Extended SAX: Extension of Symbolic Aggregate Approximation for Financial Time Series Data Representation

Battuguldur Lkhagva, Yu Suzuki and Kyoji Kawagoe

Graduate School of Science and Engineering
Ritsumeikan Univ.
Nojihigashi 1-1-1, Kusatsu, Shiga, 525-8577 Japan
{tuguldur,yusuzuki,kawagoe}@coms.ics.ritsumei.ac.jp

Abstract

Efficient and accurate similarity searching for a large amount of time series data set is an important but non-trivial problem. Many dimensionality reduction techniques have been proposed for effective representation of time series data in order to realize such similarity searching, including Singular Value Decomposition (SVD), the Discrete Fourier transform (DFT), the Adaptive Piecewise Constant Approximation (APCA), and the recently proposed Symbolic Aggregate Approximation (SAX).

In this work we propose a new extended approach based on SAX, called Extended SAX in order to realize efficient and accurate discovering of important patterns, necessary for financial applications. While the original SAX approach allows a very good dimensionality reduction and distance measures to be defined on the symbolic approach, SAX is based on PAA (Piecewise Aggregate Approximation) representation for dimensionality reduction that minimizes dimensionality by the mean values of equal sized frames. This value based representation causes a high possibility to miss some important patterns in some time series data such as financial time series data.

Extended SAX, proposed in the paper, uses additional two new points, that is, max and min points, in equal sized frames besides the mean value for data approximation. We show that Extended SAX can improve representation preciseness without losing symbolic nature of the original SAX representation. We empirically compare the Extended SAX with the original SAX approach and demonstrate its quality improvement.

Keywords
Time series, Representation, Data mining, Symbolic

1. Introduction

Efficient and accurate similarity searching for a large amount of time series data set is an important but non-trivial problem. Many dimensionality reduction techniques [1, 2, 3, 4, 5, 7] have been proposed for effective representation of time series data. Symbolic Aggregate Approximation (SAX) was proposed as a new method for time series data representation [3].

In this work, we propose a new extended approach based on SAX which we call Extended SAX. The original SAX approach allows a very good dimensionality reduction and distance measures to be defined on the symbolic approach. However, SAX is based on the PAA representation for dimensionality reduction that minimizes dimensionality by the mean values of equal sized frames. This mean value based representation causes a high possibility to miss some important patterns in some time series data such as financial time series data.

Financial time series data has its own characteristics over other time series data. One of its special characteristics is that it is typically characterized by a few critical points and multi-resolution consideration is always necessary for long-term and short-term analyses. Second one is that financial time series data is continuous, large and unbound. There are many technical analytical methods for financial time series data to identify patterns of market behavior. In those financial analytical methods, critical or extreme points, which the original SAX cannot handle, are very important to discover. To reduce a loss of these important points, Extended SAX representation especially for financial data analysis and mining tasks is proposed.

The rest of this paper is organized as follows. Section 2 briefly discusses the existing research work on time series data mining. Section 3 introduces our proposed approach, and discusses its dimensionality reduction and quality improving abilities. Section 4 contains an experimental evaluation of the approach. Finally, Section 5 offers some conclusions and suggestions for future work.

2. Background

In time series data study, the well defined and approximated representation for the original data is the most important topics in order to solve many time series data mining problems. Many approaches and techniques that address the time series data representation, have been proposed in the past decade.

Most commonly used representations are the Discrete Fourier Transform (DFT) [1], the Discrete Wavelet Transform (DWT) [2], Singular Value Decomposition (SVD) [4], Adaptive Piecewise Constant Approximation (APCA) [5], , and Piecewise Aggregate Approximation (PAA) [4,7]. Recently, one promising representation method was proposed called Symbolic Aggregate Approximation (SAX) [3].

The basic idea of our method proposed in the paper, is based on the last two approaches among these representation techniques. These two methods are the PAA and the SAX representations, which are briefly described below in this section.

2.1. Piecewise Aggregate Approximation (PAA)

Yi and Faloutsos [7] and Keogh et al. [4] independently proposed PAA. In PAA, each sequence of time series data is divided into *k* segments with equal length and the average value of each segment is used as a coordinate of a *k*-dimensional feature vector. The advantages of this transform are that 1) it is very fast and easy to implement, and 2) the index can be build in linear time.

According to [3], "to reduce the time series from *n* dimensions to *k* dimensions, the data is divided into *k* equal sized segments. The mean value of the data within a segment is calculated and a vector of these values becomes the data-reduced representation" as shown in Figure 1

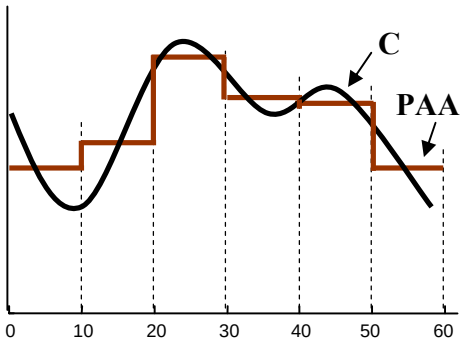

Figure 1: A time series C is represented by PAA (by the mean values of equal segments). In the example above, the dimensionality is reduced from *n* = 60 to *k* = 6.

More formally, as defined in [3], "a time series *C* of length *n* can be represented in a *k*-dimensional space by a vector *k* and the *i*th element of *C* is calculated by the following equation":

$$x_i = \frac{k}{n} \sum_{j=\frac{n}{k}(i-1)+1}^{\frac{n}{k}i} c_j \qquad (1)$$

However the PAA approach has numerous other advantages over other techniques, it also has some disadvantages. As noted in the introduction section, the PAA approach minimizes dimensionality by the mean values of equal sized frames. This mean value based representation may cause a possibility to miss some important patterns in some time series data analysis.

2.2. Symbolic Aggregate Approximation (SAX)

Lin and Keogh et al. [3] proposed new approach called SAX. SAX is based on PAA [4, 7] and assumes normality of the resulting aggregated values. As noted by authors [3], SAX is the first symbolic representation of time series with an approximate distance function that lower bounds the Euclidean distance.

In SAX, firstly the data is transformed into the PAA representation and then the transformed PAA representation is symbolized into a sequence of discrete strings.

As said in [3], there are two important advantages to doing this:

- **Dimensionality Reduction: "**Dimensionality reduction of PAA [4, 7] is automatically carried over to this representation."
- **Lower Bounding:** Distance measure between two symbolic strings can be proved "by simply pointing to the existing proofs for the PAA representation itself ".

In order to obtain string representation after a time series data is transformed into the PAA representation, symbolization region should be determined. According to [3], by empirically testing more than 50 datasets, it was defined that normalized subsequences have highly Gaussian distribution.

Definition 1 [3]: "*Breakpoints*: breakpoints are a sorted list of numbers $B = \beta_1,\ldots,\beta_{a-1}$ such that the area under a $N(0,1)$ Gaussian curve from β_i to $\beta_{i+1} = 1/a$ (β_0 and β_a are defined as $-\infty$ and ∞, respectively)."

According to [3] "These breakpoints may be determined by looking them up in a statistical table". For example, Table 1 gives the breakpoints for values of *a* changing from 3 to 5.

a	3	4	5
β1	-0.43	-0.67	-0.84
β2	0.43	0	-0.25
β3		0.67	0.25
β4			0.84

Table 1 [3]: "A lookup table that contains the breakpoints that divides a Gaussian distribution in an arbitrary number (from 3 to 5)."

As said in [3], using these defining breakpoints, a time series can be discretized as "All PAA coefficients that are below the smallest breakpoint are mapped to the symbol "**A**," all coefficients greater than or equal to the smallest breakpoint and less than the second smallest breakpoint are mapped to the symbol "**B**," etc". Figure 2 illustrates the idea.

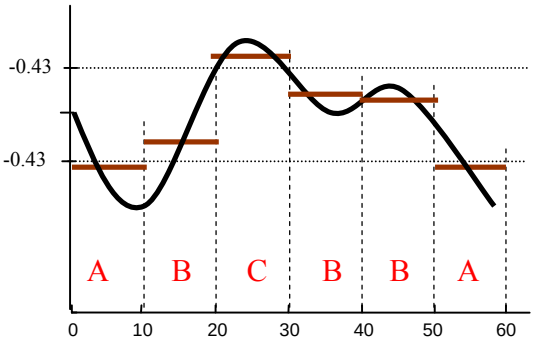

Figure 2: A time series is discretized by SAX. In the example above, with *n* = 60, *k* = 6 and *a* = 3, the time series is mapped to the word *ABCBBA*.

SAX has many advantages over other symbolic approaches such as dimensionality reduction power and lower bounding the Euclidean distance. But, as noted before, it has also some disadvantages such as the dimensionality reduction nature that has possibility to miss important patterns in some datasets.

3. Extension of Symbolic Aggregate Approximation

3.1. Adding two new values to represent time series

In this section, a new proposed time series representation is introduced. It is based on the original SAX. As briefly explained previous section, SAX is based the PAA approach to reduce dimensionality of time series data. The PAA approach minimizes dimensionality by the mean values of equal sized segments.

Minimizing dimensionality by the mean values of equal sized segments has many advantages such as fast, flexible and easy to implement. However, it has also some disadvantages for some kind of time series data, especially for financial time series data set. Our main research goal was to define effective and accurate algorithms to find extreme and unusual patterns in financial time series data by using the original SAX representation. But the original SAX approach was not suitable for financial time series data. It is because its dimensionality reduction nature, based on mean values approximation, has high possibility to miss some important patterns. Figure 3 shows intuition of this pattern missing characters of SAX.

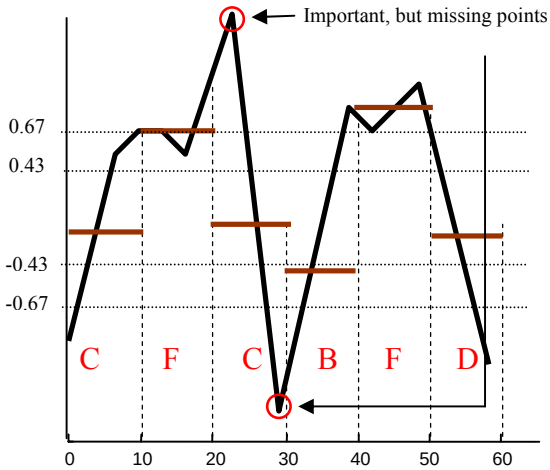

Figure 3: Financial time series data is represented by SAX. Some important points (shown in red) are missing. (US$ and Japanese yen exchange rate data of 2 months.) The SAX representation is *CFCBFD*.

In 3rd equal sized segment, shown in Figure 3, there are two very important and extreme, points, shown in small circle. In some kind of time series data

research, especially in financial time series data, these kind of important points are very meaningful. From the figure 3, we can see the miss of two important points, shown in small circle, when financial time series data is represented by the original SAX approach. The mean value of the 3rd equal sized segment is represented as symbol "C". But, in the segment, there are points which should be represented as symbol "F", the max point of the segment, and symbol "A", the min point of the segment. So, it is important to represent such important points, which is the motivation of our new extended approach.

We extend SAX by adding two special new points, that is, max and min points of each segment, for each segment to fully represent time series data. Therefore, in our Extended SAX, three values for each segment, the original mean values and these additional two new points, min and max points, are used for time series data representation.

3.2. Locating max, min, and mean values in segment

As our Extended SAX approach is based SAX, we can get the mean values of the PAA of the financial time series. After obtaining PAA, we have the equal sized segments and its mean values. Then we define max and min values in the each segment. Figure 4 illustrates the idea. Max values and min values are respectively shown in red circles and in blue squares, while mean values are shown in brown triangles in Figure 4.

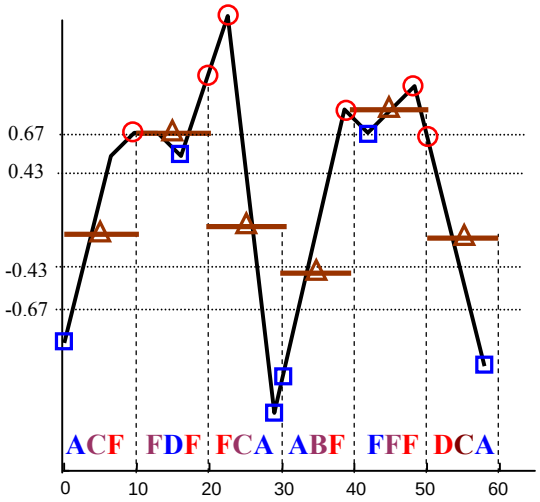

Figure 4: Financial time series data is represented by Extended SAX. The Extended SAX representation is *ACFFDFFCAABFFFFDCA*. (US$ and Japanese yen exchange rate data of 2 months.)

We have two additional values to represent time series beside mean value. So, the positions of these three values are located in the segments, as shown in Figure 5.

Locating process is done in the following manner. For any given segment C_k in a time series data , from both the beginning position S_k on the time axis, and the ending position E_k, the middle position p_{mid} of the segment is be calculated using the following equation (Eq 2).

$$p_{mid} = \frac{S_k + E_k}{2} \qquad (2)$$

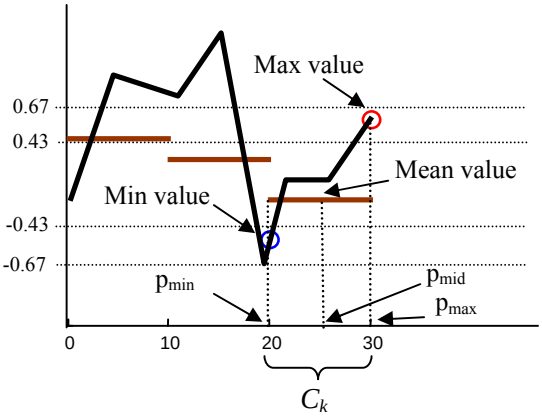

Figure 5: Intuition of locating process. Locating the positions of three important values (max, min, mean). p_{max}, p_{mid}, p_{min} are their respective positions in the segment.

Supposed that s_1, s_2, and s_3 are the first, the second, and the third the symbols in *k-th* segment and that p_{max}, p_{mid}, p_{min} are positions of max, min, mean values in the *k-th* segment on the time axis. Also we assume that s_{max}, s_{mid}, s_{min} are strings of max, min, mean values in the *k-th* segment. Ordering of three strings in the *k-th* segments can be calculated by following expression.

$$<s_1,s_2,s_3>= \begin{cases} <s_{max},s_{mid},s_{min}> \text{ if } p_{max} < p_{mid} < p_{min} \\ <s_{min},s_{mid},s_{max}> \text{ if } p_{min} < p_{mid} < p_{max} \\ <s_{min},s_{max},s_{mid}> \text{ if } p_{min} < p_{max} < p_{mid} \\ <s_{max},s_{min},s_{mid}> \text{ if } p_{max} < p_{min} < p_{mid} \\ <s_{mid},s_{max},s_{min}> \text{ if } p_{mid} < p_{max} < p_{min} \\ <s_{mid},s_{min},s_{max}> \text{ otherwise} \end{cases} \qquad (3)$$

The discretization of the three values is done as the similar way as the original SAX does, which defined in section 2.2. After the discretization, a sequence of symbols is obtained. We call it *StringSequence* S of time series data.

Definition 2. *String Sequence*: A time series $C = c_1, c_2, \ldots, c_n$ is represented by the Extended SAX using *String Sequence S*. *S* is a sequence of symbols $<s_1^1,s_2^1,s_3^1,s_1^2,s_2^2,s_3^2,\ldots,s_1^j,s_2^j,s_3^j,\ldots,s_1^k,s_2^k,s_3^k>$ where $<s_1^j,s_2^j,s_3^j>$ is an ordered collection of strings for j-*th* segment calculated in the Eq.(3).

3.3. Distance Function

Extending the original SAX representation of time series, we need to define a distance measure on it. As noted before, the SAX representation is based on the PAA representation. Distance function of PAA is defined in [4, 7]. Given a query sequence *Q,* time series data *C* of the PAA representation of the same length *n,* Eq. 4 defines lower bounding approximation of the Euclidean distance:

$$Dpaa\ (Q,C) = \sqrt{\frac{n}{k}}\sqrt{\sum_{i=1}^{k}(q_i - c_i)^2} \qquad (4)$$

A proof that *Dpaa*(*Q* ,*C*) lower bounds the true Euclidean distance can be found in [4,7]. Since we use three symbols in each segment instead of one symbol, we can define distance function as following equation.

$$D_e(Q_e, C_e) = \sqrt{\frac{n}{k}}\sqrt{\sum_{i=1}^{k}(dist(s_i, r_i))^2} \qquad (5)$$

where $<s_1, s_2,\ldots, s_k>$ and $<r_1, r_2,\ldots, r_k>$ are string sequences of Q_e and C_e, the Extended SAX representations of Q and C, respectively. As said in [3], "The *dist*() function can be implemented using a lookup tables such as illustrated in Table 1".

4. Preliminary experimental evaluation

In the evaluation work, we show our new approach's advantages, especially how accurate our approach is. We use both computer generated datasets and real datasets. We used a real financial time series sets which are downloadable from the net [8, 9]. The average length of sequences in both generated and real datasets is 5000 points.

Since our goal is to show the accuracy improvement of new approach, we performed subsequence search in datasets by the simple brute force algorithm. We searched the same length subsequence in the same datasets by both the original SAX representation and the proposed representation to show which one has the best accuracy. Experiment process can be defined more formally as following: a query sequence *Q* and a time series datasets *C*. The task is to find all the subsequences in *C* that match *Q*. Subsequence matching requires the query *Q* to be placed at every possible segment within the dataset *C*. Firstly a query sequence *Q* and a time series datasets *C* both were represented by the original SAX. A query sequence *Q* is chosen as randomly from processing dataset.

Then the subsequence search process was done. The same process was done in the new approach using the same length subsequence. We did experiments on 20 datasets. Table 2 shows the result.

Methods	Average number of matching results
SAX	4.90
ESAX	1.66

Table 2: Average number of matching results of the SAX and the Extended SAX. The average point in datasets is 5000. Dimensionality reduction on the SAX is 16 and on the Extended SAX is 48. Alphabet size is a = 6 for both approaches.

Figure 6 explains one example of the experimental results in Table 2. The picture a) is the query sequence. b) and c) are matched results by Extended SAX. d), e), f), j) and i) are matched results by the original SAX.

Extended SAX results, shown in b) and c), are much similar to the query *Q* compare with the original SAX. The original SAX gets 4 false results and only d) is similar to the query *Q*. As shown in Figure 6, it becomes clear that for the financial datasets, Extended SAX helps to detect important patterns and to improve detecting accuracy. From this preliminary evaluation, we can see that the original SAX matching will generate more false results.

This result may be said that it is natural that the better result is achieved by using more information because ESAX uses more information for representing time series. In our experiment, the dimensionality reduction of the SAX is 16 and the dimensionality reduction of our approach is 48. It means that string representation of Extended SAX is three times longer than that of SAX. But if we compare the dimensionality reduction (48) of ESAX to the dimensionality reduction (16) of the original SAX, it is

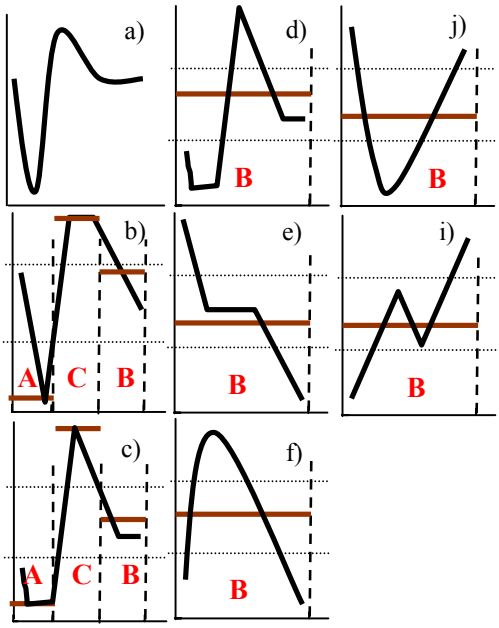

Figure 6: Example of experiment results matched by the SAX and the Extended SAX. a) Query sequence, b) and c) Matched results by the Extended SAX, d), e), f), j) and i) Matched results by SAX.

not a big difference as we compare to the original data of 5000 points. If we compare the both dimensionality reductions (16, 48) to the average data points (5000) of the datasets, it is a big difference. But our approach shows capability to represent more meaningful representations than SAX does as explained in Figure 6.

We are currently under further detail evaluation with much data set and more appropriate conditions.

5. Conclusions and suggestions for future work

In this work, we proposed the extended approach of SAX of time series data representation. By adding more important points in equal sized segments without losing symbolic nature of the original approach, Extended SAX provides a more meaningful representation for many different datasets, especially for the high frequency dataset such as financial dataset.

In financial time series analysis, extreme point movements and high frequency movements of time series are very important and critical for financial decisions. We hope that our proposed approach can improve the accuracy of the financial time series analysis by more meaningful representation.

There can be several future research directions using this approach. The preliminary experimental results presented here mainly focus on similarity searching. We think that this approach also can be effectively used for other data mining tasks such as clustering, anomaly detection, classification and other tasks. Since our approach increases dimensionality to represent time series more effectively, there may be more work to reduce dimensionality without losing the quality improvement. Lastly we are going to show low-bounding property of the distance function used in the paper.

Acknowledgments

This work was supported by MEXT.HAITEKU(2005).

References

[1] Agrawal, R., Faloutsos, C., & Swami, A. "Efficient similarity search in sequence databases" *Proceedings of the 4th Conference on Foundations of Data Organization and Algorithms*.(1993)

[2] Chan, K. & Fu, W. "Efficient time series matching by wavelets", *Proceedings of the 15th IEEE International Conference on Data Engineering*. (1999).

[3] Lin, J., Keogh, E., Lonardi, S. & Chiu, B. "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*. (2003).

[4] Keogh, E,. Chakrabarti, K,. Pazzani, M. & Mehrotra "Dimensionality reduction for fast similarity search in large time series databases", *Journal of Knowledge and Information Systems*. (2000).

[5] Keogh, E., Chakrabarti, K., Pazzani, M. & Mehrotra, S. "Locally adaptive dimensionality reduction for indexing large time series databases", *In proceedings of ACM SIGMOD Conference on Management of Data. Santa Barbara, CA, May 21-24. pp 151-162.* (2001).

[6]Keogh, E., Chu, S., Hart, D. & Pazzani, M. "An Online Algorithm for Segmenting Time Series". *In Proceedings of IEEE International Conference on Data Mining. pp 289-296.* (2001).

[7] Yi, B-K and Faloutsos, C., "Fast Time Sequence Indexing for Arbitrary Lp Norms", *Proceedings of the VLDB, Cairo, Egypt, Sept*, (2000).

[8] Time Series Data Library. *http://www-personal.buseco.monash.edu.au/~hyndman/TSDL/*

[9] Web Page for Analysis of Financial Time Series. http://gsbwww.uchicago.edu/ fac/ruey.tsay/teaching/fts/