Journal of Theoretical and Applied Information Technology
31st October 2016. Vol.92. No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

KLANG VALLY RAINFALL FORECASTING MODEL USING TIME SERIES DATA MINING TECHNIQUE

ZULAIHA ALI OTHMAN, NORAINI ISMAIL, ABDUL RAZAK HAMDAN, MAHMOUD AHMED SAMMOUR

Data Mining and Optimization Group, Center for Artificial Intelligence Technology,
Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia,
Bangi, Selangor, Malaysia

E-mail: zao@ukm.edu.my, norainismail14@yahoo.com, arh@ukm.edu.my, mahmoud.samour@gmail.com

ABSTRACT

Rainfall has influence the social and economic activities in particular area such as agriculture, industry and domestic needs. Therefore, having an accurate rainfall forecasting becomes demanding. Various statistical and data mining techniques are used to obtain the accurate prediction of rainfall. Time series data mining is a well-known used for forecast time series data. Therefore, the objective of this study is to develop a distribution of rainfall pattern forecasting model based on symbolic data representation using Piecewise Aggregate Approximation (PAA) and Symbolic Aggregate approXimation (SAX). The rainfall dataset were collected from three rain gauge station in Langat area within 31 years. The development of the model consists of three phases: data collection, data pre-processing, and model development. During data pre-processing phase, the data were transform into an appropriate representation using dimensional reduction technique known as Piecewise Aggregate Approximation (PAA). Then the transformed data were discretized using Symbolic Aggregate approXimation (SAX). Furthermore, clustering technique was used to determine the label of class pattern during preparing unsupervised training data. Three type of pattern are identified which is dry, normal and wet using three clustering techniques: Agglomotive Hierarchical Clustering, K-Means Partitional Clustering and Self-Organising Map. As a result, the best model has be able to forecast better for the next 3 and 5 years using rule induction classification techniques.

Keywords: *Time Series Data Mining, Clustering, Classification, Time Series Symbolic Representation, Rainfall Forecasting*

1. INTRODUCTION

The main characteristic of rainfall in Malaysia is predictable during dry and wet season. Quantity of rainfall supply will cause natural disasters such as drought and floods. Thus rainfall forecasting becomes a significant factor in agricultural countries like Malaysia.

Based on historical rainfall records in Federal Territory, water crisis occurred in Malaysia in February 1998 when three reservoir dams in Klang Valley which is Klang Gates Dam, Batu Dam and Semenyih Dam suffered a substantial drop in water level following the El Niño phenomenon [18]. The subsequent water shortage affected almost all the residents in the Klang Valley causing the government to impose water rationing. The shortage was blamed on El Nino despite actual rainfall in the months leading up to February 1998 in Federal Territory not being significantly below average. In fact in November 1997, Klang Gates Dam had its highest recorded rainfall. Similarly in October 1997 the Kajang station not far from the Semenyih dam had its highest rainfall in record [18]. Thus, accurate information on rainfall is essential for the planning and management of water resources.

Additionally, in the urban and suburbs areas like Klang Valley, rainfall has a strong influence on traffic, sewer systems, and other human activities. Nevertheless, rainfall is one of the most complex and difficult elements to understand and to model due to the complexity of the atmospheric processes that generate rainfall and the tremendous range of variation over a wide range of scales both in space and time [4]. Thus, accurate rainfall forecasting is one of the greatest challenges, despite many advances in weather forecasting in recent decades [3]. Therefore, this paper presents a new approach based on symbolic data representation using

Piecewice Aggregate Approximation (PAA) and Symbolic Aggregate approximation (SAX) technique to improve the distribution of rainfall forecast performance in Klang Valley.

Recently, data mining on AI technique is well known used for forecasting based on past data. Rainfall data is presented in the form of a linear graph, which is limited to discover the understandable patterns because the data is visualized based on the values in the time series. However, the pattern can be gained depends on the method used in the data representation for example by using data symbolization. Past studies have obviously indicated that PAA and SAX [1] is a good approach to change data into symbolic representation and has a high potential to provide an accurate information, so that it can improve the accuracy of rainfall forecasting. The algorithm presents a robust symbolic representation gained from the slope of information to generate many possible patterns.

On the other hand, Cluster analysis is a multivariate method which aims to classify a sample of subjects (or objects) on the basis of a set of measured variables into a number of different groups such that similar subjects are placed in the same group. A number of different measures have been proposed to measure ’distance’ for binary and categorical data. For interval data, the most common distance measure used is the Euclidean distance. The rainfall pattern can be defined and can be used as an interval value of rainfall time series data.

The experiment were conducted using daily rainfall dataset collected at 3 rain gauge stations in Langat, Malaysia: JPS Ampang, Jalan 6 Kaki and Setul Mantin. The purpose of this experiment is to extract the useful rainfall pattern over a specified region in our state and discover new knowledge.

2. RELATED WORK

Accurate and timely weather forecasting is a major challenge for the scientific community. Rainfall prediction modeling involves a combination of computer models, observation and knowledge of trends and patterns. Using these methods, reasonably accurate forecasts can be made up. Several recent research studies have developed rainfall prediction using different weather and climate forecasting methods. Many of them are using Neural Network classification technique and regression model. On the other hand, time series data mining technique is one of data mining area that well known used to forecast time series data using either indexing, clustering, classification, summarization and anomaly detection [7, 8]. Recently, many researcher have been studying on time series data sets such as rainfall. Time series is a sequence of real numbers, each number representing a value at a time point [16].

Time series databases are often extremely large and exist in high dimensional form. Data dimensionality reduction aim to mapping high dimensional patterns onto lower-dimensional patterns. The techniques used for dimensionality reduction can be classified into two: (1) feature extraction and (2) feature selection. Feature selection is a process that selects a subset of original attributes. While feature extraction techniques extract a set of new features from the original attributes through some functional mapping. For time series mining community, feature extraction is much popular than feature selection [15]. Many representation for time series data have been proposed such as Singular Value Decomposition (SVD), Discrete Fourier Transform (DFT), Discrete Wavelet Transform (DWT), Piecewise Linear Representation, Piecewise Aggregate Approximation (PAA), Regression Tree and Symbolic Representation (SAX). For choosing the appropriate dimensional reduction of the features is a challenging problem. Based on the PAA dimensional reduction and the normality aggregated values, the SAX technique have been proposed [16]. SAX is one of the discretization methods designed especially for time series data. The temporal aspect of the data is only taken into account by the preprocessing step performing via PAA. The formula shown below explains the PAA dimensional reduction representation method where n is the length of a time series C, w is the number of PAA segment and $\overline{C_i}$ is the average value of the ith segment.

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

The PAA dimensionality reduction is intuitive and simple, compared to other dimensional reduction such as DFT and DWT [7]. To transform a time series data into PAA, further transformation can be applied to obtain a discrete symbolic representation. It is desirable to have a discretization technique that will produce symbols with equiprobablity [7, 16].

Furthermore, there are two main objectives when dealing with time series data: (1) prediction of future patterns based on past data and (2) description of time series data [9]. For time series prediction, clustering can be applied either as a preprocessing in complex algorithms or as a prediction model in its own right [10, 11]. Clustering technique commonly used as an unsupervised learning process for partitioning in time series data such as cluster the similarity or dissimilarity using agglomerative hierarchical clustering, k-means and self-organizing maps (SOM) [12, 13].

Similarity search is very useful as a tool for exploring time series databases .Time series similarity mining includes similarity searching on univariate time series (one attribute) and multivariate time series (more than one attribute). Generally, to compute the similarity of two multivariate time series data directly is more difficult compared to compute the similarity of two univariate time series data. The similarity between two time series is usually measured by a distance function such as Euclidean distance, Pearson correlation coefficient, Dynamic Time Warping (DTW) and others [14]. Lin et. al. proposed a MINDIST function that return the minimum distance between the original time series of two words [16]. Nevertheless, Euclidean Distance is the most common distance measure used for time series data [14]. The formula 2 below explain two time series in numeric discretization form, Q and C of the same length n defines their Euclidean distance.

$D(Q,C)=[(d11-d21)2+(d12-d22)2+..+(d1L-d2L)2]1/2$ (2)

3. METHODOLOGY

In general, the methodology in this experiment involves three phases: (1) data collection, (2) data pre-processing and (3) model development.

3.1 Data Collection

The daily rainfall data were collected from 3 rain gauge stations in Langat area (southwest of Peninsular Malaysia), as shown in Table 1. The data were selected based on the completeness of the data and the longer period of records.

Table 1: Geographic Coordinate and Period of Records for The Selected Rain Gauged Station.

Rain gauge station	Period of records	Latitude	Longitude
17R-JPS Ampang	1953 - 1998	3° 13° N	101° 53° E
25R-Jalan 6 Kaki	1965 - 1998	3° 13° N	101° 53° E
32R-Setul Mantin	1965 - 1998	2° 48° N	101° 55° E

Peninsular Malaysia is located in the Northern latitude between 1° and 6° N and the Eastern longitude from 100° to 103° E in the equatorial zone. The climate of Peninsular Malaysia is influenced by the Southwest monsoon (May – September) and the Northeast monsoon (November – Mac). The Southwest monsoon is a drier period for the whole country, while during the Northeast monsoon, the east coast and northern areas of Peninsular Malaysia receive more heavy rains than the other parts of the country. The aim of this study is to forecast the rainfall pattern during dry season (Jun-August) in southwest region of Peninsular Malaysia.

3.2 Data Pre-processing

Working with time series data is very challenging due to its large dataset and need to dealing with subjectivity things. Therefore, we need a representation that can mine the time series data efficiently. This section focus on the preprocessing framework of rainfall time series data. This section discussed four main steps: (1) data normalization, (2) dimensionality reduction via PAA, (3) data discretization using SAX and (4) cluster using Euclidean distance measure.

Time series data can be very long, where t is the time index and n is the number of observations in the series. Commonly, data miners confine their interest to subsections of the time series, which are called subsequences. Given a time series T of length m, a subsequence C of T is a sampling of length n < m of contiguous position from T, that is, C = tp,…,tp+n-1 for 1≤ p ≤ m – n + 1. In this study, we extracted subsequences of a rainfall univariate time series data collected from three stations for 91 days (June-August), refer Table 2.

Table 2: A Rainfall Data Collected in Langat Area From 1953- 2003.

	Hari	1953	1954			1998	1999
Jan	1	27.6	6.6	…..	…..	8.5	0
	2	0	5.8	…..	…..	0	0
	.	.	.	…..	…..	.	.
Dec	1	38.3	1.2	…..	…..	0.5	0.5
	2	0	0	…..	…..	34.5	4
	.	.	.	…..	…..	.	.
	.	.	.	…..	…..	.	.
	30	0.5	10.9	…..	…..	20.5	0
	31	12.9	0	…..	…..	5	14

The raw data were normalized before applying dimensionality reduction to have mean of zero and a standard deviation of one. Then, the large and small values of time series data were coordinated to scale the values, as shown in Table 3.

Table 3: Transform Raw Data to Normalization Value.

	Days					
Patterns	1	2	3	…..	90	91
C1	-0.466	-0.466	-0.466	…..	-0.466	-0.163
C2	0.302	-0.397	-0.440	…..	-0.266	-0.440
C3	-0.427	0.533	4.158	…..	0.249	-0.427
…..	…..	…..	…..	…..	…..	…..
…..	…..	…..	…..	…..	…..	…..
C92	-0.435	-0.435	3.140	…..	-0.435	-0.435
C93	-0.464	-0.464	-0.464	…..	-0.464	0.009

The basic idea of PAA is to represents the time series as a sequence of rectangle basis functions and converts the time series into a discrete symbolic sequence. To reduce the rainfall time series data dimensionality from n dimension to w dimension, the data need to be divided into w equal 13 weeks, as provided in Table 4. The mean value for 7 days of the rainfall data falling in each segments and vector of these values becomes dimension reduction.

Table 4: Dimension Reduction Process Transform Normalization to PAA Value.

	Weeks				
Patterns	1	2	…..	12	13
C1	-0.466	0.260	…..	0.178	-0.139
C2	0.477	-0.066	…..	-0.409	-0.322
C3	1.513	-0.330	…..	-0.041	-0.330
…..	…..	…..	…..	…..	…..
…..	…..	…..	…..	…..	…..
C92	0.327	0.087	…..	0.494	1.099
C93	-0.244	-0.464	…..	-0.464	0.348

For converting PAA value to symbol discretization (SAX), we can simply determine the "breakpoints" that will produce an equal-sized areas under Gaussian curve. For example, Table 5 gives the breakpoints for value of *a* from 3 to 5.

Table 5: Table Breakpoint for Gaussian Distribution.

β \ a	3	4	5
$\beta 1$	-0.43	-0.67	-0.84
$\beta 2$	0.43	0	-0.25
$\beta 3$		0.67	0.25
$\beta 4$			0.84

There are 3 characteristics of rainfall were defined: above normal, normal, below normal [2]. In this study, we choose 3 symbols (a, b, c). All PAA coefficients breakpoint that are below and equal -0.43 (low rainfall below normal) are transformed to symbol "a", breakpoint that less than 0.43 and greater than -0.43 (normal rainfall) are transformed to symbol "b", while breakpoint that more than and equal 0.43 (heavy rainfall above normal) are transformed to symbol "c", as shown in Figure 1. Now, the symbolic data representation is defined as unsupervised training data, as shown in Table 6 below.

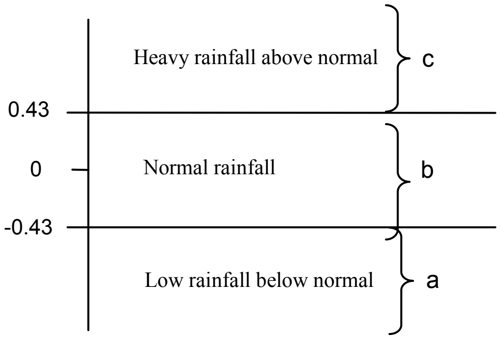

Figure 1: Rainfall Category Based on Breakpoint of Gaussian Distribution

Table 6: Discretization Process Transform PAA Value to SAX Representation.

	Attributes				
Patterns	**W1**	**W2**	**.....**	**W12**	**W13**
	1	**2**	**.....**	**12**	**13**
C1	'a'	'c'	.....	'b'	'b'
C2	'c'	'b'	.....	'a'	'a'
C3	'c'	'a'	.....	'b'	'a'
.....	.....	.....	.....	.....	.....
C92	'c'	'b'	.....	'c'	'c'
C93	'a'	'a'	.....	'a'	'c'

Then, the rainfall data were clustered using three clustering techniques: agglomerative hierarchical clustering, k-means and SOM to get the rainfall pattern. For measuring similarity or dissimilarity of the patterns, Euclidean distance measure is used. However, the discretization symbol must be converted to numeric values to allow Euclidean distance calculate the distance measure between two time series patterns. For example symbol “a” convert to “1”, symbol “b” convert to “2” and symbol “c” convert to “3”. The patterns produced can be grouped as dry, normal and wet.

The clustered result were set as a class labels of the dataset as shown in the table below. Table 7 shows three samples of training data produced by using agglomerative hierarchical clustering (H3), k-means technique (P3), and SOM technique (S3).

Table 7: Result of Class Label Patterns Using Unsupervised Learning.

a) Dataset H3 : A result class label using agglomerative hierarchical clustering

Patterns	W1	W2	…	W12	W13	Category
C1	c	a	...	a	c	normal
C2	c	b	...	c	b	dry
C3	a	a	...	b	b	normal
...	...	...	...	...	...	...
...	...	...	...	...	...	...
C92	b	c	...	b	b	wet
C93	b	a	...	c	c	normal

b) Dataset P3 : A result class label using k-means clustering

Patterns	W1	W2	…	W12	W13	Category
C1	c	a	…	a	c	normal
C2	c	b	…	c	b	normal
C3	a	a	…	b	b	normal
…	…	…	…	…	…	…
…	…	…	…	…	…	…
C92	b	c	…	b	b	wet
C93	b	a	…	c	c	normal

c) Dataset S3 : A result class label using SOM clustering

Patterns	W1	W2	…	W12	W13	Category
C1	c	a	…	a	c	normal
C2	c	b	…	c	b	normal
C3	a	a	…	b	b	wet
…	…	…	…	…	…	…
…	…	…	…	…	…	…
C92	b	c	…	b	b	dry
C93	b	a	…	c	c	dry

3.3 Model Development

3.3.1 Training Data

In order to generate the best forecasting model, the data need to be trained based on past observation to estimate the future pattern. For the experimental setup, H3, P3 and S3 is used as a training dataset. Each dataset have 93 patterns, 13 attributes and 3 category of pattern.

In this experiment, four classification algorithm were selected which is J48 and ADTree (decision tree algorithm), PART and JRip (rule based algorithm). The classification techniques is used to compare how good the clustered results fit with the data labels and built the best rainfall forecasting model by using WEKA (Waikato Environment for Knowledge Analysis). All the identified algorithms are tested to each training set by using 10-fold cross-validation method. The model that has high accuracy, minimum number of rules and minimum number of training data were chosen as the best classification model.

From the experiment, Table 8 shows the number of rules and size of trees generated by the selected algorithms. For rule algorithms, JRip produce between 2-3 rules, while PART produce between 17-20 rules. For tree algorithms, ADTree always produces a fixed size of tree (with 31 nodes), while J48 produce between 31-46 nodes. The result conclude that JRip and ADTree algorithms is more stable compared to PART and J48.

Table 8: Number of Rules or Size of Trees Produced by The Algorithms.

Dataset	JRip	PART	J48	ADTree
H3	2	20	31	31
P3	3	17	40	31
S3	3	18	46	31
Average	3	18	39	31

Table 9: Accuracy Produced by The Algorithms.

Dataset	JRip	PART	J48	ADTree
H3	66.67%	88.89%	77.78%	58.82%
P3	57.14%	62.16%	52.63%	67.86%
S3	77.78%	67.86%	57.14%	77.78%
Average	67.20%	72.97%	62.52%	68.15%

Table 9 above shows the summary of accuracy (percentage of correctly classified instances) result produced by the rule and tree algorithms. For the rule algorithms, it was observed that the PART algorithm is more accurate (having average 72.97%) compared with JRip algorithms with an average of 67.20%. While for the tree algorithms, ADTree algorithm is more accurate (having average of 68.15%) compared with J48 algorithms with average of 62.52%. Meanwhile, the highest accuracy are dominated by the PART, while its sizes are bigger then JRip algorithms. However, the best model is based on the highest accuracy and smallest size of rule. From the results, we can conclude that JRip algorithms produced the best model prediction for rainfall dataset because it’s having high accuracy and shortest rules. The extracted rule generated from the best model shown as below:

IF W4=c and W8=b => category=normal

IF W6=c and W4=c => category=dry

IF W1=a and W2=c => normal=normal

IF W7=c => category =wet

Table 9 also shows that S3 (SOM) clustering has presented the highest accuracy which is 77.78% for both decision tree and rule classification techniques.

3.3.2 Testing Data

The best model is used to predict the new data collected from six different stations in Langat area, as provided in Table 10.

Table 10: There are 6 Rain Gauged Station Selected for New Data Testing.

Rain gauge station	Period of records	Latitude	Longitude
17R-JPS Ampang	1999 - 2003	3° 13° N	101° 53° E
25R-Jalan 6 Kaki	1999 - 2003	3° 13° N	101° 53° E
32R-Setul Mantin	1999 - 2003	2° 48° N	101° 55° E
5R-Mardi Serdang	1999 - 2003	2° 59° N	101° 40° E
8R-Ampangan Semenyih	1999 - 2003	3° 04° N	101° 52° E
51R-UKM	1999 - 2003	2° 56° N	101° 47° E

Table 11 below show the prediction result based on test data. The model using JRip algorithm able to predict rainfall distribution for next 3 years period (having accuracy prediction of 88.89%) and 5 years period (having accuracy prediction of 70.00%).

Table 11: Accuracy Produced by The Classification Prediction Model on The New Data.

3 Years (1999-2001)	5 Years (1999-2003)
88.9 %	70.00%

4. RESULT AND DISCUSSION

The main objective of this study is to develop a distribution of rainfall pattern forecasting model based on symbolic data representation using PAA and SAX.

As a result, SOM clustering technique show the highest accuracy which is 77.78% for both decision tree and rule induction classification compared to agglomerative hierarchical and K-means clustering technique. From this result, we can conclude that SOM clustering technique able to determine the pattern of class label more correctly. This is because SOM clustering technique can apply competitive learning as opposed to error-correction learning and use a neighborhood function to preserve the topological properties of the input space [6]. This makes SOMs useful for visualizing low-dimensional views of high-dimensional data, akin to multidimensional scaling.

The result also shows that rainfall dataset seems to work better with rule based algorithms. We believe that its way of skewing the example distribution has different effects on divide-and-conquer decision tree learning algorithms and on separate-and-conquer rule based algorithms. The nature of this difference lies in the way a condition is selected. Typically, a separate-and-conquer rule based selects the test that maximizes the number of covered positive examples and at the same time minimizes the number of negative examples that pass the test. It usually does not pay any attention to the examples that do not pass the test [9]. Based on result, even PART rule based algorithm is more accurate than Jrip, but JRip algorithms is more stable because the rule generated is more shorter compared to PART.

The result shows that best model is Jrip algorithm. Then the testing phase was run to validate performance of the model. From the test result, the model using JRip algorithm able to predict rainfall distribution in Klang Valley for next 3 years and 5 years period.

5. CONCLUSION

This paper has proposed a method for rainfall forecasting model. In preprocessing phase, the suitable time series data representation based on symbolic data representation were used which later show the best techniques for forecasting rainfall in Klang Valley. Therefore, dimensionality reduction using PAA and SAX was used to provide three types of information, with each type being useful for a particular decision. Furthermore, clustering technique was used to determine the label of class pattern during preparing unsupervised training data.

The preprocessing step employed various approaches to obtain the appropriate form of data that fit to the classifier. The experimental results shows that JRip algorithms approach is very promising and provide valuable set of interpretable knowledge. The study has proved that the time series data mining technique is able to produce good model to forecast the distribution of rainfall pattern accurately.

REFRENCES:

[1] Ghassan Saleh Al-Dharhani, Zulaiha Ali Othman, Azuraliza Abu Bakar and Sharifah Mastura Syed Abdullah, "Fuzzy-based shapelets for mining climate change time series patterns", *Advances in Visual Informatics. Lecture Notes in Computer Science*, 2015, Vol. 9423, pp. 38-50.

[2] Bambang Widjonarko Otok and Suhartono, "*Development of rainfall forecasting model in Indonesia by using ASTAR, Transfer Function and Arima Methods", European Journal of Scientific Research*, 2009, Vol. 38(3), pp. 386-395.

[3] M. Kannan et. al., "Rainfall forecasting using data mining technique*", International Journal of Engineering and Technology,* 2010, Vol. 2(6), pp. 397-401.

[4] Mohammad Valipour, "Number of Required Observation Data for Rainfall Forecasting According to the Climate Conditions", *American Journal of Scientific Research*, 2012, Vol. 74, pp.79-86.

[5] Mohammed Hasan Abdulameer, Siti Norul Huda Sheikh Abdullah and Zulaiha Ali Othman, "Neural Gen Feature Selection for Supervised Learning Classifier", *Research Journal of Applied Sciences, Engineering and Technology,* 2014, Vol. 7(15), pp. 3181-3187.

[6] J. Fernando et. al., "Analysis of clustering and selection algorithms for the study of multivariate wave climate", *Coastal Engineering*, 2011, Vol. 58, pp. 453–462.

[7] Mustafa Gokce Baydogan and George Runger, "Learning a symbolic representation for multivariate time series classification", *Data Mining and Knowledge Discovery*, 2015, Vol. 29, pp. 400-422.

[8] Peiman Mamani Barnaghi, Azuraliza Abu Bakar and Zulaiha Ali Othman, "Enhanced Symbolic Aggregate Approximation Method for Financial Time Series Data Representation", *Journal of Soft Computing*, 2012, Vol. 8(4), pp. 261-268, Indexed Scopus.

[9] Bing Hu, Yanping Chen and Eamonn Keogh, "Time Series Classification under More Realistic Assumptions", *Proceedings of the 2013 SIAM International Conference on Data Mining*, 2013.

[10] D. Houtao, "A time series forest for classification and feature extraction", *Information Sciences: an International Journal,* 2013, Vol. 239, pp.142-153.

[11] Swasti Singhal and Monika Jena, "A Study on WEKA Tool for Data Preprocessing, Classification and Clustering", *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, 2013, Vol. 2, pp. 250-253.

[12] V. Kavitha and M. Punithavalli, "Clustering Time Series Data Stream - A Literature Survey", *International Journal of Computer Science and Information Security (IJCSIS*), 2010, Vol. 8.

[13] C. H. Li et. al., "LDA-Based Clustering Algorithm and Its Application to an Unsupervised Feature Extraction", *IEEE Transactions on Fuzzy Systems,* 2011, Vol. 19.

[14] A. Camerra et. al., "iSAX 2.0: Indexing and Mining One Billion Time Series", *IEEE 10th International Conference on Data Mining (ICDM)*, 2010, pp. 58 - 67.

[15] J. Young-Seon, "Weighted dynamic time warping for time series classification", *Computer Analysis of Images and Patterns*, 2011, Vol. 44, pp. 2231-2240.

[16] H. T. Q. Buu and D. T. Anh, "Time Series Discord Discovery Based on iSax Symbolic Representation", *Third International Conference on Knowledge and Systems Engineering (KSE)*, 2011, pp. 11 - 18.

[17] Hamidah Jantan, Abdul Razak Hamdan and Zulaiha Ali Othman, "Talent Knowledge Acquisition Using Data Mining Classification Techniques", *3rd Conference on Data Mining and Optimization (DMO2011)* IEEE Explore, 2011, pp. 32-37.

[18] Thomas Fuller, "A news from 1998 about the Klang Valley water crisis: Month of water rationing leave a lot of throats parched", *The New York Times News*, 1998.

[19] By Rui Xu and Don Wunsch, "Clustering", *IEEE Computational Intelligence Society, Sponsor*,