Australian Journal of Basic and Applied Sciences, 8(6) April 2014, Pages: 57-67

AENSI Journals

Australian Journal of Basic and Applied Sciences

ISSN:1991-8178

Journal home page: www.ajbasweb.com

Optimizing Similarity Search Using Symbolic Recently-Biased Methodology

Tamer Hassan Abd El Salam, Assoc. Prof. Dr. Zalinda Othman, Prof. Dr. Abdul Razak Hamdan

[1]*Universiti Kebangsaan Malaysia, Computer Science, Faculty of Information Science and Technology, 43600 UKM Bangi, Selangor. Malaysia.*

ARTICLE INFO

Article history:
Received 25 January 2014
Received in revised form
8 April 2014
Accepted 20 April 2014
Available online 10 May 2014

Keywords:
Time series data representation; dimensionality reduction; data mining, similarity search; optimization

ABSTRACT

An efficient and accurate similarity searching on a huge time series data set is a crucial problem in data mining preprocessing. Dimension reduction is one of the principal requirements for a successful representation to improve the efficiency of extracting the attracting trend patterns on the time series data. Symbolic representations have proven to be a very effective way to reduce the dimensionality of time series without loss of knowledge. However, symbolic representations suffer from another challenges promoted by the possibility of losing some principal patterns due to the impractical utilization of dealing with the whole data with the same weight. The main objective of this study is to integrate Symbolic Aggregate Approximation and recent biased techniques in searching a pattern similarity in the stock market data. Moreover, the data dimensionality is reduced by keeping more detail on recent-pattern data and less detail on older ones using modified sliding window controlled by the corresponding classification error rate. Experimental results were made on the UCR standard dataset comparing with the state of the art techniques such as Discrete Fourier Transform (DFT) and Dynamic Time Warping (DTW). The proposed techniques showed promising results. Furthermore, practical experiments were made on the Egyptian stock market indices EGX 30, EGX 70 and EGX 100. The discovered patterns were reviewed by professional financial experts and showed the accuracy and effectiveness of the proposed approach.

To Cite This Article: Tamer Hassan Abd El Salam, Assoc. Prof. Dr. Zalinda Othman, Prof. Dr. Abdul Razak Hamdan, Optimizing Similarity Search Using Symbolic Recently-Biased Methodology. ***Aust. J. Basic & Appl. Sci.,*** *8(6): 57-67, 2014*

INTRODUCTION

A time series is a sequence of time stamped data points. Huge time series data sets are very common in today's scientific and financial databases such as stock market applications. Similarity search is a useful tool for exploring time series databases to look for a particular pattern within a longer sequence. Efficient and accurate similarity searching for a large amount of time series data sets is nontrivial problem. However, Researches are oriented to the data compression ignoring the wastage information.

Typically, these data compression processes are only intended to reduce dimensionality on data (Han and Kamber, 2006). The problem is that they do not consider whether the new representation preserves the relevant information or not. The proper interval must be chosen carefully or this may lead to the loss of important knowledge and to hide the patterns. Consequently, if the length of the intervals is very large, some of the details that describe the data may be lost leading to miss discover the patterns from the database and if the length of each interval is too small then it will not have enough data to produce patterns. Many dimensionality reduction techniques have been proposed for effective representation of time series data.

Time series data are being generated at an unprecedented speed from daily fluctuations of stock market. As a consequence, in the last decade there has been a great interest in querying and mining such data which, in turn, resulted in a large number of works introducing new methodologies for indexing, clustering, classification and approximation of time series (Han and Kamber, 2006; Keogh, 2006). The two key aspects for achieving effectiveness when managing time series data are representation methods and similarity measures. Time series are essentially high dimensional data (Han and Kamber, 2006). Dealing with such data in its raw format is very expensive in terms of processing and storage cost. It is thus highly desirable to develop representation techniques that can reduce the dimensionality of time series, while still preserving the fundamental characteristics of the data set. Moreover, the distance between time series needs to be carefully defined in order to reflect the underlying similarity of such data. This is particularly desirable for similarity-based retrieval, classification, clustering and the other mining procedures of time series (Han and Kamber, 2006).

Corresponding Author: Tamer Hassan Abd El Salam, Phone numbers (with country and area code) 00201001717654
E-mail: tamerhassan81@hotmail.com

Similarity search task over time series data in the stock market has a great interest as it can predict the drop in the stock exchange before its happening as well as it produces a great help for investors to take the right buy or sell decision. Motivated by these observations, the state of the art representation methods and similarity measures for time series that appeared in high quality conferences and journals have been evaluated. Specifically, the proposed representation methodology for time series have been compared with various time series data sets.

Symbolic Aggregate Approximation (SAX) was proposed as a new method for time series data representation (Lin *et al*., 2003). SAX discretizes the original time sequences into symbolic strings and distance measures to be defined on the symbolic approach. However, SAX is based on the Piecewise Aggregate Approximation (PAA) representation that minimizes dimensionality by the mean values of equally sized frames (Keogh *et al*., 2001).

To improve the quality of SAX based similarity search, in this paper, a new approach is proposed which supplements the SAX based pattern matching with a recent biased technology. The proposed similarity search approach consists of the following three steps: 1) dimensionality reduction via PAA applying data segmentation using recent biased technology, 2) transforming real valued time series into symbolic representation using SAX and 3) pattern matching on the symbolic strings.

Generally, a time series reflects the behavior of the data points (monitored event), which tends to repeat periodically and creates a pattern that alters over time due to countless factors. Hence the data that contains the recent pattern are more significant than just recent data and even more significant than older data. This change of behavioral pattern provides the key to the proposed methodology in dimension reduction. Since the pattern changes over time, the most recent pattern is more significant than older ones (Phithakkitnukoon and Ratti, 2010). In this paper, a new recent pattern biased dimension reduction technique is introduced that gives more significance to the recent data by keeping it with higher resolution, while older data is kept at lower resolution.

Applying recent biased methodology the traditional dimension reduction techniques such as Single Value Decomposition SVD, Discrete Fourier Transformation DFT, Discrete Wavelet Transformation DWT, Landmarks, PAA and Adaptive Piecewise Constant Approximation APCA can be used.

This paper is distinguished from other previously proposed similarity search techniques by the following contributions:

1) A new dimension reduction is developed by keeping more detail on the data that contains the most recent pattern and less detail on the older ones.

2) Recent periodicity detection and recent pattern interval detection have been proposed applying SAX compared with the state of the art similarity search techniques.

The proposed work treats with the dimension reduction gaps in the other articles and journals applying the state of the art techniques and taking the order of the data into consideration. In a recent paper applying SAX (Fuad, 2012), there was a fine representation using the state of the art technique but the gap was the ignorance of the recent data effectiveness by dealing with the new and old data with the same weight. Another gap exists in the recent biased time series technique exist in the conference (Muruga *et al*., 2010) which is the usage of old dimension reduction technique (DWT) and ignoring the state of the art techniques such as symbolic representation that have not even being considered in the experimental comparisons.

The rest of this paper is organized as follows. Section 2 reviews the concept of time series, and gives an overview of the definitions of the different representation techniques and similarity measures investigated in this work. Section 3 and Section 4 present the main contribution of this work – the results of the extensive experimental evaluations of the different representation methods and similarity measures, respectively. Section 5 explores the conclusion of the paper and discusses possible future extensions of the work.

Background And Related Work:

Time series classification techniques can be divided in two main categories (Ding *et al*., 2008). The first category includes techniques based on shape-based similarity metrics where distance is measured directly between time series points. The principal classical example from this category is 1-NN classifier built upon Euclidean distance (xi *et al*., 2006) and DTW (Sakoe and Chiba, 1978). The second category consists of classification techniques based on structural similarity metrics, which employ a high-level representation of time series based on their global or local features such as classifiers based on time series representation obtained with DFT (Agrawal *et al*., 1993) or Bag-Of-Patterns (Lin *et al*., 2012). The development of these distinct categories can be explained by the difference in their performance: while shape-based similarity methods are virtually unbeatable on short pre-processed time series (Keogh and Kasetty, 2003), they often fail on long and noisy data, where structure-based solutions show a superior performance (Lin *et al*., 2012). Much of the world's supply of data is in the form of time series data. In the last decade, there has been an explosion of interest in mining time series data (Lin *et al*., 2007). Searching directly on these data will be very complex and inefficient. To overcome this problem, some of transformation methods should be applied to reduce the magnitude of time series databases.

Many techniques have been proposed in the literature for representing time series with reduced dimensionality, such as Discrete Fourier Transformation (Faloutsos *et al*., 1994), Single Value Decomposition (SVD) (Faloutsos *et al*., 1994), Discrete Wavelet Transformation (DWT) (Chan and Fu, 1999), Piecewise Aggregate Approximation (PAA) (Keogh *et al*. 2001), Adaptive Piecewise Constant Approximation (APCA) (Keogh *et al*. 2001), Chebyshev polynomials (CHEB) (Cai and Ng, 2004), Symbolic Aggregate approXimation (SAX) (Lin *et al*., 2007) and many other techniques. Moreover, there are over a dozen distance measures for similarity of time series data in the literature such as Euclidean distance (ED) (Faloutsos *et al*., 1994) , Dynamic Time Warping (Berndt and Clifford, 1994; Keogh *et al*., 2005), distance based on Longest Common Subsequence (LCSS) (Vlachos *et al*., 2002),Edit Distance with Real Penalty (ERP) (Chen *et al*., 2004) , Edit Distance on Real sequence (EDR) (Chen *et al*., 2005), Sequence Weighted Alignment model (Swale) (Morse and Patel, 2007), Spatial Assembling Distance (SpADe) (Chen *et al*., 2007) and similarity search based on Threshold Queries (TQuEST) (Assfalg *et al*., 2006). Many of these works and some of their extensions have been widely cited in the literature and applied to facilitate query processing and data mining of time series data as can be shown in Table 1.

Table 1: Comparison of data representation techniques

Techniques	*Authors, year*	*Advantages*	*Disadvantages /Restrictions*
DFT	Agrawal *et al*. 1993	Can convert any complex time series into terms of sine/cosine waves with high compression ratio.	"wavelets outperform the DFT" (Popivanov and Miller 2002) and the user is required to input several parameters, including the size of the alphabet.
DWT	Chan and Fu 1999	Fast to use with little storage and allows good approximation with a subset of coefficients.	DFT filtering performance is superior to DWT (Kawagoe and Ueda 2002) and Show poor performance for certain locally distributed time series data.
PAA	Keogh 2005	Surprisingly competitive with the more sophisticated transform and can apply twice as many approximating segments.	The break points parameters depend on the user assumption.
SAX	Lin *et al*. 2007	Offers effective methods that are applicable in many branches of artificial intelligence and support distance measure transform to a symbolic string.	The main drawback of SAX methodology is that it depends on applying PAA model assumptions.

However, with the multitude of competitive techniques, there is a strong need to compare what might have been omitted in the comparisons. Every newly introduced representation method or distance measure has claimed a particular superiority. However, it has been demonstrated that some empirical evaluations have been inadequate (Keogh and Kasetty, 2003) and, worse yet, some of the claims are even contradictory.

For example, one paper claims “wavelets outperform the DFT” (Popivanov and Miller, 2002), another claims “DFT filtering performance is superior to DWT” (Kawagoe and Ueda, 2002) and yet another claims “DFT-based and DWT-based techniques yield comparable results” (Wu *et al*., 2000). Surely these claims cannot all be true. The risk is that this may not only confuse newcomers and practitioners of the field, but also cause a waste of time and work efforts due to assumptions based on incomplete or incorrect claims. Two goals are clearly shown for these representations:

- Dimensionality reduction. The representations should preserve the underlying information as much as possible. Usually, such representations allow reconstructing time series as close as possible to the original time series (according to certain distance measure). The difference between those representations and the one used for compression is the need to enable some operations directly on those representation methods (for example computing distance measures).
- Information extraction. The representations should make the underlying information explicit. Those high-level representations must exhibit the relevant information. This is a step to pattern matching and symbolic process-monitoring (Love and Simaan, 1988) where episodes of interest are detected and labeled according to a predefined set of shapes.

The current symbolic approach most commonly used is the SAX (Lin *et al*., 2007; Pham *et al*., 2010; Rakthanmanon and Keogh, 2013; Keogh *et al*., 2005). SAX was the first symbolic representation that applies dimensionality reduction technique as a preprocessing step using PAA algorithm (Lin *et al*., 2007) to minimize the noise effect. However, the SAX approach causes a high possibility to miss important patterns in time series data, such as the local trend of the time series (Lin *et al*., 2007). Furthermore, the Gaussian assumption of the symbols distribution has effects on the SAX performance for non-uniform or correlated time series (Pham *et al*., 2010). Symbolic Aggregate Approximation was proposed as an effective current method for time series data representation (Lin *et al*. 2003) which enables information extraction of time series representations, where episodes are associated to symbols interpretable according to the data-mining task.

Their representation was also unique in that it allows dimensionality reduction as well as distance measures to be defined on the symbolic approach where the lower bound corresponding distance measures is defined on the original series. This method is based on PAA (John *et al*., 2001). The proposed solution utilizes PAA to achieve dimensional reduction via the proposed recent biased methodology and then transform the reduced real valued time series into a symbolic representation called RBSAX.

Concerning the global dimension reduction techniques, in many applications such as stock market prices, recent data are much more interesting and significant than old data (Phithakkitnukoon and Ratti, 2010). Thus, the dimension reduction techniques that emphasize more on the recent data by keeping recent data with high resolution and old data with low resolution have been proposed such as Tilt time frame, Logarithmic tilted-time window, Pyramidal time frame, Stream Summarization using Wavelet-based Approximation Tree SWAT (Bulut and Singh, 2003) and many other techniques. Tilt time frame has been introduced by Chen *et al*. (Chen *et al*., 2002) to minimize the amount of data to be kept in the memory or stored on the disks. In the tilt time frame, time is registered at different levels of granularity. The most recent time is registered at higher granularity, while the more distant time is registered at the lower granularity. The downgrade level depends on the application requirements. Similar to the tilt time frame concept but with more space-efficient, Giannella *et al*. (2003) have proposed the logarithmic tilted-time window model (Giannella *et al*., 2003) that partitions the time series into growing tilted-time window frames at an exponential rate of two. The concept of the pyramidal time frame has been introduced by Aggarwal *et al*. in 2003 in which data are stored at different levels of granularity depending upon the date, which follows a pyramidal pattern. SWAT has been proposed by Bulut and Singh (2003) to process queries over data streams that are biased towards the more recent values. SWAT is a Haar wavelet-based scheme that keeps only a single coefficient at each level. Zhao and Zhang have proposed the equi-segmented scheme and the vari-segmented scheme in (Zhao and Zhang, 2006). The idea of the equi-segmented scheme is to divide the time series into equal-length segments and apply a dimension reduction technique to each segment keeping more coefficients for the recent data while fewer coefficients are kept for the old data.

Methodology:

In this work, PAA is applied to achieve dimensional reduction and then transform the reduced real valued time series into a symbolic representation via the recently proposed Symbolic Aggregate Approximation algorithm. SAX allows a time series of arbitrary length n to be reduced to a string of arbitrary length w, (w < n, typically w << n). The alphabet size is also an arbitrary integer a, where a > 2. Table 2 summarizes the major notation used in this and subsequent sections.

Table 2: A summarization of the notation used

C	A time series C = $c_{1,\dots\dots,}c_n$
$\overline{C}$	A Piecewise Aggregate Approximation of a time series C = $\bar{c}_{1,\dots\dots,}\bar{c}_w$
$\hat{C}$	A symbol representation of a time series $\widehat{C_i} = \widehat{c_1}\ ,\dots\dots,\widehat{c_w}$
w	The number of PAA segments representing time series C
a	Alphabet size

Discretization procedure uses an intermediate representation between the raw time series and the symbolic strings. In the first step the data is transformed into the Piecewise Aggregate Approximation representation and then the PAA representation is symbolized into a discrete string. There are two important advantages which can be summarized in the following:

Dimensionality Reduction:

The well-defined and well-documented dimensionality reduction power of PAA (Keogh, 2006; (Zhao and Zhang, 2006) can be used and the reduction is automatically carried over to the symbolic representation.

Lower Bounding:

The key observation that allows us to prove lower bounds is to concentrate on proving that the symbolic distance measure lower bounds the PAA distance measure. Then the desired result can be deduced by simply pointing to the existing proofs for the PAA representation itself (Zhao and Zhang, 2006).

The PAA dimensionality reduction is intuitive and simple, yet has been shown to rival more sophisticated dimensionality reduction techniques like Fourier transforms and wavelets (Keogh, 2006; Keogh *et al.*, 2001; Zhao and Zhang, 2006). The framework of this paper is shown in Figure 1 exploring the main contribution of this paper.

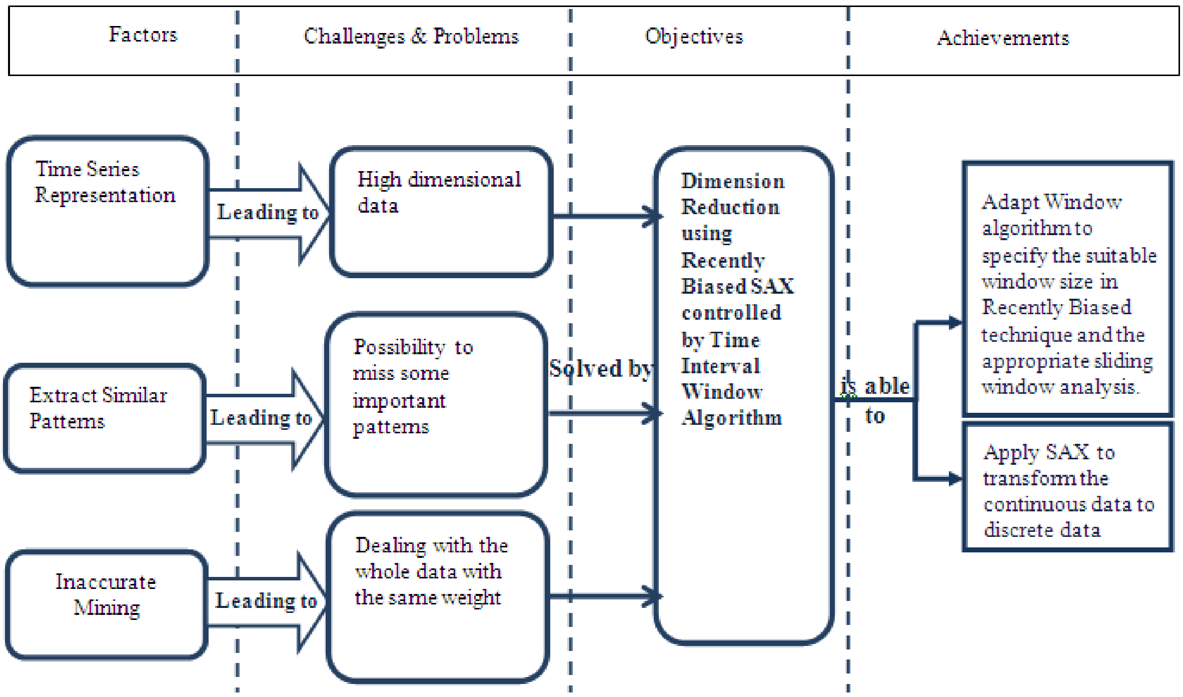

Factors
Challenges & Problems
Objectives
Achievements
Time Series Representation
Leading to
High dimensional data
Extract Similar Patterns
Leading to
Possibility to miss some important patterns
Inaccurate Mining
Leading to
Dealing with the whole data with the same weight
Solved by
Dimension Reduction using Recently Biased SAX controlled by Time Interval Window Algorithm
is able to
Adapt Window algorithm to specify the suitable window size in Recently Biased technique and the appropriate sliding window analysis.
Apply SAX to transform the continuous data to discrete data

Fig. 1: Study framework

Appropriate representation of time series is critical for various mining and learning tasks. Due to the inherent high dimensionality of time series data, dimensionality reduction is an important issue for time-series representation. High dimensional time series data is common in stock market exchange data. The direct use of the raw time series data can often lead to wrong results. The time series consisted of 10 data points can be divided into 5 segments. If the segments are equal it is obvious that point 1, 2 should be in segment 1, point 3, 4 should be in segment 2 and so on. Detection has been studied for a long time in statistics literature where signal with a series of best fitting lines return the end points of the segments as a change point known as window or pattern (Epifani *et al*., 2010). Changing the window size of the data points using the recently biased technique combined with the modification of the segment window size, the data points could be divided into 4 points putting point 1, 2 in segment 1, point 3, 4 in segment 2, point 5, 6 in segment 3 and putting the points 7, 8, 9 and 10 in segment 4 keeping more detail in the recent data points and less details in the old data points. Moreover the segment window size can be changed to get the most effective results. A time series C of length n can be represented in a w-dimensional space by a vector

$\overline{C} = \overline{C}_{1,\dots,}\overline{C}_{w}$, the i th element of $\overline{C}$ is calculated by the following equation:

$$\overline{C}_i = \frac{w}{n} \sum_{j=(i-1)\frac{n}{w}+1}^{i\,n/w} C_j \qquad (1)$$

To reduce the time series from n dimensions to w dimensions, the data is divided into w equally sized “frames.” The mean value of the data falling within a frame is calculated and a vector of these values becomes the data-reduced representation. PAA can be visualized as an attempt to approximate the original time series data with a linear combination of box basis functions as shown in Figure 2. Applying PAA on a specified time series can be visualized as an attempt to model a time series with a linear combination of box basis function. In this example in Figure 2 a sequence of length 128 is reduced to 8 dimensions.

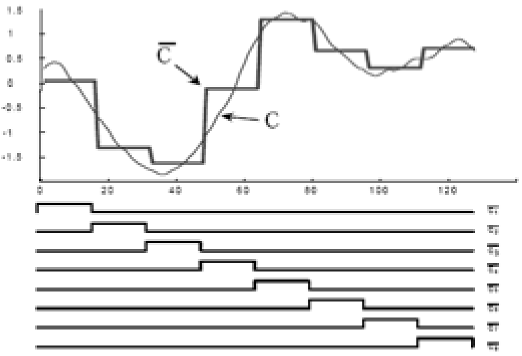
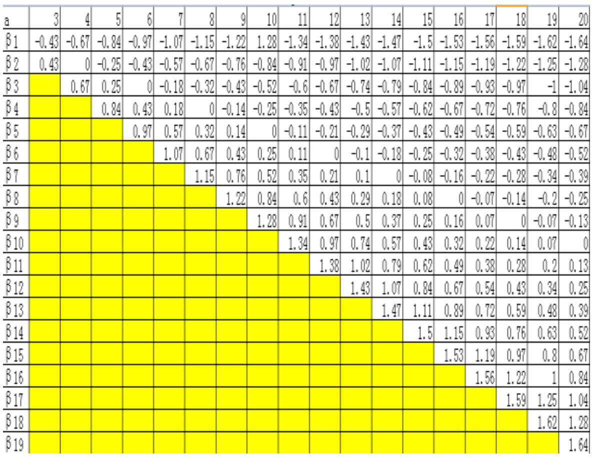


Fig. 2: The PAA representation of Time Series

It is required to have a discretization technique that can produce symbols with equal probability (Chen *et al.*, 2007; Aggarwal *et al.*, 2003). This is easily achieved since a normalized time series have a Gaussian distribution (Giannella *et al.*, 2003). The breakpoints can be simply determined which will produce an equal sized area under the Gaussian curve (Giannella *et al.*, 2003).These breakpoints may be determined by looking them up in a statistical table. For example, Table 3 gives the breakpoints for the values of 'a' from 3 to 20. Each time series is normalized to have a mean of zero and a standard deviation of one before converting it to the PAA representation, since it is well understood that it is meaningless to compare time series with different offsets and amplitudes (Keogh *et al.*, 2001).

Table 3: A lookup table that contains the breakpoints that divide a Gaussian distribution in an arbitrary number (from 3 to 20) of equal probability regions



a	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
β1	-0.43	-0.67	-0.84	-0.97	-1.07	-1.15	-1.22	1.28	-1.34	-1.38	-1.43	-1.47	-1.5	-1.53	-1.56	-1.59	-1.62	-1.64
β2	0.43	0	-0.25	-0.43	-0.57	-0.67	-0.76	-0.84	-0.91	-0.97	-1.02	-1.07	-1.11	-1.15	-1.19	-1.22	-1.25	-1.28
β3		0.67	0.25	0	-0.18	-0.32	-0.43	-0.52	-0.6	-0.67	-0.74	-0.79	-0.84	-0.89	-0.93	-0.97	-1	-1.04
β4			0.84	0.43	0.18	0	-0.14	-0.25	-0.35	-0.43	-0.5	-0.57	-0.62	-0.67	-0.72	-0.76	-0.8	-0.84
β5				0.97	0.57	0.32	0.14	0	-0.11	-0.21	-0.29	-0.37	-0.43	-0.49	-0.54	-0.59	-0.63	-0.67
β6					1.07	0.67	0.43	0.25	0.11	0	-0.1	-0.18	-0.25	-0.32	-0.38	-0.43	-0.48	-0.52
β7						1.15	0.76	0.52	0.35	0.21	0.1	0	-0.08	-0.16	-0.22	-0.28	-0.34	-0.39
β8							1.22	0.84	0.6	0.43	0.29	0.18	0.08	0	-0.07	-0.14	-0.2	-0.25
β9								1.28	0.91	0.67	0.5	0.37	0.25	0.16	0.07	0	-0.07	-0.13
β10									1.34	0.97	0.74	0.57	0.43	0.32	0.22	0.14	0.07	0
β11										1.38	1.02	0.79	0.62	0.49	0.38	0.28	0.2	0.13
β12											1.43	1.07	0.84	0.67	0.54	0.43	0.34	0.25
β13												1.47	1.11	0.89	0.72	0.59	0.48	0.39
β14													1.5	1.15	0.93	0.76	0.63	0.52
β15														1.53	1.19	0.97	0.8	0.67
β16															1.56	1.22	1	0.84
β17																1.59	1.25	1.04
β18																	1.62	1.28
β19																		1.64

Having transformed a time series database into PAA, a further transformation can be applied to obtain a discrete representation which means that $alpha_1$ = a and $alpha_2$ = b then the mapping from PAA approximation $\bar{C}$ to a word $\hat{C}$ is obtained as the following equation:

$$\hat{C}_i = alpha_j, \text{ iif } \sum j-1 \leq \bar{C}_i < \sum j \qquad (2)$$

The sequence C of length n can be represented as a word

$$\bar{C} = \bar{C}_1, \ldots\ldots \bar{C}_w$$

Once the breakpoints have been obtained, the time series can be discretized in the following manner. After obtaining the Piecewise Aggregate Approximation of the time series, 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” and so on. Figure 3 illustrates the idea. The time series database is transformed into PAA and then a further transformation can be applied to obtain a discrete representation.

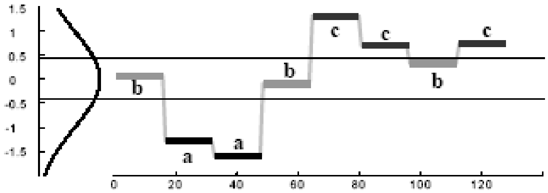


Fig. 3: A time series is discretized by first obtaining a PAA approximation and then using predetermined breakpoints to map the PAA coefficients into SAX symbols. With n = 128, w = 8 and a = 3, the time series is mapped to the word "baabccbc**".**

Lin *et al*. (2003) proposed a method called SAX which is based on PAA and assumes normality of the resulting aggregated values. SAX is a process which maps the PAA representation of the time series into a sequence of discrete symbols. SAX is a symbolization method that involves placing a symbol for each segment obtained by using PAA. In order to do that, it is essential to specify the number of symbols and the interval of the values for each symbol. The number of symbols to be used is generally determined by an expert having knowledge about the studied domain. Using too many symbols will cause to end up with a string keeping too much of the original data and will not simplify the series; on the other hand, too less symbols will cause considerable amount of information loss. After choosing the number of symbols, some histograms of the data values can be helpful to specify the intervals for each symbol. The original Egyptian index can be seen in Figure 4 before applying the dimension reduction technique. In Figure 5, the PAA normalization of the Egyptian index can be shown.

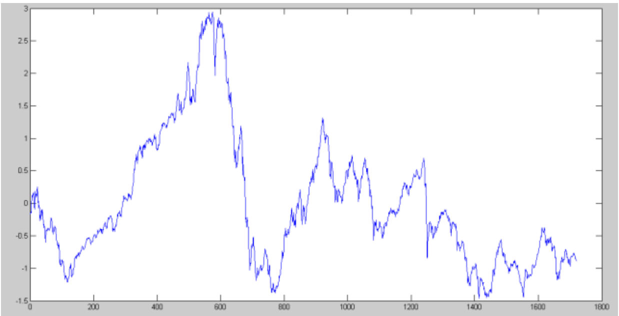


Fig. 4: EGX 100 before normalization

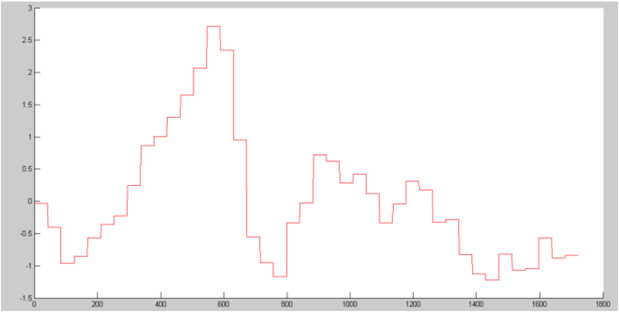


Fig. 5: EGX 100 after PAA normalization

Finally, after applying SAX the resulted data can be shown in Figure 6.

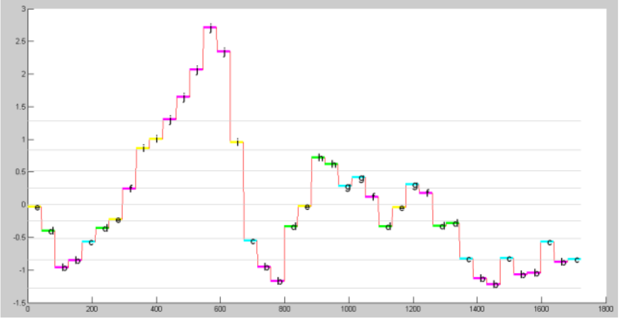


Fig. 6: EGX 100 after SAX Discretization

Applying the Symbolic representation on one of the stock market indices EGX 100 is to discretize the data and discover the similar patterns by first obtaining a PAA approximation and then using predetermined breakpoints to map the PAA coefficients into SAX symbols. The symbolic representation is applied on the Egyptian index in Figure 5 with n = 2022, number of segments = 41 and alpha size = 10, the time series is mapped to the word edbbcdefiijjjjjicbbdehhggfdegfddcbbcbbcbc.

The Egyptian stock market indices data is divided into two segments. The position of the two segments is positioned and controlled by using a sliding window to specify the most important data information in the stock market index data applying the recent biased methodology which gives an advantage to the recent data with much more interesting and significant information than the old data (Phithakkitnukoon and Ratti, 2010).

The sliding window is modified beginning from the most recent data towards the direction of the older data while evaluating the classification error rate during the window adjustment until an appropriate error below certain threshold is obtained. Furthermore, each segment is divided into sub-segments in which the number of sub-segments inside each segment depends on the recentness of the data inside the segments. In this work the data was divided into two segments controlling the size and the number of the sub-segments of the first recent segment using a sliding window size corresponding to the least classification error obtained. Similarly, the number of the sub-segments in the older data segment is also obtained using the controlled window size which yields the appropriate classification accuracy.

Experimental Results:

The recent biased technology combined with the symbolic representation was first tested using a UCR standard data set comparing with the state of the art techniques such as DFT and DTW. The recent biased symbolic representation showed promising results in many data sets as shown in Table 4.

Table 4: Experimental Results of the Standard UCR Data set comparing classification error using RBSAX methodology with the state of the art techniques

Data Set	Classes	ED	DTW	DFT	SAX	RBSAX
50words	50	0.407	0.375	0.521	0.341	0.431
Adiac	37	0.464	0.465	0.476	0.89	0.388
Beef	5	0.467	0.433	0.493	0.567	0.337
Car	4	0.275	0.333	0.316	0.333	0.297
CBF	3	0.087	0.003	0.16	0.104	0.002
chlorineconcentration	3	0.349	0.38	0.415	0.41	0.474
cinc_ECG_toeso	4	0.051	0.165	0.087	0.195	0.326
Coffee	2	0.193	0.191	0.249	0.464	0.337
diatomsizereduction	4	0.022	0.015	0.13	0.152	0.08
ECG200	2	0.162	0.221	0.283	0.12	0.264
ECGFiveDays	2	0.118	0.154	0.21	0.263	0.244
FaceFour	4	0.149	0.064	0.172	0.17	0.169
Faces(all)	14	0.225	0.192	0.216	0.33	0.285
fish	7	0.319	0.329	0.493	0.474	0.308
Gun Point	2	0.146	0.14	0.281	0.18	0.23
Lighting2	2	0.341	0.204	0.338	0.213	0.211
Lighting7	7	0.377	0.252	0.363	0.397	0.384
OliveOil	4	0.15	0.133	0.623	0.833	0.115
OSULeaf	6	0.448	0.401	0.535	0.467	0.394
plane	7	0.051	0.001	0.112	0.038	0.033
SwedishLeaf	15	0.295	0.256	0.319	0.483	0.395
synthetic control	6	0.142	0.019	0.134	0.02	0.011
Trace	4	0.368	0.016	0.427	0.46	0.007
TwoPatterns	4	0.095	0	0.163	0.081	0.84
wafer	2	0.005	0.015	0.09	0.004	0.08
yoga	2	0.16	0.151	0.184	0.195	0.143

Practical experiments were made on the Egyptian stock market indices EGX 30, EGX 70 and EGX 100.The data was divided in this work after applying the recent biased technology according to the controlled sliding window corresponding to the appropriate classification error rate by keeping 16% of the data beginning from the recent data direction inside the first segment and split the first segment into 33 sub-segments. Moreover, the data of the older data segment is kept in the 84% of the data splitting the second segment into 8 segments. The integration between recent biased methodology and symbolic representation showed a promising results as well as shown in Table 5.

Table 5: Experimental Results of the Egyptian Stock Market Indices classification error using RBSAX methodology compared with the state of the art techniques

Index	Number of instances	ED error	DTW	DFT	SAX error	RBSAX error
EGX 30	2375	49.78	43.22	44.23	39.54	29.48
EGX 70	1232	47.15	41.93	42.84	40.28	39.27
EGX 100	1724	48.83	45.71	47.49	38.13	37.82

Discussion:

Comprehensiveness is critical for several applications of classification, because without interpretable features, domain experts may often reluctant to adopt a classification approach in the decision making process (Xing *et al.*, 2011; Batyrshin and Sheremetov, 2008). In order to produce a meaningful model a Symbolic Aggregate Approximation dimension reduction technique is applied with the recent biased weight methodology. Other works in the literature have used the instance based classification, such as the nearest neighbor algorithm, and have presented accurate results on a great variety of time series data sets.In this context, a new symbolic representation method is proposed to take into account the weight of the time series data. The standard UCR data sets and the Egyptian stock market indices is descretized according to the proper classification accuracy accomplished applying a proper sliding window beginning with 1% of the data increased by 1% repeatedly beginning from the most recent data directed towards the older ones. The statistical evaluation of recent biased SAX indicates a significant difference between this work and other classic state of the art techniques. According to the experimental results it can be explored that this work has a superior performance of dimension reduction technique for most data sets. In the accuracy analysis there was significant difference for comparisons between applying recent biased technology comparing with the classic SAX and other state of the art techniques such that DFT, DTW, where the classic SAX was better in many data sets this technique called RBSAX presented in this work was better in significant number of data sets. On the other hand, the results presented in Table 4.1 demonstrate the ability of recent biased symbolic representation to classify the time data whether having superior accuracy or not. Applying the new methodology RBSAX on the standard UCR datasets improve the classification error rate in 9 datasets from 1% until 28% compared with the other classification techniques. Moreover comparing the results in the Table 4.2 it can be easily observed that RBSAX fulfill a significant percentage of improvement. The classification error rate reduced from 1% until 10% in the Egyptian stock market indices.

The proposed approach was tested on Egyptian stock market time series data sets with the three indices EGX 30, EGX 70 and EGX 100 using the Matlab tool. Analyzing the stock market data corresponding to the discovered patterns it was found that the pattern is repeatedly significantly raised in certain intervals as a direct effect of declaring any kind of elections which may means stability until the election ends as happened in the intervals around the election dates in 19th of March 2011,26th of December 2011 and 20th of June 2012.As an expectation of the discovered pattern it is also expected for the stock market to be significantly up during the coming election intervals beginning in 15th of January 2014. On the other side it was detected that a repeated fall pattern is invoked as a result of any kind of significant demonstration such as happened in the intervals around the dates of 11th of January 2011, 9th of June 2011 and 28th of November 2012. Consequently, it is expected for the stock market to go down in the interval around advertised demonstration intervals. The discovered patterns were reviewed by professional financial experts and were expected to be accurate with an accuracy rate exceeds 70%.

REFERENCES

Aggarwal, C., J. Han, J. Wang and P. Yu, 2003. ”A Framework for Clustering Evolving Data Streams,” *Procs. 29th Very Large Data Bases Conference (VLDB’03)*, pp: 81-92.

Agrawal, R., C. Faloutsos and A. Swami, 1993. Efficient Similarity Search in Sequence Data bases. *International Conference on Foundations of Data Organization (FODO).*

Assfalg, J., H.-P. Kriegel, P. Kr¨oger, P. Kunath, A. Pryakhin and M. Renz, 2006. Similarity search on time series based on threshold queries. *In EDBT.*

Batyrshin, I.Z. and L. Sheremetov, 2008. “Perception-based approach to time series dat mining,” Applied Soft Computing, 8(3): 1211-1221.

Berndt, D. and J. Clifford, 1994. Using Dynamic Time Warping to Find Patterns in Time Series. *In Proc. AAAI Workshop on Knowledge Discovery in Databases*.

Bulut, A. and A.K. Singh, 2003. ”SWAT: Hierarchical Stream Summarization in Large Networks,” *Procs. 19th Int’l Conf. Data Eng. (ICDE’03).*

Cai, Y. and R.T. Ng, 2004. Indexing spatio-temporal trajectories with chebyshev polynomials. *In SIGMOD Conference*.

Chan, K. and A.W. Fu, 1999. Efficient time series matching by wavelets. Proceedings of 15th IEEE International Conference on Data Engineering; Sydney, Australia, pp: 126-133.

Chen, L., M.T. Ozsu and V. Oria, 2005. Robust and fast similarity search for moving object trajectories. *In SIGMOD Conference*.

Chen, L. and R.T. Ng, 2004. On the marriage of lp-norms and edit distance. *In VLDB*.

Chen, Y., G. Dong, j. Han, B.W. Wah and J. Wang, 2002. "Multi-Dimensional Regression Analysis of Time Series Data Streams," *Procs. 2002 Int'l Conf. Very Large Data Bases (VLDB'02).*

Chen, Y., M.A. Nascimento, B.C. Ooi and A.K.H. Tung, 2007. SpADe: On Shape-based Pattern Detection in Streaming Time Series. *In ICDE*.

Ding, H., G. Trajcevski, P. Scheuermann, X. Wang and E. Keogh, 2008. Querying and mining of time series data: experimental comparison of representations and distance measures. In Proc. VLDB, 1542–1552.

Epifani, I., C. Ghezzi and G. Tamburrelli, 2010. Change-point detection for black-box services. *In International symposium on Foundations of Software engineering, (FSE),USA*, pp: 227-236.

Faloutsos, C., M. Ranganathan and Y. Manolopoulos, 1994. Fast subsequence matching in time-series databases. *In proceedings of the ACM SIGMOD Int'l Conference on Management of Data*; 25-27; Minneapolis, MN, pp: 419-429.

Fuad, M.M.M., 2012. "Genetic Algorithms-Based Symbolic Aggregate Approximation",*in Proc. DaWaK*, pp: 105-116.

Giannella, C., J. Han, J. Pei, X. Yan and P.S. Yu, 2003. "Mining Frequent Patterns in Data Streams at Multiple Time Granularities," *Data Mining: Next Generation Challenges and Future Directions, H. Kargupta, A. Joshi, K. Sivakumar, and Y. Yesha, eds., AAAI/ MIT Press*.

Han, J. and M. Kamber, 2006. Data Mining: Concepts and Techniques. *2nd edition. Morgan Kaufmann*.

John, F.R., H. Kathleen and S. Myra, 2001. An Updated Bibliography of Temporal, Spatial, and Spatio-temporal Data Mining Research. *In: Roddick, J., Hornsby, K.S. (eds.) TSDM 2000.LNCS (LNAI), vol. 2007, pp. 147–163. Springer, Heidelberg*.

Kawagoe, K. and T. Ueda, 2002. A Similarity Search Method of Time Series Data with Combination of Fourier and Wavelet Transforms. *In TIME*.

Keogh, E., K. Chakrabarti, M. Pazzani and S. Mehrotra, 2001. Dimensionality reduction for fast similarity search in large time series databases. Journal of Knowledge and Information Systems, 3(3): 263-286.

Keogh, E.J., 2006. A Decade of Progress in Indexing and Mining Large Time Series Databases. *In VLDB*.

Keogh, E.J. and S. Kasetty, 2003. On the Need for Time Series Data Mining Benchmarks: A Survey and Empirical Demonstration. *Data Min. Knowl. Discov.,7(4).*

Keogh, E., J. Lin and A. Fu, 2005. "Hot sax: Efficiently finding the most unusual time series subsequence," *in Data Mining, Fifth IEEE International Conference on. IEEE, pp. 8–pp*.

Lin, J., E. Keogh, S. Lonardi and B. Chiu, 2003. "A Symbolic Representation of Time Series, with Implications for Streaming Algorithms", *8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, San Diego, California,* pp: 2-11.

Lin, J., E.J. Keogh, L. Wei. and S. Lonardi, 2007. "Experiencing sax: a novel symbolic representation of time series," Data Min. Knowl. Discov., 15(2): 107-144.

Lin, J., R. Khade and Y. Li, 2012. Rotation-invariant similarity in time series using bag-of-patterns representation. J. Intell. *Inf. Syst.* 39(2): 287-315.

Love, P.L. and M. Simaan, 1988. Automatic recognition of primitive changes in manufacturing process signals. Pattern Recognition, 21(4): 333-342.

Morse, M.D. and J.M. Patel, 2007. An efficient and accurate method for evaluating time series similarity. In SIGMOD Conference.

Muruga, D. Radha Devi, V. Maheswari and P. Thambidur, 2010. "Similarity Search In Recent Biased Time Series Databases Using Vari-DWT and Polar Wavelets", pp: 398-404.

Pham, N.D., Q.L. Le and T.K. Dang, 2010.Two novel adaptive symbolic representations for similarity search in time series databases,*in APWeb*, pp: 181-187.

Phithakkitnukoon, S. and C. Ratti, 2010. A recent-pattern biased dimension-reduction framework for time series data. Journal of Advances in Information Technology, 1(4): 168-180.

Popivanov, I. and R.J. Miller, 2002. Similarity search over time series data using wavelets. *In proceedings of the 18th Int'l Conference on Data Engineering; Feb 26-Mar 1; San Jose, CA*, pp: 212-221.

Rakthanmanon, T. and E. Keogh, 2013. Fast shapelets: A scalable algorithm for discovering time series shapelets," *in SDM*.

Sakoe, H. and S. Chiba, 1978. Dynamic Programming Algorithm Optimization for Spoken Word Recognition. IEEE Trans. Acoustics, Speech and Signal Processing,ASSP-26(1).

Xi, X., E. Keogh, C. Shelton, L. Wei and C. Ratanamahatana, 2006. Fast time series classification using numerosity reduction. *In Proc. ICML*.

Xing, Z., J. Pei, P.S. Yu, and K. Wang, 2011. “Extracting interpretable features for early classification on time series,” *in SDM*, pp: 247-258.

Zhao, Y. and S. Zhang, 2006. ”Generalized Dimension-Reduction Framework for Recent-Biased Time Series Analysis,” IEEE Trans. *on Knowledge and Data Eng.*, 18(2): 231-244.

Wu, Y., D. Agrawal and A. El Abbadi, 2000. A comparison of DFT and DWT based similarity search in time-series databases*. In proceedings of the 9th ACM CIKM Int’l Conference on Information and Knowledge Management; McLean*, VA, pp: 488-495.