International Journal of Soft Computing 8 (4): 261-268, 2013
ISSN: 1816-9503

Enhanced Symbolic Aggregate Approximation (EN-SAX) as an Improved Representation Method for Financial Time Series Data

Peiman Mamani Barnaghi, Azuraliza Abu Bakar and Zulaiha Ali Othman
Data Mining and Optimization Research Group, Center for Artificial Intelligence Technology,
Faculty of Information Science and Technology,
University Kebangsaan Malaysia, 43600 Bangi, Malaysia

Abstract: Data representation is one of the most important tasks in time series data pre-processing. Time series data representation is required to make the data more suitable for data mining specifically for prediction. Time series data is characterized by its numerical and continuous values. One of the data representation methods for time series is the Symbolic Aggregate Approximation (SAX) which uses mean values as the basis of representation of the data. However, representing the time series financial data with the mean value often causes the loss of patterns that can describes important pieces of information. The aim of this study is to propose an enhancement of SAX representation purposely for the financial time series data. The Enhanced SAX (EN-SAX) adds two new values to the original mean value for each segment in SAX. These values enable better representation for each segment in a lower dimension and keep some of the important patterns that are meaningful in financial time series data. The experimental results show that the EN-SAX representation manages to give lower error rates compared to SAX and improves the prediction accuracy.

Key words: Financial time series data, dimensionality reduction, Symbolic Aggregate Approximation (SAX), (EN-SAX), pre-processing, Malaysia

INTRODUCTION

There is an important class of data object that is pertaining to time which is called time series data. Time series data can be defined by specific characteristics including: large size of data that is characterized by its numerical values with high dimensionality and continuous updating. The essential problem in the context of time series data mining is how to represent this kind of data. Time series data mining is one of the important issues to extract useful information such as prominent patterns, rules and structured descriptions for a data domain. Because of large size of time series data, it is difficult to find out such prominent knowledge without reduce the dimension. Transforming time series data to another domain is one of the frequent solutions for dimensionality reduction followed by indexing mechanisms. Basically the algorithms for feature selection and dimensionality reduction are used to reduce the dimensions without losing the important patterns and information content of the domain in a dataset. Feature selection basically consist two main categories: supervised and un-supervised. The supervised method is used where a class label associate each illustration and unsupervised method is used where there is not any relation between instances with any class label. In order to dimensionality reduction as a pre-processing of machine learning techniques such as classification, clustering or association rule mining, unsupervised method is used in the domain space with keeping the prominent patterns and avoiding the loss of information content. Dimensionality reduction as one of the related search in time series data is mentioned in this study for improving accuracy to predict and forecasting time series data using data mining techniques. Some important patterns have significant rule to prediction and impact on future plans and policies. It has a high probability of missing some of these kinds of patterns. Therefore, keeping these significant patterns during the reduction of financial data dimensionality is very important. Dimensionality reduction method based on common method is proposed in this study to improve the accuracy of prediction for financial time series data.

Literature review: There are various methods for representation time series data. The simplest one is the

Corresponding Author: Peiman Mamani Barnaghi, Data Mining and Optimization Research Group,
Center for Artificial Intelligence Technology, Faculty of Information Science and Technology,
University Kebangsaan Malaysia, 43600 Bangi, Malaysia

shape of sampled time series with using length of a time series and dimension after dimensionality reduction. This method can be used whilst the rate of sampling is too low and missing some values is not important. An improved method that is called Piecewise Aggregate Approximation (PAA) presents the time series using the mean (avg) value of each segment (Fu, 2011). An extended version of representation that is proposed by Keogh *et al.* (2004) is called Adaptive Piecewise Constant Approximation (APCA). This method of representation is different with earlier method (PAA) in which the length of each segment is not fixed but totally versatile and adaptive to the shape of the original time series data but totally it is not suitable for financial time series data because of high possibility to miss prominent patterns in different segments which the main reason is the length of each segment that is not fixed.

Piecewise aggregate approximation: Piecewise Aggregate Approximation (PAA) is one of the representation time series data methods which uses the average value for each equal sized segment. This method uses the segmented means to represent time series data (Yi and Faloutsos, 2000; Keogh and Pazzani, 2000). Figure 1 shows the PAA Method. This method was originally called piecewise constant approximation (Keogh and Pazzani, 2000; Buu and Anh, 2011).

Symbolic Aggregate Approximation (SAX): Symbolic Aggregate Approximation (SAX) is a common method for representation of time series data. SAX transfers numeric time series to a new form of data. As mentioned before one of the attributes of time series data is including numerical values. This approach transforms the time series from numerical state to symbolic form. SAX discretizes time series data into segments and then transforms each segment into a symbol by introducing a new form of representation the time series data (Aref *et al.*, 2004; Abu Bakar *et al.*, 2010). SAX transforms numerical data to symbolic form by using the result from Piecewise Aggregate Approximation (PAA).

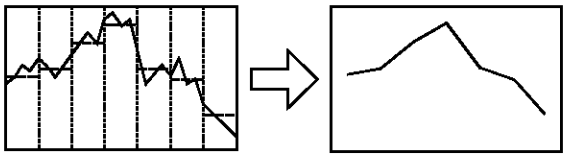


Fig. 1: Dimensionality reduction on time series by Piecewise Aggregate Approximation (PAA) that mean value of each segment displayed with horizontal dotted lines (Fu, 2011)

PAA uses the average value for each equal sized segment to represent time series data (Yi and Faloutsos, 2000). In other words SAX converts the results that are obtained from PAA to string of symbols.

Time series data is shown by two major parameters. One is time that can be based on hour, day, month, etc. which is shown at x-axis and the other is numerical value of data which is shown at y-axis. The segmentation of data is done on the time axis. The distribution space of data (y-axis) is divided into regions that are equi-probable. Each region is represented in symbolized form and each segment of data can be mapped into a certain symbol that is matching with related region's symbol that it resides. In order to transform time series to symbolic form, two parameters must be determined. One is the length of each segment and the other is the number of symbols that are used for this conversion. Figure 2 shows the mapping of time series to symbolize form based on SAX.

Extended SAX (ESAX): Lkhagva introduces new time series data representation ESAX for financial applications which uses Min, Avg, Max values as string of symbols to search similarity of shapes between financial time series data. The accuracy is measured just by the number of shapes that are similar to original shape of financial time series data. But there is no numerical values as output to measure the similarity between original data and represented data using ESAX Method. The Enhanced SAX (EN-SAX) as proposed method is totally different with Extended SAX (ESAX) which is proposed by Lkhagva *et al.* (2006a, b). Some main difference are shown in the Table 1.

The main differences of the proposed EN-SAX with the ESAX by Lkhagva *et al.* (2006a, b) are three folds. First, researchers use Min, Avg, Max values as vector of data to have similarity search of data whereas they used Min, Avg, Max values as string of symbols to search for shapes. Secondly, in this study K-means clustering method is used to determine the symbols zones while the previous research use the Gaussian curve to determine symbolization region. Thirdly, cosine similarity is to determine similarity between vectors of data while they

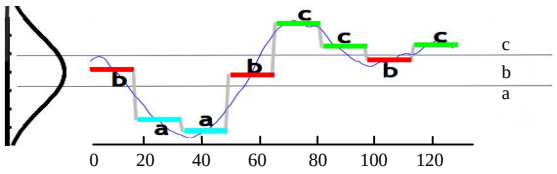


Fig. 2: Time series with length of 128 data that distribution zone divided to 8 segments (x-axis) and 3 regions (y-axis) and is mapped to the word "baabccbc"

Table 1: Main difference between Extended SAX (ESAX) and Enhanced SAX (EnSAX)

Extended SAX	Enhanced SAX
Using Min, Avg, Max values as string of symbols to search similarity of shapes	Using Min, Avg, Max values as vector of data to similarity search of data
Using Gaussian curve to determines symbolization region	Using K-means clustering method to determines symbols zones
Using Euclidean distance to evaluate the similarity between shapes of time series data	Using cosine similarity to determines similarity between vectors of data

use Euclidean distance to evaluate the similarity between shapes of time series data (Lkhagva *et al*., 2006a, b).

The reason of using clustering method to determine symbolization region is mapping each vector of data at each segment to relate symbol for that segment. Also, cosine similarity is used for similarity measurement instead of Euclidean distance because prominent points are very important in financial time series data which these points will be ignored using Euclidean distance (Megalooikonomou *et al*., 2005). Cosine similarity between two vectors determines whether two vectors are pointing in roughly the same direction. The next section will present the proposed enhanced SAX.

ENHANCED SYMBOLIC AGGREGATE APPROXIMATION

Although, SAX is an appropriate method for time series dimensionality reduction but it is not suitable for financial time series data because it is based on mean values estimate and there are some important patterns such as extreme value and unusual patterns in financial time series that with applying SAX, it has high possibility to miss some of these kinds of patterns. Based on characteristic of financial time series data (Si-Zhi *et al*., 2006) the important patterns have significant role to prediction and decide on future plans and policies. Therefore, keeping these significant patterns during the reduction of financial data dimensionality is very important (Lkhagva *et al*., 2006b). Figure 3 shows some important patterns that are missing while SAX representation has been used on a sample financial time series data.

In the example above, there are two extreme values in third equal sized segment that are missed by using SAX representation that relies on mean value of each segment. These extreme values indicate some fluctuation in the sample exchange market rates and where in the real world applications keeping these values is very important. The mean values of third segment [20~30] is represented as symbol "C" but there are two important points that must be presented as "A" and "F". It is very important to represent these kinds of meaningful important values.

Enhanced SAX (EN-SAX) uses two additional Min and Max points with original mean value for each segment in time series data. This helps to preserve some important

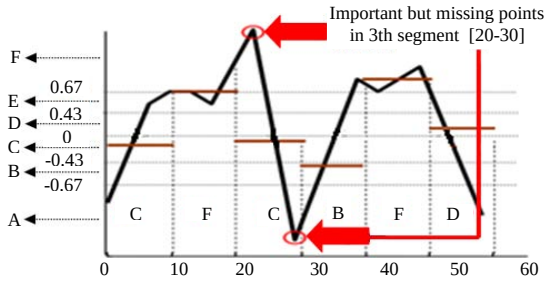


Fig. 3: Some missed important financial patterns (shown in circle) represented by SAX (Lkhagva *et al*., 2006a)

points that are meaningful (especially in financial time series). The example (Fig. 4) illustrates how EN-SAX using mean values in each equal sized segment that are obtained from PAA process (i.e., similar to SAX) and by adding two new Min and Max points helps to detect important values and improves the accuracy.

As it is shown in the Fig. 4, in addition to mean value (shown by rectangular) in each segment, two other Min and Max values (Min are shown by diamond and Max by triangle) are considered for each segment to preserve the important patterns on financial time series data. Also, there is a sample diagram to show the difference between actual data and predicted results using Min, Avg and Max value together (Fig. 5).

As mentioned before EN-SAX is based on SAX and includes two new Min and Max points that help to detect important values and improve the accuracy. There are four main steps to implement the representation method for time series data. The steps are representation and indexing, similarity measurement, segmentation and pattern discovery. A combination of these steps is used in a special sequence to implement EN-SAX.

Data representation: By applying clustering, the current data is grouped in 20 different clusters. At this step, to implement EN-SAX Method, it is needed to map each set of data to a special symbol. There are 20 clusters for the data set (cluster_0-cluster_19). Table 2 shows mapping of each clusters to a symbol.

Similarity measurement: Based on the earlier steps, original exchange data is transferred to new symbolized

form. Researchers describe evaluation of the error rate between original data and the new symbolized data in the following section. Researchers calculate mean value for each symbol using number of the records for each symbol. For example if the number of records for cluster_0 that is mapped to symbol A equal with X, researchers have a matrix with X rows and 3 columns [Min, Avg, Max]. In order to calculate mean value for this matrix, researchers estimate Min (Min(A)), average (Avg(A)) and max (Max(A)) value for each column of matrix that is calculated Eq. 1-3:

$$\text{Min (A)} = (\text{average [cells (A1... Ax , 1)]}) \quad \text{//Min value column shows with 1} \tag{1}$$

$$\text{Avg (A)} = (\text{average [cells (A1... Ax, 2)]}) \quad \text{//Avg value column shows with 2} \tag{2}$$

$$\text{Max (A)} = (\text{average [cells (A1...Ax, 3)]}) \quad \text{//Max value column shows with 3} \tag{3}$$

The mean value provides a vector with 3 values [Min, Avg, Max] as in Eq. 4:

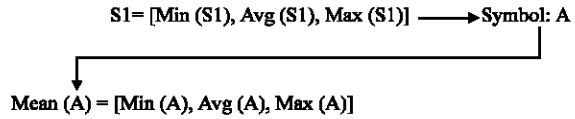
$$\text{Mean (A)} = [\text{Min(A), Avg(A), Max(A)}] \tag{4}$$

These steps are repeated for all remaining values for each vector. Finally researchers will have 20 vectors with [Min, Avg, Max] that each of them represents a mean value for each cluster that is mapped to a symbol. Figure 6 shows calculating Min, Avg and Max value for each symbol.

Table 2: Mapping clusters to symbols

Cluster	0	1	2	3	4	5	6	7	8	9
Symbol	A	B	C	D	E	F	G	H	I	J
Cluster	10	11	12	13	14	15	16	17	18	19
Symbol	K	L	M	N	O	P	Q	R	S	T

In order to calculate error rate, at the first record of each segmented data that is mapped to a symbol is compared with equal symbol mean value. For example, suppose that values for segment 1 of data are shown with S1 = [Min, Avg, Max] that corresponds to a cluster symbol such as A. The value for mean (A) is extracted from the related table. Now researchers have two vectors that consist:



One of the vectors is as a record from segmented data with related symbol after clustering and the other is

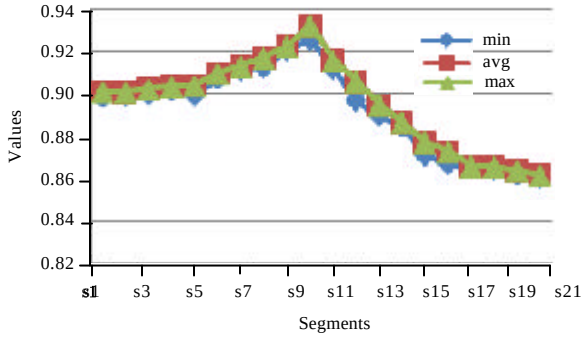


Fig. 4: Financial time series data is represented by Enhanced SAX (EN-SAX)

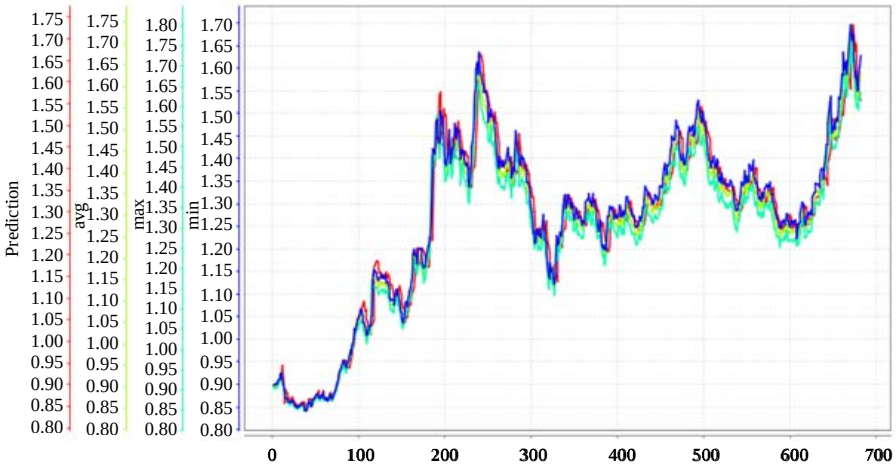


Fig. 5: Prediction financial time series data using Min, mean and Max values of segmented data which is used EnSAX to dimensionality reduction of data

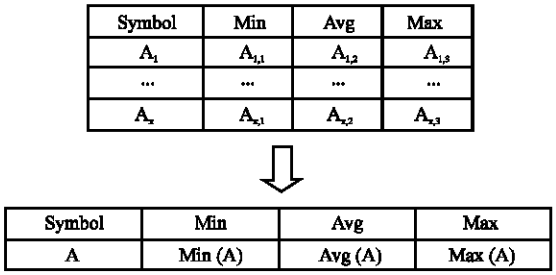


Fig. 6: Calculation of Min, Avg and Max value for each symbol

also a single vector that shows mean value for related symbol. To determine the similarity between these two data items, researchers use cosine distance value between the vectors.

In order to measure the similarity between two vectors, the cosine of the angle between them is measured. The latter is called cosine similarity. In the field of data mining, cosine similarity is used to measure the cohesion within clusters. Cosine similarity between two vectors determines whether two vectors are pointing in roughly the same direction. The cosine of 0 is 1 and less than 1 for any other angle (Tan *et al*., 2006). The cosine of two vectors is derived by using the below formula. Given two vectors of attributes, A and B, the cosine similarity (θ) is shown in Eq. 5:

$$\text{Similarity} = \text{Cos}(\theta) = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \quad (5)$$

The resulting similarity ranges between -1 to 1. The value of -1 means exactly opposite and 1 means exactly the same and between values indicates intermediate similarity or dissimilarity of the two vectors.

Data segmentation: The implementation of EN-SAX for financial time series has various steps that some primary steps are nearly similar to SAX. At first, the data must be divided to equal sized parts which are defined as segment. The size of segments is determined based on type of data and application domain. Since, the data is based on daily exchange rates, the size of each segment for this study is chosen as 7. This enables us to survey the data based on weekly rates and also by merging 4 segments that represent 4 weeks, it can be surveyed monthly and so on. An attribute is also defined as id for each segment.

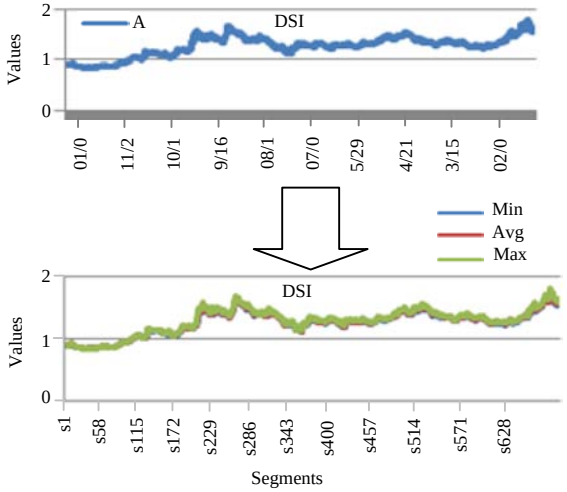


Fig. 7: Segmentation of DS1 with adding Min, Avg and Max values

Segmentation values: As mentioned before for SAX, the mean values for each segment is calculated and used for representation of the data for each segment. EN-SAX uses two new Min and Max points in addition to the original mean value for each segment in time series data segments. The single value for each segment as it is represented in SAX is changed to a vector in EN-SAX.

The vector representation in EN-SAX is represented with [Min, Avg, Max] values with another attribute that is used as id for each segment. Using the data set, the original data is transformed to 682 segments (i.e., as described above the segment size is 7 so the whole data set is divided to 628 segments that each segment holds the exchange data for 7 days). Researchers then create a vector representation for each segment that is represented with [Min, Avg, Max] values. Figure 7 shows the segmentation of DS1 compared with original data.

Pattern discovery: The next step to implement EN-SAX is categorising and grouping different cases on the available data. Researchers use RapidMiner (http://rapid-i.com/content/view/181/190/) data mining tool for the clustering method. Researchers cluster the data that is changed to segments represented with three values for each segment. The parameters for clustering are determined based on the data type and application domain. In this study, the number of clusters is set to 20 which is determined based on kind of data and type of application with arbitrary experience. That means the data is categorised to 20 different clusters using clustering process algorithm (Fu, 2011). Figure 8 is an illustration of clustering for DS1.

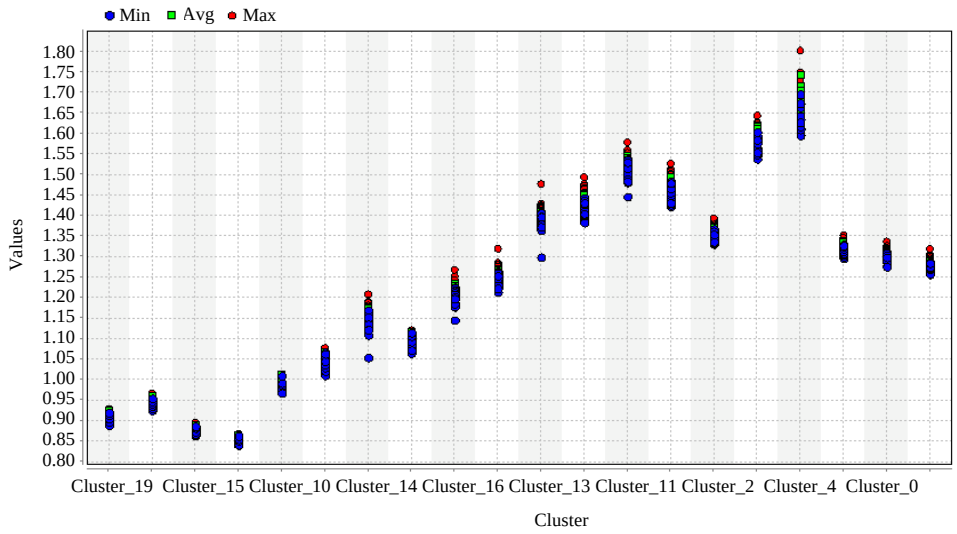


Fig. 8: Distribution diagram of clusters between segments based on Min, mean and Max values for each segment and related cluster for DS1

EXPERIMENTS

The performance of EN-SAX is evaluated in terms of error rate and accuracy of prediction. It is compared with the original SAX. The exchange rate data that is transformed to new form of representation in order to reduce the dimensionality while providing higher accuracy than the existing methods and keeping important patterns that have high possibility to be missed using the original SAX Method. At first the data is prepared using EN-SAX and then researchers apply linear regression and SVM Method to perform the prediction.

Data collection and preparation: The current dataset is exchange time series data which is downloaded from the official website of the World Bank. Various exchange rates in a special period of time in a vast domain with high dimensions are selected as the financial time series steps of the method. The world bank database currently covers 280,000 working papers, 900 journals, 2,800 book chapters, 2,400 books and 1,700 software components. The archived exchange data from a high validity resource is selected to be used in the experiment. The period of time is 1980 to 1998 and the attributes are various exchange rates compared with the United States dollar. These rates consist: Australian dollar (AUSTRUS), British pounds (BRITPUS), Canadian dollar (CDNDLUS), Japan yen (JAPYNUS) and Swiss franc (SWISFUS) which are shown in Table 3.

The number of fields for each exchange rate is around 4770 and the start date is 2 January 1980 and the end date is end of year 1998. The data is downloaded in CSV format from the website (The World Bank) and is saved as main data set in an MS Excel sheets for starting various steps of this study (WorldBank). The first item is Australian-US dollars.

Table 3: The number of data and the dates within the dataset

Data set	Data set	Total No.	Data of data
DS1	Australian dollar	4770	1980~1998
DS2	British pounds	4770	1980~1998
DS3	Canadian dollar	4770	1980~1998
DS4	Japan yen	4770	1980~1998
DS5	Swiss franc	4770	1980~1998

The financial data is divided to equal sized parts which are defined as segment. The size of each segment is determined based on kind of data and the nature of this research. Because the data includes daily exchange rates, the chosen size of each segment for this study is 7 which it enables us to survey the data on weekly basis and also with merging 4 segments, it can be studied on monthly basis. A unique attribute is defined as id for each segment to refer to each segment. Using the values included in each segment, researchers calculate minimum (min), mean (average) and maximum (max) values for each segment. These values are calculated in order to implement EN-SAX Method as representation solution for time series data. In the first step of representation, Min, mean and Max values are calculated for each segment (Table 4).

By using the triple attributes segmentation mechanism and applying a clustering method the current data is changed to 20 different sets. At this stage to implement EN-SAX Method, it is needed to map each set of data to a special symbol. There are 20 clusters for data set (cluster_0~cluster_19). Each segment is mapped to relate symbol based on Min, mean and Max values. The mean value is calculated for each segment that indicates the domain of each cluster which maps to the related symbol.

Table 4: Average error rate between original and symbolized segmented data by EN-SAX

Cluster	Symbol	DS1	DS2	DS3	DS4	DS5
0	A	0.59	0.15	0.22	0.62	0.27
1	B	0.86	0.51	0.23	0.66	0.84
2	C	0.84	0.58	0.15	0.58	0.81
3	D	0.46	0.92	0.89	0.59	0.35
4	E	0.59	0.81	0.81	0.66	0.23
5	F	0.42	0.63	0.73	0.39	0.31
6	G	0.82	0.27	0.22	0.69	0.83
7	H	0.89	0.38	0.71	0.62	0.46
8	I	0.85	0.21	0.78	0.55	0.57
9	J	0.76	0.64	0.75	0.63	0.92
10	K	0.81	0.07	0.88	0.65	0.54
11	L	0.55	0.45	0.48	0.60	0.46
12	M	0.07	0.18	0.78	0.54	0.38
13	N	0.14	0.32	0.28	0.59	0.75
14	O	0.20	0.70	0.75	0.70	0.46
15	P	0.88	0.76	0.22	0.83	0.87
16	Q	0.46	0.88	0.28	0.68	0.57
17	R	0.23	0.92	0.84	0.53	0.57
18	S	0.82	0.31	0.50	0.76	0.54
19	T	0.70	0.01	0.22	0.62	0.43

Similarity measurement: Based on earlier steps, original exchange data is transferred to new symbolize form (EN-SAX) that is based on SAX Method which is a common method to representation time series data. The next step is to measure similarity between original data and new symbolized data. It is performed by calculating the error rate between the two sets of original and new data.

There are two sets of vectors for each record of data which one of them is a vector as a record from segmented data with related symbol after clustering and second one is another single vector that shows mean value for related symbol. In order to measure similarity between two vectors the cosine of the angle between them is measured that is called cosine similarity. In the field of data mining, it is used to measure cohesion within clusters. Cosine similarity between two vectors determines whether two vectors are pointing in roughly the same direction. Table 5 shows average error rate between original and symbolized segmented data using EN-SAX. Using mean value within SAX Method, similarity measurement can be done by calculating error rate between the mean values and the original values.

Comparison results between SAX and EN-SAX: The result of error rate for our data using SAX and EN-SAX shows that the accuracy of EN-SAX is better than SAX. The following results are obtained by calculating average error rate values between all segments and symbols at each related data set. Calculating error rate between SAX and EN-SAX shows that in a large amount of time series data using some additional values causes decrease of error rate between original data with segmented data. Table 6 shows the result of experiment.

Table 5: Average error rate between original and symbolized segmented data by EN-SAX

Symbol	DS1	DS2	DS3	DS4	DS5
A	0.77	0.68	0.79	0.83	0.71
B	0.84	0.43	0.34	0.96	0.37
C	0.45	0.39	0.62	0.65	0.83
D	0.53	0.85	0.16	0.95	0.91
E	0.56	0.28	0.23	0.89	0.79
F	0.38	0.34	0.48	0.40	0.31
G	0.43	0.09	0.91	0.88	0.12
H	0.28	0.41	0.73	0.85	0.78
I	0.71	0.13	0.81	0.78	0.14
J	0.65	0.47	0.86	0.68	0.83
K	0.58	0.63	0.59	0.76	0.89
L	0.63	0.38	0.33	0.84	0.34
M	0.74	0.76	0.42	0.82	0.67
N	0.82	0.89	0.54	0.63	0.22
O	0.56	0.64	0.37	0.24	0.36
P	0.96	0.37	0.74	0.58	0.83
Q	0.85	0.28	0.92	0.79	0.24
R	0.92	0.89	0.81	0.82	0.93
S	0.77	0.19	0.79	0.78	0.81
T	0.42	0.37	0.28	0.13	0.45

Table 6: Error rate of SAX and EN-SAX

Data set	SAX	EN-SAX
DS1	0.64	0.59
DS2	0.47	0.42
DS3	0.58	0.54
DS4	0.71	0.63
DS5	0.57	0.54

Table 7: Linear regression and SVM error rate for EN-SAX

Data set	Linear regression	SVM
DS1	0.42	0.46
DS2	0.51	0.52
DS3	0.45	0.48
DS4	0.61	0.67
DS5	0.49	0.52

Prediction error rate on discretized data: There are some algorithms that can be used to predict data. At first the data is prepared using SAX and EN-SAX that is read from related file and then by applying linear regression (Montgomery *et al*., 2001) and SVM (Lee and Huang, 2007) methods prediction is done.

The results are shown in Table 7. The comparison between performance of Linear Regression (LR) and Support Vector Machine (SVM) Methods shows that the linear regression has better performance for prediction on the exchange data sets (i.e., financial time series data). The results are obtained by determining similarity between the real segment values and predicted values using each method. Table 7 shows the prediction error rate using linear regression and SVM Methods.

Statistical results: The results are compared using t-test as statistical test to show improving the results by using EN-SAX as representation method and linear regression as classification method for prediction. There is an α value in t-test which is selected 0.05 at this test ($\alpha = 0.05$). It means if the final p-value be less than α value, the

Table 8: t-test (SAX vs. EnSAX) and (LR vs. SVM)

					95% confidence interval of the difference	
Phase	t	df	p-value	Mean difference	Lower	Upper
SAX vs. En-SAX	5.9761	4	0.0039	0.008	0.0268	0.0732
LR vs. SVM	4.1851	4	0.0139	-0.034	-0.0566	-0.0114

probability of this results is 95% in using similar method on this kind of data. The t-test results are shown in Table 8.

The two-tailed p value between SAX and EN-SAX data sets equals 0.0039 which by conventional criteria, this difference is considered to be very statistically significant. The two-tailed p value between LR and SVM data sets equals 0.0139 which by conventional criteria this difference is considered to be statistically significant.

As it is discussed in earlier study, EN-SAX Method is more effective in comparison with other dimensionality reduction methods such as SAX. The features of EN-SAX include keeping important patterns on financial time series data and avoiding missing key values which have significant role on decision for future predictions. Using EN-SAX is helpful in almost all types of time series data, especially in the financial time series data.

CONCLUSION

This study proposes an improved method called Enhanced Symbolic Aggregate Approximation (EN-SAX) for symbolic representation of time series financial data. It uses the segmentation and mean value representation for financial time series data. The results show that EN-SAX produces more accurate results and is effective in using the represented data for processing and prediction purposes in time series financial data.

REFERENCES

Abu Bakar, A., A.M. Ahmed and A.R. Hamdan, 2010. Discretization of time series dataset using relative frequency and K-nearest neighbor approach. Proceedings of the 6th International Conference on Advanced Data Mining and Applications, November 19-21, 2010, Chongqing, China, pp: 193-201.

Aref, W.G., M.G. Elfeky and A.K. Elmagarmid, 2004. Incremental, online and merge mining of partial periodic patterns in time-series databases. IEEE Trans. Knowledge Data Eng., 16: 332-342.

Buu, H.T.Q. and D.T. Anh, 2011. Time series discord discovery based on iSAX symbolic representation. Proceedings of the 3rd International Conference on Knowledge and Systems Engineering, October 14-17, 2011, Hanoi, Vietnam, pp: 11-18.

Fu, T.C., 2011. A review on time series data mining. Eng. Appl. Artif. Intell., 24: 164-181.

Keogh, E., S. Chu, D. Hart and M. Pazzani, 2004. Segmenting Time Series: A Survey and Novel Approach. In: Data Mining in Time Series Databases, Last, M., A. Kandel and H. Bunke (Eds.). Chapter 1. World Scientific, Singapore, ISBN-13: 9789812382900, pp: 1-22.

Keogh, E.J. and M.J. Pazzani, 2000. A simple dimensionality reduction technique for fast similarity search in large time series databases. Proceedings of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining: Current Issues and New Applications, April 18-20, 2000, Kyoto, Japan, pp: 122-133.

Lee, Y.J. and S.Y. Huang, 2007. Reduced support vector machines: A statistical theory. IEEE Trans. Neural Networks, 18: 1-13.

Lkhagva, B., Y. Suzuki and K. Kawagoe, 2006a. Extended sax: Extension of symbolic aggregate approximation for financial time series data representation. Proceedings of the 4th Annual Meeting of the Database Society of Japan 17th Data Engineering Workshop of Electronics, Information and Communication Engineers, March 1-3, 2006, Okinawa Convention Center, Japan.

Lkhagva, B., Y. Suzuki and K. Kawagoe, 2006b. New time series data representation ESAX for financial applications. Proceedings of the IEEE 22nd International Conference on Data Engineering Workshops, April 3-7, 2006, Atlanta, GA., USA, pp: x115--x115.

Megalooikonomou, V., Q. Wang, G. Li and C. Faloutsos, 2005. A multiresolution symbolic representation of time series. Proceedings of the 21st International Conference on Data Engineering, April 5-8, 2005, Tokyo, Japan, pp: 668-679.

Montgomery, D.C., E.A. Peck and G.G Vining, 2001. Introduction to Linear Regression Analysis. 3rd Edn., John Wiley and Sons, New York, USA., ISBN-13: 9780471315650, Pages: 641.

Tan, P.N., M. Steibach and V. Kumar, 2006. Introduction to Data Mining. Pearson Addison Wesley, Boston, MA., USA., ISBN-13: 9780321420527, Pages: 769.

Yi, B.K. and C. Faloutsos, 2000. Fast time sequence indexing for arbitrary L_p norms. Proceedings of 26th International Conference on Very Large Data Bases, September 10-14, 2000, Cairo, Egypt, pp: 385-394.