International Journal of Basic and Applied Sciences, 3 (2) (2014) 101-109
©Science Publishing Corporation
www.sciencepubco.com/index.php/IJBAS
doi: 10.14419/ijbas.v3i2.2358
Research paper

Similarity search optimization using recently-biased symbolic representation

Tamer Hassan Abd El Salam*, Zalinda Othman, Abdul Razak Hamdan

FTSM-UKM-Malaysia
**Corresponding author E-mail: tamerhassan81@hotmail.com*

Abstract

Dimension reduction is one of the important requirements for a successful representation to improve the efficiency of extracting the attracting trend patterns on the time series. Furthermore, an efficient and accurate similarity searching on a huge time series data set is a crucial problem in data mining preprocessing. 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 methodology utilized in this paper is proposed to overcome symbolic representation pattern mismatch. 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. 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 showed the accuracy and effectiveness of the proposed approach.

Keywords: *Time Series Representation, Dimensionality Reduction, Data Mining, Pattern Discovery, Optimization.*

1. Introduction

A time series is a set of unique time points with each value assigned to time point. In the recent financial databases, huge time series data sets are very common such as stock market applications. Similarity search is a useful tool for exploring time series databases to look for a specific pattern within a longer sequence. Efficient similarity searching in a large amount of time series data is nontrivial problem. Searching for a similar pattern in a huge time series data is usually attached by data dimension reduction. However, researches are oriented to the time series data compression ignoring the wastage information. Typically, these data compression processes are only intended to reduce data dimension [5]. The problem is that most of data reduction techniques 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 patterns. Consequently, if the length of the intervals is very large, some of the details that describe the data may be lost leading to patterns mismatch and if the length of each interval is too small then it will not have enough data to produce patterns. Time series data are being generated from daily transactions of stock market. Moreover, there has been a great interest in mining such data with a lot of works introducing new methodologies for clustering and classifying time series data [17],[22]. Han and Kamber noted the time series are essentially high dimension data [17]. Dealing with this raw data format is very expensive in terms of processing and storage cost. Consequently, it is required to develop time series data representation strategy that can reduce the time series dimension, while preserving the fundamental characteristics of the data set. Furthermore, the distance between time series data must be defined to reflect the underlying data similarity. This is desirable for similarity, classification and the other mining time series methodologies [17]. Time series similarity search task in the stock market exchange 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, many time series data representation techniques and similarity measures applied in the high quality conferences and journals have been evaluated. Specifically, the proposed representation methodology for time series data have been compared with different time series data sets.

Symbolic Aggregate Approximation (SAX) was proposed as a new time series data representation technique [8]. SAX is based on the Piecewise Aggregate Approximation (PAA) time series representation technique that reduces the dimension of the data by the mean values of equally sized frames [4]. SAX discretizes the original time series data into specified symbolic strings and distance measures. In this work, a new approach is proposed which integrates the SAX pattern matching with a recent biased technology to improve the quality of SAX similarity search. The proposed approach consists of the following steps:

- Dimension reduction using PAA applying data segmentation using recent biased technology
- Time series data discretization into symbolic representation using SAX.
- Symbolic strings pattern matching.
- Post processing using Piecewise Linear Approximation (PLA) to select between the candidate patterns given the valid patterns by the third step which are approximately similar in shape with the query pattern but in fact they are not a real similar patterns.

Time series tends to repeat periodically and creates a pattern that alters over time. Since the pattern changes over time, the most recent pattern is more significant than older ones [1]. In this study, a new recent 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 Discrete Fourier Transformation DFT, Discrete Wavelet Transformation DWT, Single Value Decomposition SVD, PAA and SAX can be used.

This study is distinguished from the other previously proposed similarity search methodologies by the following contributions:

- Propose novel technique to solve the similarity search problem in the high dimensional data, mainly time series data. The similarity search is also addressed from a perspective that is as close to financial time series data as possible by optimizing the similarity search using the recent biased weight of the data.
- To overcome the symbolic representation point of weakness by applying an integration of symbolic representation and Piecewise Linear Approximation to minimize pattern mismatch.

The proposed study 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 [36], 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 [37] which is the usage of old dimension reduction technique (DWT) and ignoring the other state of the art techniques such as SAX that have not even being considered in the experimental comparisons.

The rest of this work is organized in 5 sections. Section 2 gives a background about the different time series similarity measures and representation. Section 3 and Section 4 present the main contributions of this study and the results of the experimental evaluations of the different representation methods and similarity measures. Section 5 discusses the conclusion of this study and possible future extensions.

2. Background and related work

Time series data are often extremely large. Consequently, searching on these data will be complex and inefficient. To overcome this problem, some of transformation methods should be applied to reduce the magnitude of time series databases. In the literature, many techniques have been proposed to represent time series with reduced dimensionality, such as PAA [24], DFT [13], Indexable Piecewise Linear Approximation (IPLA) [11], SVD [13], DWT [12], SAX [30] and many other techniques. Moreover, there are many distance measures for similarity of time series data in the literature such as Euclidean distance (ED) [13] and Dynamic Time Warping (DTW) [7, 26]. Classification techniques can be also divided into two categories [44]. The first category includes techniques based on shape similarity metrics where distance is measured directly between time series points. The principal classical example from this category is 1NN classifier built upon Euclidean distance [45] and Dynamic Time Warping (DTW) [46]. The second category is based on structural similarity metrics such as classifiers based on time series representation obtained with DFT [47] or Bag-Of-Patterns [48]. The current most commonly used approach is the SAX [38], [39], [40], [41]. SAX applies dimension reduction technique as a preprocessing step using PAA [38] 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 [42]. There has recently been a growing interest in SAX [15] proposed by Eamonn et al. (2003) which enables information extraction of time series representations based on PAA representation [16]. In many applications such as stock market prices, concerning the dimension reduction techniques, recent data are much more interesting and significant than old data [1]. Zhao and Zhang have proposed the equi-segmented scheme [35] by dividing 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. Many of these works and some of their extensions have been widely cited in the literature and applied to facilitate query processing of time series data. 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. Comparison between the different data representation techniques can be shown in Table1. Since SAX is based on PAA approach, it suffers from the following disadvantages:

Concerning MINDIST function that returns the minimum distance between the original time sequences S, Q of the two symbol strings reducing the dimension from n-dimension to w-dimension S = $\grave{s_1}, \ldots\ldots, \grave{s_w}$ and Q = $\grave{q_1}, \ldots\ldots, \grave{q_w}$

$$\text{MINDIST(S, Q)} = \sqrt{\frac{n}{w}}\sqrt{\sum_{i=1}^{w}(dist(\grave{s_\iota}, \grave{q_\iota})^2} \quad (1)$$

When the reduction scale (n/w) is large, the chance of information loss increases and if n/w is small, PAA becomes meaningless.

In the SAX based pattern matching, if the predefined distance R is too small, no candidate subsequences can be found even though there are many subsequences having the same shape with the query pattern. Furthermore, if R is too large, there are many candidates that are not correct.

To overcome these disadvantages, PLA is used in order to compare the patterns in more direct manner.

Table 1: Comparison of Data Representation Techniques

Techniques	Authors, year	Advantages	Disadvantages /Restrictions
DFT	Agrawal et al. 1993	Convert any complex time series into sine/cosine waves with high compression ratio.	"Wavelets outperform the DFT" [34] and the users are 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 [19] 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.
PLA	Keogh et al. 2001	Minimize pattern mismatch when applied in a hybrid with SAX [50].	The hardness of choosing the optimal number of segments used to represent a particular time series which has no general solution.
SAX	Lin et al. 2007	Offers effective methods that are applicable in many applications.	The main SAX disadvantage is that it depends on applying PAA model assumptions.
SVD	Faloutsos et al. 1994	Efficient dimension reduction technique.	The hardness of computation especially for time series data.

3. Methodology

Lin and Keogh et al. [8] proposed the SAX technology which is based on PAA and assumes normality of the resulting aggregated values. SAX maps the PAA representation of the time series data into a sequence of discrete symbols. SAX is a symbolization method that involves placing a symbol for each segment obtained by 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. In this study, dimension reduction is applied using PAA and then the reduced time series is transformed into a symbolic representation using the recently proposed SAX (RBSAX) algorithm as shown in Fig.1. SAX allows a time series of arbitrary length n to be reduced to a string of arbitrary length w and the alphabet size is an integer a where a > 2.

Since SAX is based on PAA with its disadvantages as previously explored, PLA is used in order to compare the patterns in more direct manner but still not on raw data. PLA denotes the advantage of minimizing pattern mismatch which is the main drawback of applying SAX technology. The PAA dimensionality reduction is intuitive and simple, yet has been shown to rival more sophisticated dimensionality reduction techniques like Fourier transforms and wavelets [22, 23 and 35]. The flow chart of analyzing the data using PLA is shown in Fig.2. In this figure, PLA is used to interpolate the stock market index data EGX 30 while applying symbolic representation with recent biased technology. This interpolates a function of one variable using the 'linear' method which used for the piecewise linear interpolation.

A time series C of length n (C = $c_{1,\ldots\ldots,}c_n$) can be represented in a w dimensional space by a vector $\overline{C}$ = $\overline{C}_{1,\ldots,}\overline{C}_w$ which is the PAA representation of the time series with the i Th element of $\overline{C}$ *is* calculated by the following equation:

$$\overline{C}_i = \frac{\mathrm{w}}{\mathrm{n}} \sum_{\mathrm{j}=(\mathrm{i}-1)\frac{\mathrm{n}}{\mathrm{w}}+1}^{\mathrm{i\ n/w}} C_j \quad (2)$$

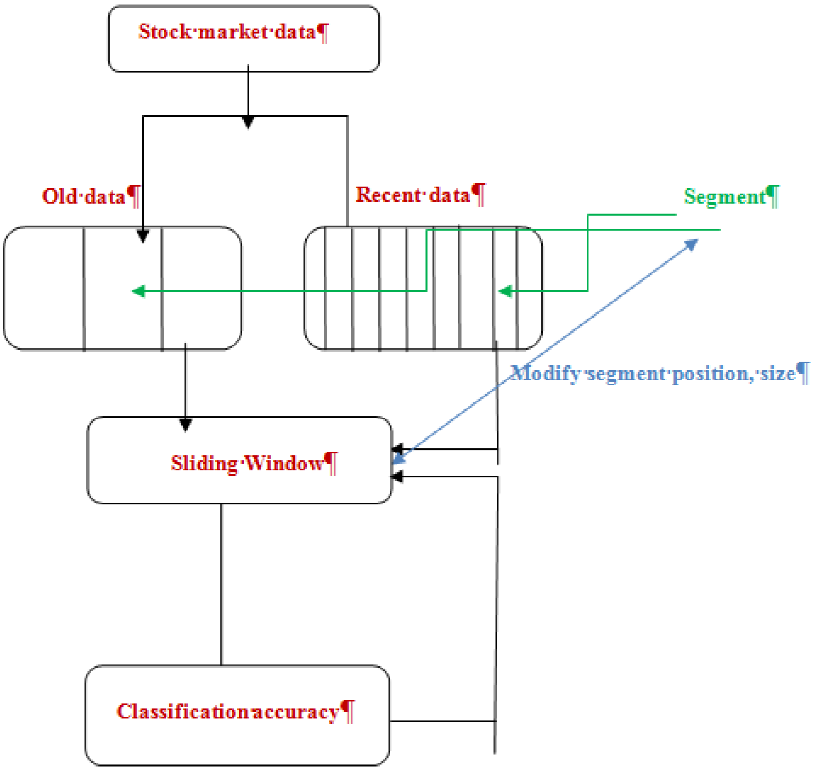

Fig. 1: Applying RBSAX on the Stock Market Index Data

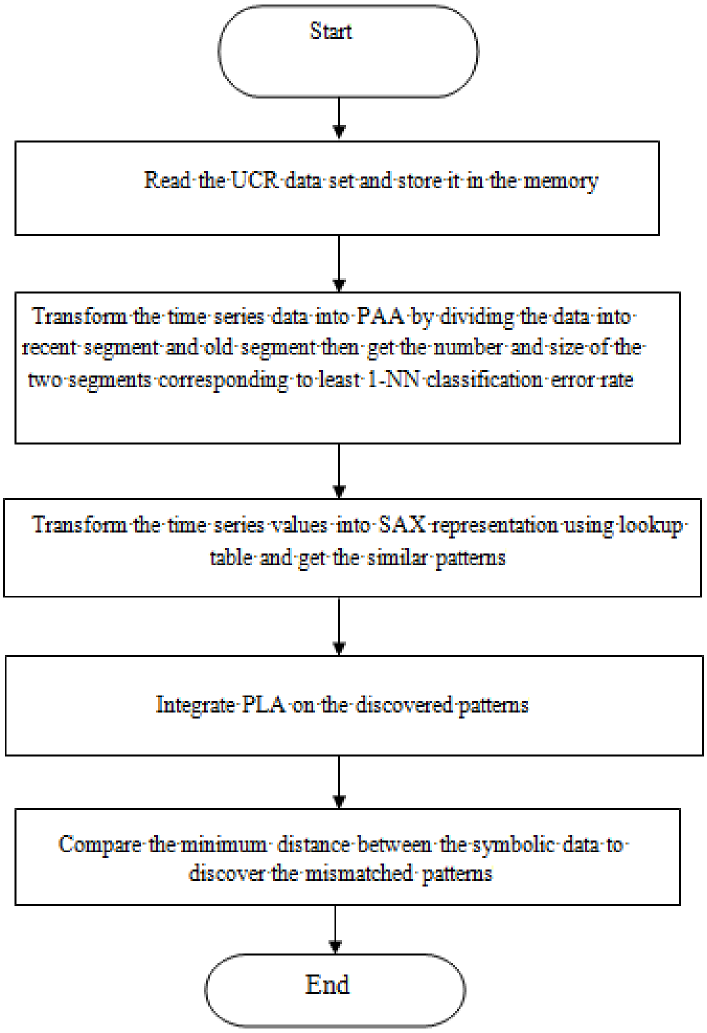
Start
Read·the·UCR·data·set·and·store·it·in·the·memory
Transform·the·time·series·data·into·PAA·by·dividing·the·data·into·recent·segment·and·old·segment·then·get·the·number·and·size·of·the·two·segments·corresponding·to·least·1-NN·classification·error·rate
Transform·the·time·series·values·into·SAX·representation·using·lookup·table·and·get·the·similar·patterns
Integrate·PLA·on·the·discovered·patterns
Compare·the·minimum·distance·between·the·symbolic·data·to·discover·the·mismatched·patterns
End

Fig. 2: The Flow Chart of the Outlined Proposed RBSAX Optimized By PLA

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 reduced data representation. PAA can be visualized as an approximation of the original time series data with a linear combination of box basis functions as shown in Fig. 3.

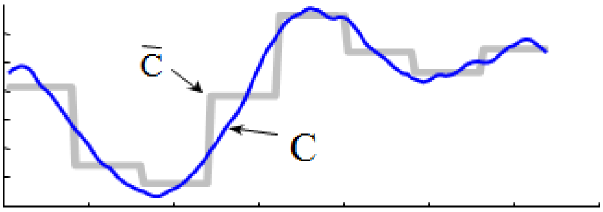

Fig. 3: The PAA Linear Combination of Box Basis Functions Representing Time Series Data

It is required to discretize the data to produce symbols with equal probability [3, 28]. This can be easily achieved since a normalized time series have a Gaussian distribution [27]. The breakpoints can be simply determined which will produce an equal sized area under the Gaussian curve [27]. These breakpoints may be determined by looking them up in a statistical table. Table2 denotes the breakpoints for the values of 'a' from 3 to 10.

Table 2: Lookup Table Contains the Breakpoints That Divide a Gaussian distribution in an Arbitrary Number (From 3 to 10) of Equal Probability Regions

a	3	4	5	6	7	8	9	10
β_1	-0.43	-0.67	-0.84	-0.97	-1.07	-1.15	-1.22	-1.28
β_2	0.43	0	-0.25	-0.43	-0.57	-0.67	-0.76	-0.84
β_3		0.67	0.25	0	-0.18	-0.32	-0.43	-0.52
β_4			0.84	0.43	0.18	0	-0.14	-0.25
β_5				0.97	0.57	0.32	0.14	0
β_6					1.07	0.67	0.43	0.25
β_7						1.15	0.76	0.52
β_8							1.22	0.84
β_9								1.28

Each time series is normalized to have a mean of zero and a standard deviation of one before converting it to the PAA representation to compare time series with same offsets and amplitudes [23] then 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 $\overline{C}$ to symbol $\hat{C}$ knowing that $\widehat{C_\iota} = \widehat{c_1}$,…., $\widehat{c_w}$ is obtained as the following equation:

$\widehat{C}ı = alpha_j$ iif $\sum$ j-1≤ $\overline{C_i}$ ‹ $\sum j$ (3)

Once the breakpoints have been obtained, the time series can be discretized in the following manner. After obtaining the PAA 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. Fig. 4 illustrates the idea.

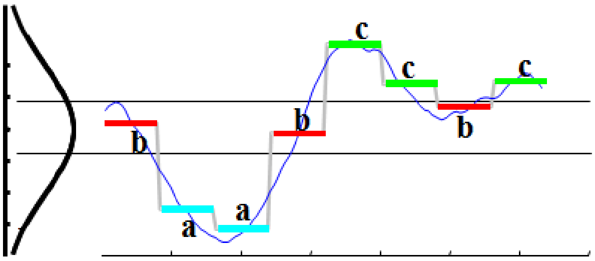

Fig. 4: Time Series Discretization Using Predetermined Breakpoints PAA Approximation into Sax Symbols

The time series database is transformed into PAA and then a further transformation can be applied to obtain a discrete representation. The framework of this study can be explored in Fig.5 showing the main contribution of this study.

Applying SAX on the stock market index EGX 100 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 Fig.6 with n = 2022, number of segments = 41 and alpha size = 10, the time series is mapped to the word "edbbcdefiijjjjjicbbdehhggfdegfddcbbcbbcbc". The Egyptian stock market index data is divided into two segments. The position of the two segments is positioned and controlled 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 [1] optimized by PLA. 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 the minimum error is obtained. Furthermore, each segment is divided into sub-segments in which the number of these 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 maximum classification accuracy.

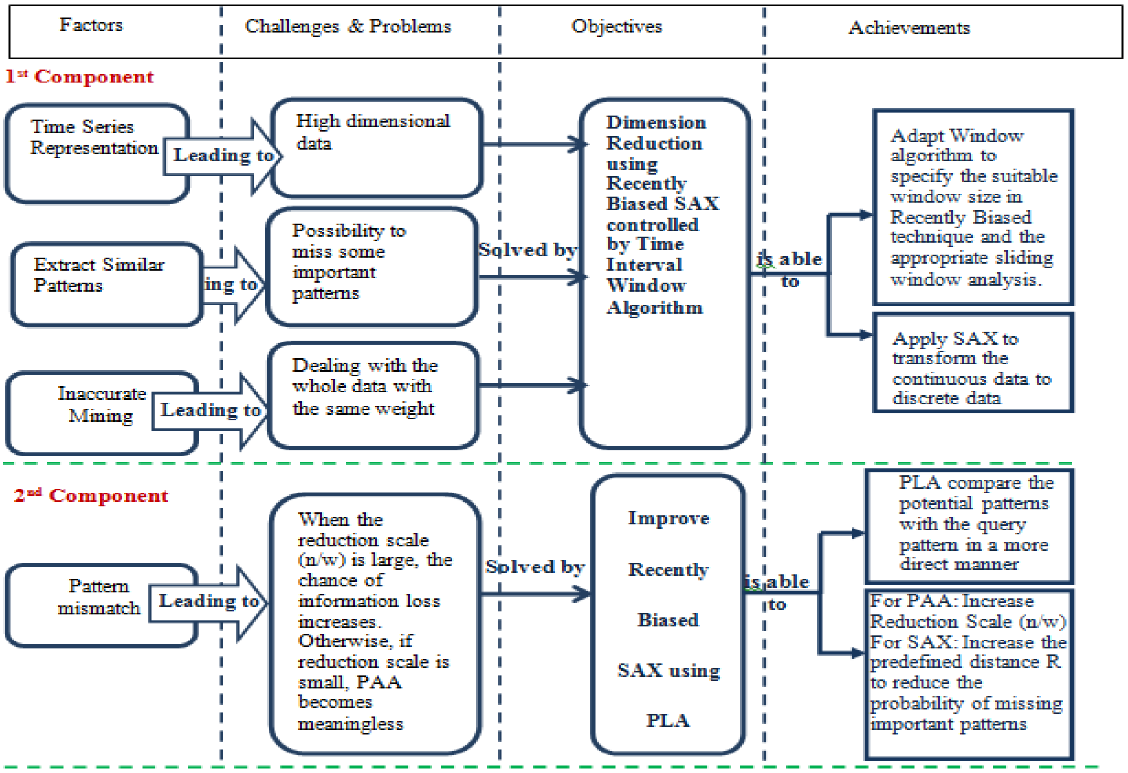
Factors
Challenges & Problems
Objectives
Achievements
1st Component
Time Series Representation
Leading to
High dimensional data
Extract Similar Patterns
ing to
Possibility to miss some important patterns
Inaccurate Mining
Leading to
Dealing with the whole data with the same weight
Solved by
Dimension 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
2nd Component
Pattern mismatch
Leading to
When the reduction scale (n/w) is large, the chance of information loss increases. Otherwise, if reduction scale is small, PAA becomes meaningless
Solved by
Improve Recently Biased SAX using PLA
is able to
PLA compare the potential patterns with the query pattern in a more direct manner
For PAA: Increase Reduction Scale (n/w) For SAX: Increase the predefined distance R to reduce the probability of missing important patterns

Fig. 5: Paper Framework

The original Egyptian index can be seen in Fig.6 before applying the dimension reduction technique. In Fig.7, the PAA normalization of the Egyptian index can be shown.

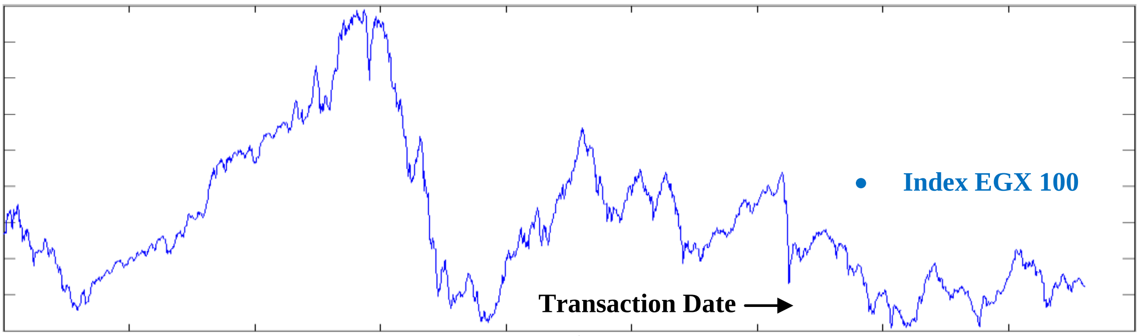

Fig. 6: EGX 100 before Normalization

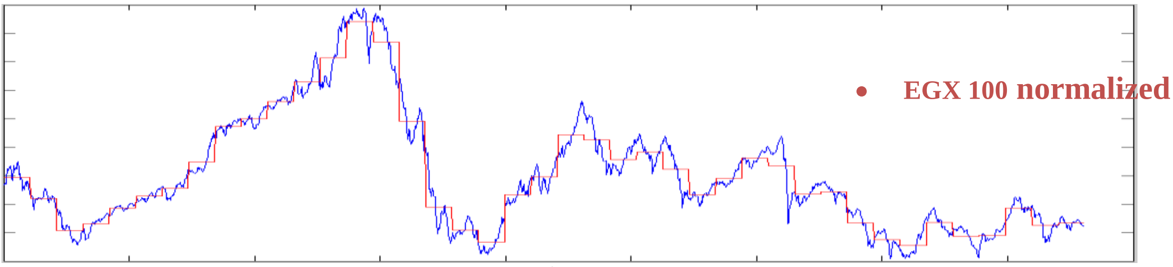

Fig. 7: EGX 100 after PAA Normalization

Finally, after applying SAX the resulted data can be shown in Figure 8.

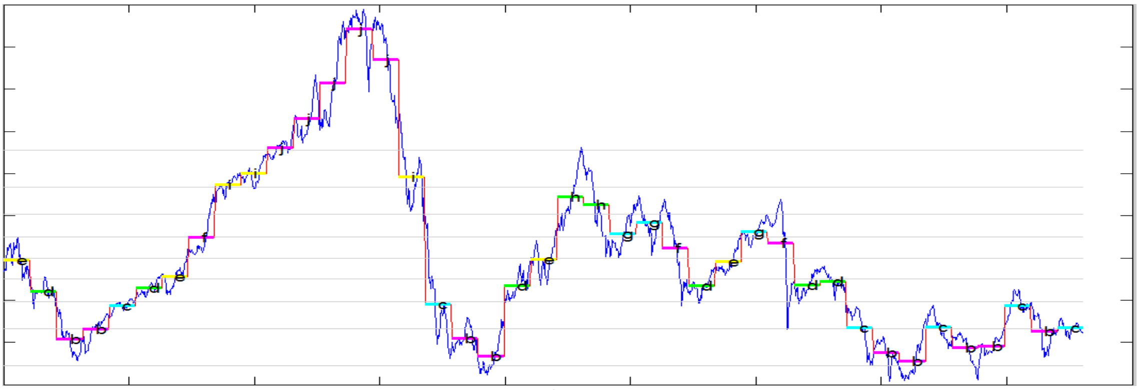

Fig. 8: EGX 100 after SAX Discretization

4. Experimental results

The recent biased technology combined with the symbolic representation and optimized using PLA 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 integration showed promising results in many data sets as explored in Table 3.

Table 3: Standard UCR Data Set Comparing Classification Error Using RBSAX Optimized By PLA Methodology

Data Set	Classes	ED	DTW	DFT	SAX	RBSAX+PLA
50words	50	0.407	0.375	0.521	0.341	0.429
Adiac	37	0.464	0.465	0.476	0.89	0.352
Beef	5	0.467	0.433	0.493	0.567	0.422
Car	4	0.275	0.333	0.316	0.333	0.341
CBF	3	0.087	0.003	0.16	0.104	0.029
chlorineconcentration	3	0.349	0.38	0.415	0.41	0.438
cinc_ECG_toeso	4	0.051	0.165	0.087	0.195	0.238
Coffee	2	0.193	0.191	0.249	0.464	0.316
diatomsizereduction	4	0.022	0.015	0.13	0.152	0.077
ECG200	2	0.162	0.221	0.283	0.12	0.25
ECGFiveDays	2	0.118	0.154	0.21	0.263	0.215
FaceFour	4	0.149	0.064	0.172	0.17	0.166
Faces(all)	14	0.225	0.192	0.216	0.33	0.134
fish	7	0.319	0.329	0.493	0.474	0.316
Gun Point	2	0.146	0.14	0.281	0.18	0.088
Lighting2	2	0.341	0.204	0.338	0.213	0.208
Lighting7	7	0.377	0.252	0.363	0.397	0.381
OliveOil	4	0.15	0.133	0.623	0.833	0.097
OSULeaf	6	0.448	0.401	0.535	0.467	0.379
plane	7	0.051	0.001	0.112	0.038	0.036
SwedishLeaf	15	0.295	0.256	0.319	0.483	0.411
synthetic control	6	0.142	0.019	0.134	0.02	0.193
Trace	4	0.368	0.016	0.427	0.46	0.014
TwoPatterns	4	0.095	0	0.163	0.081	0.93
wafer	2	0.005	0.015	0.09	0.004	0.003
yoga	2	0.16	0.151	0.184	0.195	0.128

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 combined with PLA methodology according to the controlled sliding window corresponding to the minimum classification error 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 optimized by PLA showed a promising results as well as shown in Table 4.

Table 4: Egyptian Stock Market Indices Classification Error Using RBSAX Optimized By PLA Methodology

Index	Number of instances	ED error	DTW	DFT	SAX error	RBSAX+PLA error
EGX 30	2375	49.78	43.22	44.23	39.54	23.14
EGX 70	1232	47.15	41.93	42.84	40.28	21.37
EGX 100	1724	48.83	45.71	47.49	38.13	18.43

5. Conclusion

In order to produce a meaningful model a SAX classifier 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 datasets. However, in many real world domains the classification task should focus not only on accuracy, but also to produce a classification result more comprehensive, which global features of the nearest neighbor cannot provide [49].

In this context, a new symbolic representation methodology 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 using a proper sliding window beginning from 1% of the data and increased by 1% repeatedly beginning from the most recent data directed towards the older ones.

The statistical evaluation of the integrated recent biased SAX optimized by PLA indicates a significant difference between this work and the other classic state of the art techniques. According to the experimental results it can be seen that this work have a superior performance for most datasets. In the accuracy analysis there was significant difference for comparisons between applying the optimized recent biased technology comparing with the classic SAX and the other state of the art techniques such that DFT and DTW, where the classic SAX was better in many data sets this technique called RBSAX optimized by PLA presented in this work was better in significant number of data sets. On the other hand, the results presented in Table 4 explore the ability of recent biased optimized symbolic representation to classify the time series data whether having superior accuracy or not. Applying the new methodology RBSAX integrated with PLA on the standard UCR datasets improve the classification accuracy rate in 10 datasets from 1% until 33% compared with the other classification techniques. Moreover comparing the results in the Table 5 it can be easily observed that RBSAX fulfill a significant percentage of improvement. The classification error is reduced from 17% until 47% in the Egyptian stock market indices.

The proposed approach was tested on the Egyptian stock market 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 raised in a specified intervals as a direct effect of declaring any kind of elections which gives the feeling of more 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 future election intervals beginning from May 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 25th of January 2011, 9th of June 2011 and 28th of November 2012. Consequently, it is expected for the stock market index to go down in the intervals around the advertised demonstration. The discovered patterns were reviewed by professional financial experts and were expected to be accurate with an accuracy rate exceeds 70%.

Acknowledgements

First and foremost, all praise to Almighty Allah who gave me the courage, patience, health and ability to complete this research, Alhamdulillah.

My sincere appreciation and gratitude goes to my main supervisor Assoc. Prof. Dr. Zalinda Othman and to my co-supervisor Prof. Dr. Abdul Razak Hamdan for their continuous guidance, understanding, encouragement and valuable advice throughout my PhD study.

To my family, who have been there for me from the first day, thank you for all the love, support, encouragement and continuous prayers for me to complete my study? Without your help I definitely would have not completed my doctoral degree. May Allah bless you all?

References

[1] Phithakkitnukoon, Santi and Ratti, Carlo (2010). A recent-pattern biased dimension-reduction framework for time series data. *Journal of Advances in Information Technology*, 1(4), pp. 168–180.

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

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

[4] E. Keogh, K. Chakrabarti, M. Pazzani and S. Mehrotra,“Dimensionality reduction for fast similarity search in large time series databases”, *Journal of Knowledge and Information Systems*, Vol. 3, No. 3, 2000, pp. 263-286.

[5] Han, J.: Data mining techniques. *In: Proceedings of the 1996 ACM SIGMOD International Conference on Management of data, Montreal, Quebec, Canada*, p. 545 (1996).

[6] Y. Cai and R. T. Ng. Indexing spatio-temporal trajectories with chebyshev polynomials. *In SIGMOD Confer*ence, 2004.

[7] D. J. Berndt and J. Clifford. Using dynamic time warping to find patterns in time series*. In KDD Workshop*, 1994.

[8] J. Lin, E. Keogh, S. Lonardi, and B. Chiu, “A Symbolic Representation of Time Series, with Implications for Streaming Algorithms”, *8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discover (DMKD 2003),* June 13, 2003, pp. 2-11.

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

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

[11] Q. Chen, L. Chen, X. Lian, Y. Liu, and J. X. Yu.Indexable PLA for Efficient Similarity Search. *In VLDB*, 2007.

[12] K. pong Chan and A. W.-C. Fu. Efficient Time Series Matching by Wavelets. *In ICDE*, 1999.

[13] C. Faloutsos, M. Ranganathan, and Y. Manolopoulos.Fast Subsequence Matching in Time-Series Databases.*In SIGMOD Conference*, 1994.

[14] E. Frentzos, K. Gratsias, and Y. Theodoridis. Index-based most similar trajectory search. *In ICDE*, 2007.

[15] Jessica, L., Eamonn, K.: 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, San Diego, California*, pp. 2–11 (2003).

[16] John, F.R., Kathleen, H., Myra, S.: 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* (2001).

[17] Jiawei Han and Micheline Kamber. *Data Mining:Concepts and Techniques*.2nd edition. Morgan Kaufmann Publishers, CA, 2006.

[18] Y.-L. Wu, D. Agrawal, and A. E. Abbadi. A Comparison of DFT and DWT based Similarity Search in Time-Series Databases. *In CIKM*, 2000.

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

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

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

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

[23] E. J. Keogh, K. Chakrabarti, S. Mehrotra, and M. J. Pazzani. Locally Adaptive Dimensionality Reduction for Indexing Large Time Series Databases. *In SIGMOD Conference*, 2001.

[24] E. J. Keogh, K. Chakrabarti, M. J. Pazzani, and S. Mehrotra. Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases.Knowl. *Inf. Syst.,* 3(3), 2001.

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

[26] E. J. Keogh and C. A. Ratanamahatana. Exact indexing of dynamic time warping. Knowl. *Inf. Syst*., 7(3), 2005.

[27] C. Giannella, J. Han, J. Pei, X. Yan, and P.S. Yu, ”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, 2003.

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

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

[30] J. Lin, E. J. Keogh, L. Wei, and S. Lonardi. Experiencing SAX: a novel symbolic representation of time series. Data *Min. Knowl. Discov*., 15(2), 2007.

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

[32] M. Vlachos, D. Gunopulos, and G. Kollios.Discovering similar multidimensional trajectories. *In ICDE*, 2002.

[33] Charest, M and S. Delisle: Ontology-guided intelligent data mining assistance: Combining declarative and procedural knowledge. *Artificial Intelligence and Soft Computing* 2006: p 9-14.

[34] I. Popivanov and R. J. Miller. Similarity Search over Time-Series Data Using Wavelets. *In ICDE*, 2002.

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

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

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

[38] J. Lin, E. J. Keogh, L. Wei, and S. Lonardi, “Experiencing sax: a novel symbolic representation of time series,” *Data Min. Knowl. Discov*., vol. 15, no. 2, pp. 107–144, 2007.

[39] N. D. Pham, Q. L. Le, and T. K. Dang, “Two novel adaptive symbolic representations for similarity search in time series databases,” *in APWeb*,2010, pp. 181–187.

[40] T. Rakthanmanon and E. Keogh, “Fast shapelets: A scalable algorithm for discovering time series shapelets*,” in SDM*, 2013.

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

[42] B. Lkhagva, Y. Suzuki, and K. Kawagoe, “New time series data representation esax for financial applications,” *in ICDE Workshops*, 2006, p. 115.

[43] Keogh, E., Kasetty, S.: On the need for time series data mining benchmarks: a survey and empirical demonstration. *DMKD*, 7, 4, (2003).

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

[45] Xi, X., Keogh, E., Shelton, C., Wei, L., Ratanamahatana, C.: Fast time series classification using numerosity reduction. *In Proc. ICML* (2006).

[46] Sakoe, H. Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 1, 43–49 (1978).

[47] Agrawal, R., Faloutsos, C., Swami, A.: Efficient Similarity Search In Sequence Databases. *In Proc. FODO,* 69–84 (1993).

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

[49] L. Ye and E. J. Keogh, “Time series shapelets: a novel technique that allows accurate, interpretable and fast classification,” *Data Min. Knowl. Discov*, vol. 22, no. 1-2, pp. 149–182, 2011.

[50] Nguyen & Duong T. Anh. 2007. "Combining SAX and Piecewise Linear Approximation to Improve Similarity Search on Financial Time Series" *Information Technology Convergence, International Symposium on In Information Technology Convergence*, pp.58-62.