Journal of Engineering and Applied Sciences 12 (Special Issue 10): 8844-8850, 2017
ISSN: 1816-949X

An Optimal Stream Prediction Using Adaptive Regression Neural Network

[1]Nabeel Al-A'araji, [2]Eman Al-Shamery and [2]Alyaa Abdul-Hussein
[1]Ministry of Higher Education, Apparatus of Supervision and Scientific Evaluation, Baghdad, Iraq
[2]Department of Software, Faculty of Information Technology, University of Babylon, Hillah, Iraq

Abstract: Data stream is concerned in industry engineering, finance, economy, traffic and many other fields. The main challenging problems in stream are changed stream with time, time of data arrival and space required for storage stream. Prediction in stream is used to forecast the new data from available data. An Adaptive Regression Neural Network (ARNN) is proposed as a new model in this study based on regression neural network with feedback which it is added to the time variable to make model adaptive for the prediction process with minimum error and high accuracy. The proposed system consists of three main stages, data comes from complex environments may be noisy, redundant, contain outliers and missing values. Thus, the polynomial regression with the segmentation and thresholding are employed for preprocessing stage in the interpolation of missing values and remove outlier points from data stream. The ARNN represents the second and main stage for prediction problem. The evaluation process represents the final stage. The proposed method is compared with traditional regression method for prediction and the results show that the proposed method indicates better accuracy.

Key words: Interpolation, polynomial regression, stream prediction, Adaptive Regression Neural Network (ARNN), traditional regression, thresholding

INTRODUCTION

Stream management is used to process a large amount of time-varying data effectively and is applied in some fields such as analysing, network monitoring, telecommunication management and sensor network. In research for stream management, continuous query and stream mining are concerned mostly but stream prediction is paid less attention relatively. Stream is considered one kind of time series, stream is arrived multiple, rapid, time-varying and unbounded, so, stream prediction is needed to update with time to satisfy the data changing (Kong *et al*., 2008).

Stream prediction is involved predicting the system behaviour in future based on information of the current and the past status of the system (Agarkar *et al*., 2016). Stream prediction is used in many types of situation. For example, a global positioning system in a car is not send only the current position of the car but also, its vector of movement or expected trajectory. Similarly, the values measured by a temperature sensor in home are not hoped for change in normal conditions. In general, predictions are used mostly when the values are expected to change according to predefined patterns or when the interest is required in the detection of unexpected changes.

Prediction methods are classified into two types time series methods and regression methods. A time series is defined as a sequence of data points, consisting of observations made over a time interval and ordered in time. Each observation is represented as x(t) where the observed value x is indexed by the time (t) in which it is made. These predictions are represented as a function of the past observations and their respective time model usually is defined by parameters calculated using past observations. Time series methods are explained in this study.

Naive approaches are assumed that future values are computed in constant time, given the historical data. They are considered simple but these methods are became more inaccurate and imprecise when data is high variability.

Autoregressive (AR) is used to represent the expected value of a variable given the values from a set of correlated variables. It is considered very efficient for short-term because it is differentiated from naive prediction less sensitivity against the data noise. Long-term prediction using AR is tended to be inaccurate, due to uncertainty about the order of the model, its coefficients and unobserved errors.

Moving Average (MA) method is used for events that are influenced a values observed at time (t), also, it is used to remove casual noise from the data, given that its

Corresponding Author: Eman Al-Shamery, Department of Software, Faculty of Information Technology, University of Babylon, Hillah, Iraq

prediction is considered less sensitive against outliers than the prediction which it is made using AR Models. The model is used to calculate the parameters by linear least squares and iterative non-linear fitting procedures which is computationally more complex than the AR alone.

Exponential Smoothing (ES) Exponentially Weighted Moving Average (EWMA) is the simplest version. The value is predicted for the time (t+i) and it is calculated using only the most recent observation and the most recent forecast. The space and time complexities are looked smaller than the AR and MA methods, it has some of the same limitations such as the low efficiency its confidence intervals increased exponentially.

Autoregressive Integrated Moving Average (ARIMA), it is considered a stationary process. It is composed by the combination of an AR and a MA Models. It has higher accuracy than the previous methods. The time and space complexities are bounded by the worst complexity between the AR and the MA methods (Dias *et al*., 2016).

Regression method is the second approach for prediction. Instead of relying only on past values to make prediction in the previous approach, the measurements are predicted based on other types. A given a value is observed by one sensor node, a regressive model is used to predict value which can be observed by another sensor node. The advantage of regression is combined different data types for prediction. Error reduction is considered main factor for the accuracy of regression method (Dias *et al*., 2016). Regression methods are in this study.

Linear regression is considered the simplest kind of regression. It is used to characterize linear relations between the observed variables. It is used a measurement (x) to predict the value of (y) based on a linear function in Eq. 1:

$$y = \beta_0 + \beta_{1x} \quad (1)$$

The coefficients β_0 and β_1 is calculated using the least squares method (other methods shown by Diez *et al*. (2015). Linear regressions are assumed normally distributed data which may rarely occur when considering several sources of data within-study correlation estimates (Dias *et al*., 2016).

Kernel regression is a non-parametric model. It is used to estimate the probability density function of the observed data, starting with no assumptions about the data distribution. The goal of the kernel regression founding the value of E[Y|X] = m(X) for an unknown function m(.). To achieve that a regression is implemented based on the given values of X with the help of a Kernel function that is quantified the similarity between their data points. Finally, a new probability density function is drawn based on the observed values and it is used to predict the value of E[Y|X]. The main advantage of this method no assumption is required about any distribution of data and tend to has smaller errors when it is compared with the linear models.

The main problem in this method, it is need more data to find a proper approximation to the real distribution (Dias *et al*., 2016).

Gaussian Process regression (GP) is a collection of random variables. There is a finite set of such random variables with a joint multivariate Gaussian distribution, i.e., it is defined by the means and the covariance of the distributions. Each random variable (f(x)) is indexed by x and a covariance function that incorporates prior assumptions about the relation with the other distributions. Thus, no prior knowledge about the data is required for making accurate prediction. The good point it has higher accuracy when it is compared to other regression methods. The main drawback of the GP regression a computation time is required to make a prediction (Dias *et al*., 2016).

Also, regression neural network is one of regression methods. In this study, ARNN is suggested for developing RNN by adding feedback for stream time to reduce error and increase prediction accuracy.

Literature review: A two identical prediction models, one at the coordinator and the other at remote nodes are introduced by Tian and Zou (2006). The communication consumption of distributed stream is reduced by using the output from the prediction model to answer on the applied queries which are used by the coordinator. The remote nodes check the deviation between the predicted and the actual value. If significant deviation is founded, the updated message will be sent to the coordinator.

A prediction algorithm based on wavelet transform and Least Squares Support Vector Machine (LS-SVM) is introduced by Kong *et al*. (2008) to predict the time series data stream.

A linear regression model is proposed by Meng and Zhuang (2009). The proposed model is based on tendency correction for stream prediction. The results showed that the exponential smoothing method to adjust the prediction function parameters achieved better results than the weighted moving average method.

Alzghoul *et al*. (2012) employ a data stream predictor to improve the functionality of the fault detection system. It is used to improve the possibility for detecting failures in advance. Thus, increase the availability of industrial systems.

Two approaches are introduced by Lian and Chen (2006) for prediction. The polynomial and probabilistic are used to predict unknown values that have not arrived at the system and to facilitate the prediction and similarity search on future time series.

An adaptive prediction method is proposed by Sing, Mark and Leungb, the proposed system combines the strengths of NNs and multivariate regression models. This hybrid approach consists of two steps. In the first step a time series model generates estimates of the exchange rates. In the second stage, general regression neural network is used to correct the errors of the estimates (Chen and Leung, 2004; Chen *et al*., 2016).

General regression neural network is employed by Agarkar *et al*. (2016) to deal with the remote terminal units problem in power systems. The model handles noisy data for practical applications and has good performance in removing the unintended changes to the original data.

Many measures are listed by Neil. Timm to attest the quality of a prediction model in a certain use case. Considering the multiple options to make predictions common way to choose a model among a list of options. This is done in several ways by using measures such as Mean Square Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), the Relative MAE (RelMAE), the Mean Absolute Percentage Error (MAPE), the symmetric Mean Absolute Percentage Error (sMAPE) (Johnson and Wichern, 2007).

MATERIALS AND METHODS

Theoretical background: An Artificial Neural Network (ANN) is an information processing model that simulates by the way human brain system. It consists of a big number of greatly interrelated neurons that research in harmony to resolve particular problems. ANNs are learned by example. They are not programmed to accomplish a explicit task. A multilayer ANN is contains three layers. Input layer is linked to a level of hidden units which is joined to a layer of output units. The input units are used to characterize the raw data. The action of hidden unit is determined by the activities of the input units and the weights on the links between them. The output units behavior depends on the activity of the hidden units and the weights between the hidden and output units (Villa *et al*., 2016).

Regression is considered as a form of predictive techniques which explores the relation-ship between dependent (target) and independent variable's (predictor). This technique is used for estimating, time series modeling and finding the causal effect relationship between the variables. For example, relationship between sudden driving and number of road accidents by a driver is best predicted through regression. Linear and Logistic regressions are established as traditional algorithms in predictive modeling due to their popularity. Other types are founded such as polynomial regression, stepwise regression, ridge regression and others, more detailed is explained by Draper and Smith (1998).

A Regression Neural Network (RNN) is one of the applications in machine learning which it is used to predict the values of a numeric variables based on the values of one or more numeric or categorical predictor variables. For example, it is required to predict a yearly salary of a person based on age, gender and year of teaching. Regression types are classified into polynomial regression model, general linear model regression and Regression Neural Network (RNN). The last type of regression is considered the most influential form of regression in prediction. The RNN consists of a single output node that holds the predicted value of the dependent numeric variable, output unit is determined by its input values and a set of constants are called the weights and biases. The process of updating the weights values is called training the model. The basic idea is to try different values of the weights to decide where the computed output values of the NN closely match the known correct output values of the training data (Yi *et al*., 2016).

The proposed system: ARNN has been proposed to reduce the error and increase the accuracy of prediction model. Figure 1 shows the architecture of the proposed system.

For input stream, real dataset is named (PAMAP 2) for physical activity. The dataset contains data of nine healthful persons. Three Inertial Measurement Units (IMUs) are worn by each person. The IMUs are placed; one over the wrist on the dominant arm, one on the chest and one on the dominant side's ankle. Also, a heart rate has been measured. The data stream contains 2872532 measurements. Each person has 54 attributes values. The 54 attributes in the data files are formed in the following Table 1.

The performed activities are cycling, standing, vacuum cleaning, ironing, walking, running, Nordic walking, lying, sitting ascending stairs, descending stairs and rope jumping.

Interpolation using polynomial regression method has been applied for finding missing values. All details has been explained in our pervious study (Al-A'araji *et al*., 2016). The interpolation stage has been divided into three sub-stages (segmentation, thresholding, polynomial regression model). The segmentation algorithm has been contained the following steps.

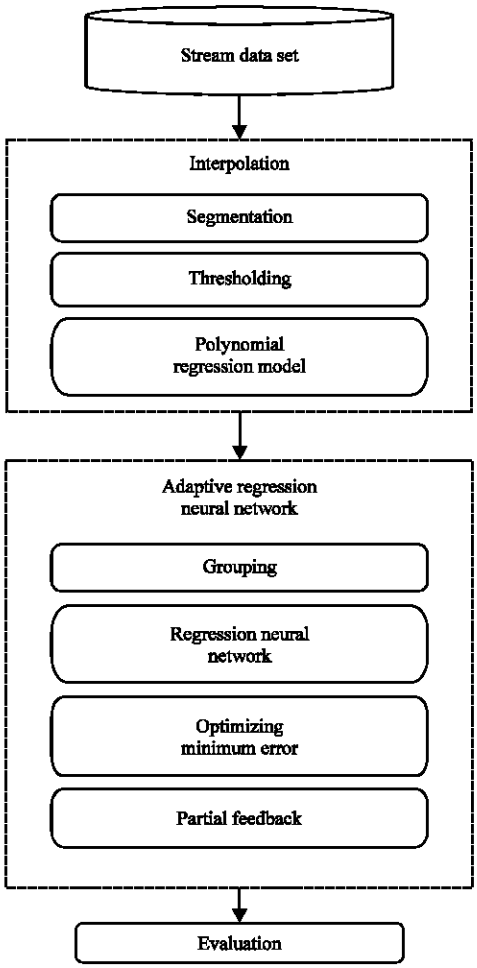

Fig. 1: Architecture of the proposed method

Table 1: The attributes in PAMAP 2 dataset

Column number	Attribute description
1	Timestamp
2	ActivityID
3	Heart rate (bpm)
4-20	IMU hand
21-37	IMU chest
38-54	IMU ankle

Segmentation algorithm:

Input: Matrix (D_{ij}) of stream with missing points where (I): time (j): Attribute

Output: B_{kn}; Matrix of all segment for known points and missing points data where k: number of known point (n = 1, 2, 3, …, 6): values that related to known point k

Step 1: extract two vectors from Dij data matrix

Vector xt: where t time value

Vector ya: where a is attribute value at time t

Step 2: Form vector R_t depend on ya if ya = NaN then R_t = 1 otherwise R_t = 0

Step 3: Locate the start and end for each missing value by finding known segment (B_{kn}) where k represent number of known point and (n = 1, 2, 3, …, 6) represent values that related to known point k

Do

If current point is not NaN (R_t = 1)then store the following in matrix B_{kn} otherwise increment the counter for NaN points N

$B_{k1} = x_t$ (time value)

$B_{k2} = y_a$ (attribute value)

B_{k3} = t (location of not NaN value)

B_{k4} = t-N (start of NaN)

B_{k5} = t-1 (end of NaN)

B_{k6} = k (counter of known point)

k = k+1

N = 0

while k< = Length (Rt)

Step 4: Return matrix of all segments B_{kn}

The polynomial curve fitting based on Segmentation of Variable Points and Variable Models (SVPVM) algorithm that has been proposed in our pervious study (Al-A'araji *et al*., 2016). This proposed model succeeded with high accuracy. Segmentation is the first sub-stage in this model. The polynomial regression model has been applied not as single model for all data points but the data has been segmented into overlapped segments based on threshold, then the model is applied on that known points segment to interpolate the missing points. The new segment has been taken and new polynomial parameters have been calculated to find missing values in current segment, the degree of polynomial has been changed based on the threshold value for each segment. The threshold has been selected to specify the polynomial degree. This process has been continued until no missing points.

The thresholding is the second sub-stage, the degree of polynomial has been specified based on number of known points that have been determined by some threshold.

The polynomial regression model is last stage in the preprocessing for finding missing points and preforming interpolation. Polynomial regression has been used with order of Threshold-1 for each segment of data. Polynomial models are applied for each segment concurrently. The SVPVM Model is shown below:

Polynomial regression model:

Input: B_{kn}: Matrix of all segment for known points and missing points data without missing points where k represent number of known point and (n = 1, 2, 3, …, 6) represent values that related to known point k

Output: $\dot{D}_{ij}$; Matrix of data without missing points

C_{pq}: Matrix store coefficient of polynomial model where

p: Counter determine the model number

q: Polynomial coefficients p = 1, ..., (T-1) where T is threshold

Step 1: From B_{kn} matrix select number of points according to threshold (T) with shift and overlapping

Do

Form p1 from (x1, x2, x3, ..., xT)

Form p2 from (y1, y2, y3, ..., yT)

Apply equation of polynomial regression model of order (T-1) using Eq. 2:

$$y = 0+\sum_{n=1}^{(T-1)} n^{x^n} \quad (2)$$

Where (0, m) are coefficient of polynomial and stored in matrix Cpq, x represent vector p1,y represent vector p2

Find $\hat{y}$ current segment by substitution the time of missed value and replace NaN by the result of applied model

While k<= Number of points in B_{kn}
Step 2: Return D_{ij}^{y} with new attribute

ARNN is the core of this study. Also, it consists of sub-stages (grouping, regression neural network, optimizing minimum error, partial feedback) . Grouping has been applied on the dataset, the data has been grouped according to features relationship. Four groups have been used (heart rate and information of each sensor IMU 1-IMU 3).

ARNN has been proposed to has many nodes for input, the number of nodes has been determined by the number of attributes for selected dataset (52 node). Single node for output represent the estimated output. Weights of the RNN have been initialized by the coefficients of multiple regression that have been calculated for each data attribute as in Eq. 1. The error between actual output and the estimated output is computed by Eq. 2. A Contribution Update Factor (CUF) represents the adaptation process of weights update according to the time and contribution of error in addition to the feedback operation. The ARNN has been trained on the selected training set of data. The dataset have been collected in 10 h. This time period is divided into several intervals each with a length of half an hour and then every half hour has been partitioned into training set and testing set (70% for training and 30% for testing). This is best partition to get minimum error. The updating process continues on the training data and adjusting the prediction result until minimum error has reached. The general structure of ARNN illustrates in Fig. 2.

Where, A_1, A_2, ..., A_n represent the attributes of the data set, β represents weight vector which is calculated by the following Eq. 3:

$$\beta = (A^T A)^{-1} A^T \quad (3)$$

The E value represent error value between actual output (P) and Predicted output ($\dot{P}$) as in Eq. 4:

$$E = \dot{P} - P \quad (4)$$

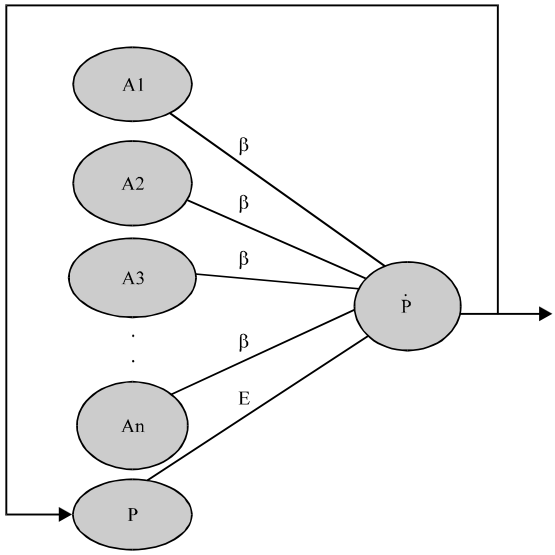

Fig. 2: Structure of ARNN

where, $\dot{P}$ and P represent estimated prediction and desired prediction, respectively. The feedback in ARNN is used to reduce the error in each epoch of training. The value of P is computed by the following Eq. 5:

$$\dot{P} = A\beta + E \quad (5)$$

In partial feedback T has been used to specify the currently period time.

RESULTS AND DISCUSSION

The SVPVM is applied for interpolation. Figure 3 shows example for heart rate of one file which is taken from (PAMAP 2) dataset. Figure 3a reviews the original file and Fig. 3b illustrates the file after interpolating missing values.

The ARNN has been applied on the dataset, the data is grouped according to features relationships as shown.

Grouping stage (give result for one file): The time for dataset is 10 h, split it to many half hour by Eq. 6:

$$\text{Interval} = \frac{\text{Number of records}}{20} \quad (6)$$

The size of data in file is (213598×54), each half of size (10680). Split each interval (half hour) into 70% training set and 30% test set. Segment the attribute of data set into four related groups:

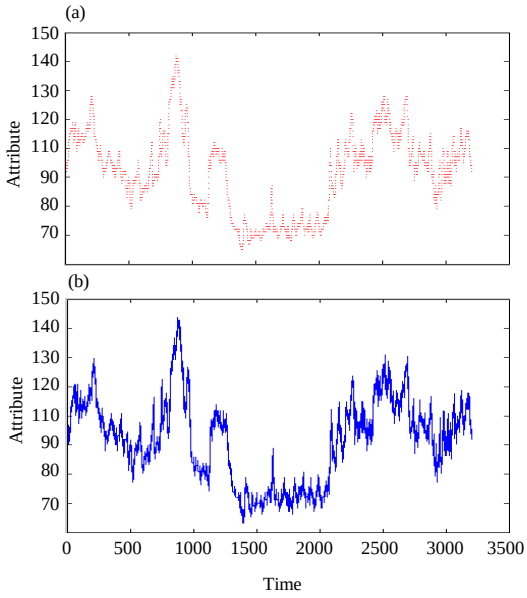

Fig. 3: Fitting missing points: a) Original data with missing value and b) Interpolated data

Table 2: Evaluation measure of ARNN

	Dataset file 1 (1st epoch)		Dataset file 2 (2nd epoch)	
Epoch	MSE	MAE	MSE	MAE
G1	25.509	4.467	40.374	5.012
Heart rate	7.117	1.984	4.101	1.557
G2	4.625	1.667	0.000	0.000
IMU1	2.565	1.260	0.000	0.000
G3	7.227	2.122	0.000	0.000
IMU2	1.209	0.905	0.000	0.000
G3	7.226	1.920	0.000	0.000
IMU2	3.436	1.562	0.000	0.000

Table 3: Results of traditional regression methods

Dataset file/Evaluation measures	MR	LR
File 1		
MSE	1.6335	1.7035e+05
MAE	1.0216	323.3523
File 2		
MSE	37.3264	7.1210e+05
MAE	4.8719	0

- G1: Attribute (3) for hart rate
- G2: Attribute (4-20) for IMU hand
- G3: Attribute (21-37) for IMU chest
- G4: Attribute (38-54) for IMU ankle

Furthermore, Table 2 shows the (MSE) Mean Square Error and (MAE) absolut error for prediction stage for samples of only two files from dataset.

The proposed ARNN has been compared with two traditional methods for regression (Multiple Regression MR, Linear Regression LR) on same files of dataset and the results is stated by Table 3.

The results refer that the proposed ARNN gave less error in a comparison to the traditional regression methods.

CONCLUSION

Stream prediction has been considered an important field in many applications. A developed polynomial regression models using segmentation, thresholding and multi models regression have proved as an excellent method in the interpolation for processing missing and outlier values. Also, the ARNN using partial feedback is suggested as a new approach for minimizing error in the RNN. Consequently, the ARNN has satisfied promise results in the stream prediction with high accuracy.

REFERENCES

Agarkar, P., P. Hajare and N. Bawane, 2016. Optimization of generalized regression neural networks using PSO and GA for non-performer particles. Proceedinsg of the 2016 IEEE International Conference on Recent Trends in Electronics Information and Communication Technology (RTEICT'16), May 20-21, 2016, IEEE, Bangalore, India, ISBN: 978-1-5090-0775-2, pp: 103-107.

Al-A'araji, N.H., E. Al-Shamery and A.H. Alyaa, 2016. A new polynomial curve fitting based on segmentation of variable point and variable modes for reconstructing missing values. Res. J. Appl. Sci., 11: 1089-1094.

Alzghoul, A., M. Lofstrand and B. Backe, 2012. Data stream forecasting for system fault prediction. Comput. Ind. Eng., 62: 972-978.

Chen, A.S. and M.T. Leung, 2004. Regression neural network for error correction in foreign exchange forecasting and trading. Comput. Oper. Res., 31: 1049-1068.

Chen, K., Y.S. Koh and P. Riddle, 2016. Proactive drift detection: Predicting concept drifts in data streams using probabilistic networks. Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN'16), July 24-29, 2016, IEEE, Vancouver, British Columbia, Canada, ISBN:978-1-5090-0621-2, pp: 780-787.

Dias, G.M., B. Bellalta and S. Oechsner, 2016. A survey about prediction-based data reduction in wireless sensor networks. ACM. Comput. Surv., Vol. 49, 10.1145/2996356.

Diez, D., C. Barr and M. Cetinkaya-Rundel, 2015. OpenIntro Statistics. 3rd Edn., OpenIntro Inc., Rumford, Maine, ISBN: 9781943450053, Pages: 436.

Draper, N.R. and H. Smith, 1998. Applied Regression Analysis. Vol. 1, John Wiley & Sons, Hoboken, New Jersey, ISBN:9780471170822, Pages: 706.

Johnson, R.A. and D.W. Wichern, 2007. Applied Multivariate Statistical Analysis. 6th Edn., Prentice Hall, Upper Saddle River, New Jersey, ISBN: 9780131877153, Pages: 773.

Kong, Y., Y. Shi and J. Yuan, 2008. Prediction method of time series data stream based on wavelet transform and least squares support vector machine. Proceedings of the 4th International Conference on Natural Computation (ICNC'08) Vol. 2, October 18-20, 2008, IEEE, Jinan, China, ISBN: 978-0-7695-3304-9, pp: 120-124.

Lian, X. and L. Chen, 2006. Efficient methods on predictions for similarity search over stream time series. Proceedings of the 18th International Conference on Scientific and Statistical Database Management, July 3-5, 2006, IEEE, Vienna, Austria, ISBN:0-7695-2590-3, pp: 241-250.

Meng, F. and P. Zhuang, 2009. Stream prediction model based on tendency correction. Proceedings of the 6th Conference on Web Information Systems and Applications (WISA'09), September 18-20, 2009, IEEE, Xuzhou, China, ISBN:978-0-7695-3874-7, pp: 189-193.

Tian, L. and P. Zou, 2006. Prediction models over distributed data streams. Proceedings of the International Conference On Web Information Systems Engineering (WISE'06), October 23-26, 2006, Springer, Wuhan, China, pp: 25-36.

Villa, A.E.P., P. Masulli and A.J.P. Rivero, 2016. Artificial Neural Networks and Machine Learning-ICANN 2016: 25th International Conference on Artificial Neural Networks, Barcelona, Spain, September 6-9, 2016, Proceedings. Springer, Berlin, Germany, ISBN:978-3-319-44780-3, Pages: 557.

Yi, W., I.V. Gerasimov, S.A. Kuzmin and H. He, 2016. An intelligent algorithm of Support Vector Regression parameters Optimization in soft measurements. Proceedings of the 19th IEEE International Conference on Soft Computing and Measurements (SCM'16), May 25-27, 2016, IEEE, St. Petersburg, Russia, ISBN:978-1-4673-8920-4, pp: 404-406.