
11. Pham, D.T., Chan, A.B.: Control chart pattern recognition using a new type of self organizing neural network. Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering 212, 115–127 (1998)
12. Keogh, E., Ratanamahatana, C.A.: Exact indexing of dynamic time warping. Knowledge and Information Systems 7, 358–386 (2005)