ISSN No. 0976-5697

Volume 3, No. 1, Jan-Feb 2012

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

An Efficient Model for Time Series Prediction using Instance Based Learning Technique

Pooja M R
Associate Professor
Department of Computer Science and Engineering
Vidyavardhaka College of Engineering, Mysore,
Karnataka, India
pooja_m_r@yahoo.co.in, poojamr.cs@vvce.ac.in

Abstract: Multi-step ahead time series forecasting has become an important activity in various fields of science and technology, due to its usefulness in future events management. Long-term or multi-step prediction problem is a challenging area as it predicts several steps ahead into the future - starting from information at current instant. Storing and using specific instances improves the performance of several supervised learning algorithms. Instance-Based learning is a framework and methodology that can be applied to generate time series predictions using specific instances. In this paper we propose a new concept to improve the performance of prediction model. The proposed learning technique implemented here extends the nearest neighbour algorithm to include the concept of pattern matching to identify similar instances thus implementing a nonparametric regression approach. Pattern matching in the context of time-series forecasting implies the process of matching current state of the time series with its past states. Specific instances are chosen using hybrid distance measure which includes correlation measure with that of Euclidean distance measure. The instances chosen are combined using multiple regressions to generate multi-step ahead predictions. Bench mark data set of Mackey-Glass series and real time series of Nord Pool are used to test and validate the proposed technique. Experimental results show remarkable enhancement in the performance of our prediction model.

Keywords: Instance Based Learning (IBL), Multistep ahead Prediction, Time series forecasting.

I. INTRODUCTION

Time series forecasting is the use of a prediction model to forecast future events based on known past events: to forecast future data points before they are measured. A standard example in econometrics is the opening price of a share of stock based on its past performance. The term time series analysis is used to distinguish a problem from more ordinary data analysis problems where there is no natural ordering of the context of individual observations and from spatial data analysis where there is a context that observations relate to geographical locations [1]. There are possibilities where space-time models are formed analyzing spatial-temporal characteristics. A time series model will generally reflect the fact that observations close together in time will be more closely related than observations further apart. In addition, time series models will often make use of the natural one-way ordering of time so that values in a series for a given time will be expressed as deriving in some form, from past values rather than from future values [1, 10].

Forecasting is a central component in many organizations. For example, one of the most critical aspects of inventory and supply chain management is the ability to accurately predict demand [5]. Inventory levels, reorder points, probability of stock outs, and production levels are all based on forecasted demand. The ability to forecast the behavior of a system hinges, generally, on the knowledge of the laws underlying a given phenomenon. When this knowledge is expressed as a solvable equation, one can predict the behavior along the future with the initial condition is given. However, phenomenological models are often unknown or extremely time consuming. As it is possible to predict the dynamic behavior of the system along the future by extracting knowledge from the past, the time series behavior can be captured by expressing the value x (k+1) as a function of the d previous values of the time series, x (k).., x (k-d) and expressed as: X (k+1) =F(x (k), x (k-d)) (1)

In many time series applications, one-step prediction schemes are used to predict the next sample of data, x (k+1), based on previous samples. The disadvantage of one-step prediction is that it may not provide enough information especially in situations where a broader knowledge of the time series behavior is desirable to anticipate the behavior of the time series process. Hence the long-term or multi-step prediction model is used as it obtains predictions several steps ahead into the future i.e., x (k+1), x (k+2), x (k+3), starting from information at current instant k [5, 6, 8].

The proposed system aims at developing a model for Multi-step ahead time series forecasting using an improved instance based learning approach. The model implements a learning technique that extends the nearest neighbor algorithm to include the concept of pattern matching to identify similar instances thus implementing a nonparametric regression approach.

The paper is organized as follows. In Section II we focus on Instance based learning concepts. The proposed methodology and the algorithm for multi-step prediction are presented in Section III. Results of experimentations for Mackey-Glass series and Nord Pool time series for real time are discussed and shown in section IV. Finally, in Section V we have the conclusions and future works.

II. INSTANCE BASED LEARNING

Storing specific instances and using the instances improves the performance of several supervised learning algorithms, as in algorithms that learn decision trees, classification rules, and distributed networks. Instance-based learning is a methodology that generates classifications or predictions using only specific instances. Instance-based learning algorithms do not maintain a set of abstractions derived from specific instances; however, no investigation has analyzed algorithms that use only specific instances to solve incremental learning tasks. The proposed IIBL technique approach extends the nearest neighbor algorithm to include the concept of pattern matching to identify similar instances. IBL is an illustration of how storage requirements can be significantly reduced with little compromise in learning rate and classification accuracy.

Using specific instances in supervised learning algorithm decrease the cost incurred while updating concept descriptions, increases learning rate, allows the representation of probabilistic concept descriptions, and focuses theory- based reasoning in real-world applications. However, no investigation has analyzed algorithms supervised learning tasks [2].

IBL algorithms haves certain limitations that they don t' construct explicit abstractions such as decision trees or rules [2, 6]. Most learning algorithms derive generalizations from instances when they are presented and use simple matching procedures to classify subsequently presented instances. This incorporates the purpose of the generalizations at presentation time.IBL algorithms perform comparatively less computations at presentation time as they do not store explicit generalizations. However, the computation is higher when presented with subsequent instances for classification, as they have to compute the similarities of their saved instances with the newly presented instance. This obviates the need for IBL algorithms to store rigid generalizations in concept descriptions, which can require large updating costs to account for prediction errors. Hence, there is substantial scope to improve the performance in terms of prediction accuracy. In our work presented, we look at the problem of analyzing and synthesizing the model from a simple perspective. Instead of making the algorithm too complex for claiming better accuracy such as in [5], we propose to introduce a novel idea to build an optimal prediction model as explained in the next section.

III. PROPOSED METHODOLOGY

The proposed system aims at developing a model for Multi-step ahead time series forecasting using an improved instance based learning approach. The model implements a learning technique that extends the nearest neighbor algorithm to include the concept of pattern matching to identify similar instances thus implementing a nonparametric regression approach. [2, 3] Our approach uses a hybrid distance measure that combines both correlation and Euclidean distance to select the similar instances.

Description of the proposed system:

The instance based learning approach is the concept of identifying a set of data that are most similar to the past information available before the data that is to be predicted or forecasted. The primary assumption in the time series data is that the series copies its own behavior and past pieces of information on the series that have symmetry with the available past information [1, 2]. The instance based classification aims at locating similar pieces of information independent of their location in time. While locating the most similar neighbors the method tries to eliminate outliers present in the time series data.

Given the time series, initially the time series data set is partitioned into training set and test set. If the total number of observation in the training set is S, and the training set is divided into N instances each of a specific size, say L. Where N is computed dividing S by L, H is the number of nearest neighbors that will be chosen as the best instances out of N instances for the process of forecasting. The values of H and L are determined by conducting several trials so as to yield better results in terms of RMSE and MAPE. H thus indicates the size of the nearest neighborhood and L indicates the window size. The instance that lies just before the time value to be predicted is chosen as the critical instance and considered as reference pattern against which the similarity of other instances called candidate patterns is to be estimated. Euclidean correlation is used as the similarity metric to choose similar instances. We first select those patterns that exhibit positive correlation with the reference pattern and then apply Euclidean similarity metric to estimate the similarity between reference pattern and candidate patterns. H most similar instances are chosen by selecting patterns that have least Euclidean distance. Once the similar instances are chosen they are combined using multiple regression method to predict future forecast pattern. The forecast instance is then added to the training set, and is now treated as the new reference pattern, against which the similarity of other candidate patterns is estimated. This process is repeated till the desired number of forecast values is generated. Algorithm for training the model along with the notations used is given in the following subsections.

A. Notations:

TR size of the training data set
L size of the individual patterns
H number of windows chosen the best instances
Forecast_Corr -Forecast vector containing the
Forecast values for the testing period
Error_CorrError vector containing the difference between the actual test data set values and forecast values

B. Algorithm for the proposed learning technique:

Step 1: Set train_ set = 50 percent of total data set size and test_ set = 0.5 of total data set size
Step 2: Set L and H to suitable values. Step 3: Set Ref_pattern = $X_{cur.}$
Step 4: Set Candidate_ patterns = X_i
Where i =1, 2, 3…n-1
Step 5: for j=1: L

If (Corr[t]>0)
(Corr [t] = correlation index obtained by estimating correlation between reference pattern and candidate pattern)
Euclid_Corr [t] = $(X_{j,cur} - X_{j,i})^2$ ...
For t=1, 2, 3 . (S/L) where S is the total number of observations in the training set.
Step 6: Choose H lowest values from Corr[i] For j=1…. (S/L)
Low[i] = min (Euclid_Corr Corr[j])); Where i = 1: H
Step 7: $Xcur_{+1} = \alpha_0 + \alpha_1 X_{t\alpha 1} + \alpha_2 X_{t\alpha 2} + \ldots + \alpha_L X_{t-L}$
Step 8: Add X_{cur+1} to train_period
And $X_{cur.}$ To X_{cur+1}
Step 9: Set Candidate patterns = X_i, i=1, 2, 3...n
Step 10: Repeat steps 5 to 9 for all values in test_period so that the forecast values are generated.

IV. EXPERIMENTATIONS

A. *Data sets:*

The performance of the model has been tested on Mackey-Glass Chaotic benchmark time series and a real time series data set employed for Day-ahead electricity price forecasting in Nord Pool [4]. The details of the datasets are provided.

B. *Parameter Selection (H & L):*

The proposed improved instance based learning (IIBL) technique extends the k-nearest neighbor to include the pattern matching algorithm. The number of instances which contain patterns and the size of individual patterns present in each instance must be predefined. To check the sensitivity of the performance of model on the number of nearest neighbors, various numbers of patterns and instances of different sizes were investigated to obtain optimal value. The number of patterns in each instance was varied from 2 to 5, assuming initially fixing the number of nearest instances to be 10. Investigations show that the RMSE values initially a start decreasing as the number of patterns increases in each instance and later RMSE tends to increase with increase in pattern size as indicated in Table I. A pattern size of 3 was opted as it provided the best performance (the smallest RMSE value). With a pattern size of 3, the number of nearest instances is varied to find the effect of number of neighbors on the performance of the system. As can be observed from Table II, there is a significant improvement in the performance of the system when the number of instances (H) chosen is 30, 40 or 50. However, this would result in decrease in the speed at which the learning or training takes place as more number of observations will be involved. Hence, H value of 20 is chosen as the number of nearest instances that will be used for forecasting process accounting for negligible compromise in performance accuracy.

Table 1: RMSE values obtained for different pattern sizes with nearest neighbors (instances) to be 10 for Mackey-Glass Chaotic time series.

Pattern size(p)	RMSE
2	0.0093
3	0.0077
4	0.0082
5	0.0092
6	0.0093
7	0.0186
8	0.0127

Table 2: RMSE values obtained by varying the number of nearest neighbors (instances) by fixing pattern size of 3 for Mackey-Glass Chaotic time series.

Number of nearest neighbors(K)	RMSE
10	0.0077
20	0.0070
30	0.0068
40	0.0067
50	0.0067

a. ***Mackey-Glass Chaotic System****:* The time series used in this simulation is generated by the chaotic Mackey-Glass differential delay equation:

$$dx(t)/dt = -0.1x(t) + 0.2x(t-)/[1+x(t-)]^{10}$$

Where $x(0) = 1.2$, $= 17$, and $x(t) = 0$ for $t < 0$.

In our experimentations 1000 data points have been chosen, and the simulations are done considering training set comprising of 75%, 50% and 25% of the total data points.

b. **Nord pool electricity prices**: This dataset consists of 988 observations and was downloaded from the Nord Pool server. The observations correspond to the daily electricity prices recorded at the end of the day [4].

Table V shows the performance of our proposed IIBL with that of linear and neural network model for Nord pool time series [4]. Our proposed model for Nord pool series depicts with less prediction error as shown in Table VI. The total data set used here is 988 samples values which are equally partitioned.

V. RESULTS AND DISCUSSIONS

To observe the impact on the amount of dataset used and also to analyze the learning characteristics of the proposed model, experiments were conducted considering data set with 25 and 75 percentage of patterns separately for training and testing processes. A data set of 250 samples and 750 samples were used for training and its impact on the observations of RMSE and MAPE have been shown in Table III. It is observed that increase in the training data set size would result to decrease in error.

Table IV shows the comparison of proposed IIBL model with the Anfis model for Mackey Glass series. Our proposed model outperforms better compared to Anfis Experimentations with the series partitioned considering

equal number of data points in the training set and test set resulted in reduction in RMSE values are as shown by the results tabulated in Table VI. Based on the observations, it is also evident that the proposed IIBL model achieves reasonably good prediction when trained with different sizes of training and test sets and prediction accuracy level is close to 95 %.

Table 3 : RMSE values with different training size for Mackey- Glass series

	% of Training	Training size	Test size	RMSE	MAPE
Mackey-Glass Series	75	750	250	0.0019	0.0017
	25	250	750	0.0021	0.0018

Table 4: Comparison of IIBL model with that of Anfis system for Mackey-Glass series

Type of mode	Training size	Test size	RMSE
Anfis model	500	500	0.0332
IIBL model	500	500	0.0020

Table 5: RMSE values of IIBL model with that of Linear and Neural Network estimator for Nord Pool Series

Type of model	Training size	Test size	RMSE
Naïve Linear model	1080	24	6.22
Neural network estimator	1080	24 5	5.20
Proposed IIBL	494	49	1.8313

Table 6: RMSE & MAPE values with equal partition of data

Data sets	Training size	Test size	RMSE	MAPE
Mackey-glass	500	500	0.0020	0.0017
Nord pool	494	494	1.8313	0.0127

Fig.1 shows the plot of the complete original Mackey Glass series and the Fig.2 shows the forecast plot using the proposed IIBL technique trained with data set size of 500, considering first 500 points for training the model and predicting the next 500 points. Fig.3 shows the complete Nord Pool series and Fig.4 the actual plot of the predicted points and Fig.5 shows the plot obtained for Nord Pool series using the proposed IIBL technique considering the training set size of 494. The results shows from the predicted plots that for both the experiments the predicted plots exactly matches the test data points, resulting in achieving prediction accuracy of above 95%.

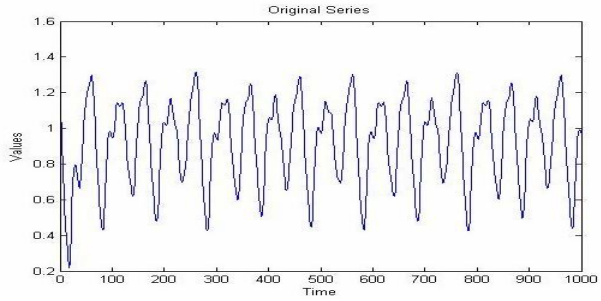


Figure. 1 Mackey Glass series

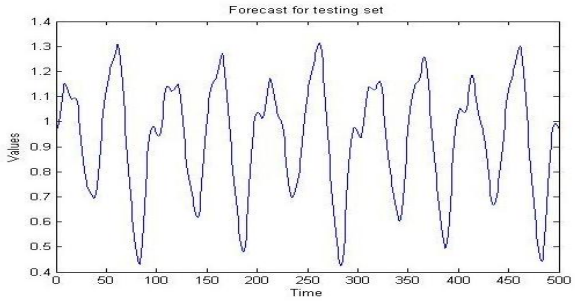


Figure. 2 Forecast Mackey Glass series

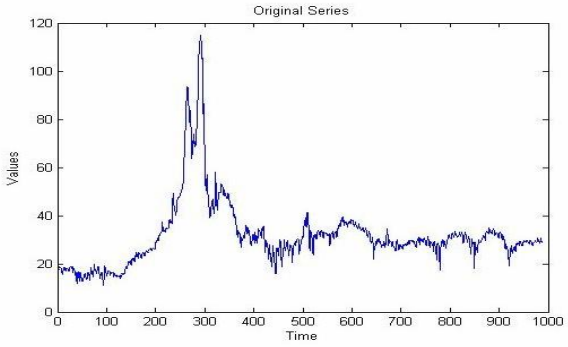


Figure. 3 Nord Pool series

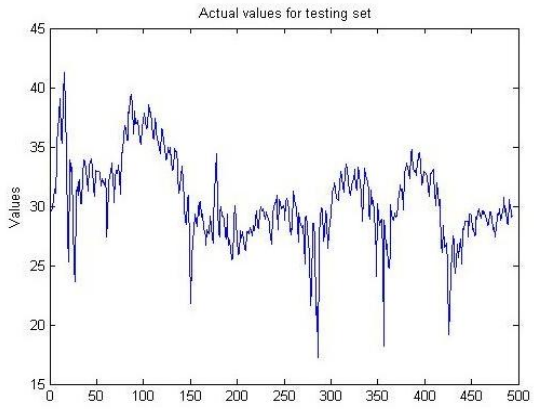
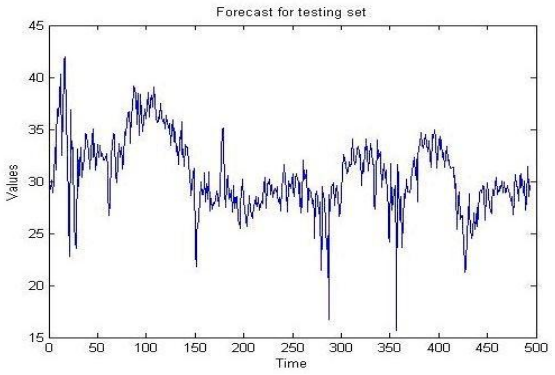


Figure: 4. Actual Nord Pool Series



Figure, 5 Forecast Nord Pool series

VI. CONCLUSION

In this paper, we have discussed the model for multi-step ahead time series forecasting using improved Instance based learning to achieve better performance in terms of prediction accuracy. The proposed learning technique extends the naïve nearest neighbor technique to incorporate a pattern matching methodology which retrieves the significant patterns from the historic data using instance based learning approach. The patterns are then used in the forecasting process by applying the method of multiple regressions. The model parameters are optimized by several experimentations. The hybrid distance measure adopted yields better results as against simple Euclidean distance used in naïve instance based techniques. The performance of the model is evaluated on Mackey-Glass benchmark time series and a real time series of Nord pool data used in day ahead forecast of electric prices. Further, the results obtained from the proposed model shows that it outperforms the anfis fuzzy prediction system in the case of Mackey-Glass Chaotic time series and the linear model and the neural network estimator non-linear model in the case of Nord Pool data**.** The storage-reducing algorithm performs well on several real world databases, but their performance degrades rapidly with the level of noise in training instances. Our future work is to investigate techniques to distinguish noisy instances.

REFERENCES

[1]. Sameer Singh and Jonathan Fieldsend Financial time series forecasts using fuzzy and long memory pattern recognition Systems- ,Computational Intelligence for Financial Engineering, 2000. (CIFEr) Proceedings of the IEEE/IAFE/INFORMS 2000 Conference Volume, pp.166 169, Issue 2000

[2]. David W Aha, Dennis Kibler, Marc K Albert, Instance- Based Learning Algorithms- Machine Learning (1991) ©1991 Kluwer Academic Publishers, Boston. 1991, pp37-66

[3]. Syed Rahat Abbas, Muhammad Arif Modified nearest neighbor method for Multi-step ahead time series prediction, International Journal of Pattern Recognition and Artificial Intelligence, Year: 2007 Vol: 21 Issue: 3 pp: 463 - 481 .

[4]. Achilleas Zapranis, Stratos Livanis, Forecasting the day ahead electricity pricing in Nord Pool with Neural networks: Some preliminary results, h t t p ://www.valueinvest.gr www.valueinvest.gr, pp 50-66.

[5]. Sameer Singh, Fuzzy nearest neighbor method for time series forecasting, Proc. 6th European Congress on Intelligent Techniques and Soft Computing (EUFIT'98), Aachen, Germany, vol. 3, pp. 1901-1905,1998

[6]. David W Aha, Dennis Kibler, Marc K Albert, Instance- Based Learning Algorithm, Machine Learning, Kluwer Academic Publishers, and Boston. Vol- 6,pp 37-66,1991