Prediction of Univariate Time Series Based on Clustering

Sivo Daskalov, Ventsislav Nikolov

Abstract: *In this paper an approach is proposed for the prediction of the behavior of time series. In order to achieve that various subseries with fixed length are formed from the initial data series, which are then grouped into clusters based on their shape. The data kept in the clusters for each of the subseries are the relative differences between all of its consecutive values. This enables these differences subseries to be averaged to form a single series for each cluster. These cluster centers are used for the prediction of nonexistent future values.*

Keywords: *Prediction, clustering, self-organizing map, time series, artificial intelligence*

ПРЕДСКАЗВАНЕ НА ВРЕМЕВА СЕРИЯ БАЗИРАНО НА КЛЪСТЕРИЗАЦИЯ

Сиво Даскалов, Венцислав Николов

Резюме: *В този доклад е представен подход за предсказването на поведението на времеви серии. За да се постигне това, от началната серия се образуват множество подсерии с фиксирана дължина, които в последствие се групират в клъстъри в зависимост от формата им. Данните съхранявани в отделните клъстъри за всяка от подсериите са относителните разлики между всеки два съседни елемента. Това позволява осредняването на групата от подсерии в една за всеки клъстър. Тези клъстърови центрове се използват за предсказването на несъществуващи бъдещи стойности.*

Ключови думи: *Предсказване, клъстеризация, самоорганизиращи се карти, времева серия, изкуствен интелект.*

1. Introduction

The prediction is of great importance in many domains. Economics, finance, the public sphere, technologies and many physical processes are amongst the ones that benefit greatly from predictions. The predictions can often be inaccurate but there are a number of different approaches to overcome this problem. First of all indicators for predictability of the available historical data exist. If the data is not identified as predictable then mathematical methods could be applied but would hardly produce good results. Secondly, the predictions cannot be considered as sole indicator values but rather as most probable values with a given confidence levels. Thus the predicted values can be substituted with ranges of values in accordance with the level of confidence.

The prediction methods are divided into two main categories: univariate and multivariate. The univariate prediction methods take into consideration only the data available as historical information while multivariate methods use dependencies between a number of additional factors as well as the historical data. This paper presents an approach for univariate time series prediction. The time series used here are considered to be discrete, that is, their values have been obtained in equal time intervals.

There are many well-known approaches for univariate time series prediction: the classical Box and Jenkins methodology [1]; regression based methods [3] like autoregressive (AR), autoregressive moving average (ARMA), autoregressive integrated moving average (ARIMA), seasonal autoregressive moving average (SARIMA), self-exciting threshold autoregressive (SETAR) [8]; soft-computing methods (neural networks) [7, 9, 11, 13], decomposition based methods (e.g. Holt-Winter), Kalman and Wiener filtering, etc. Each one of them has its specific characteristics and performance, having both advantages and disadvantages, and some of them are based on the extrapolation principle.

The approach proposed here is based on building of a mathematical model which combines subseries grouping obtained from the historical series and performing of multistep iterative prediction. The overall time series prediction consists of two stages. In the first stage the historical time series is used to build the model and in the second stage the model itself is used for the production of an arbitrary number of future values. The method is applicable only for series for which there are dependencies in the time development of the series. If the time series is completely chaotic with nonrelated values the method will also work but will not produce satisfactory results.

1. Time series preprocessing

In order to avoid problems related to infinitely increasing trend and series nonstationarity, the time series has to be preprocessed. Linear approximation is used to find such a line that the root-mean-square error (1) between the line and the time series is as low as possible – fig. 1.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y)^2}{n}} \tag{1}$$

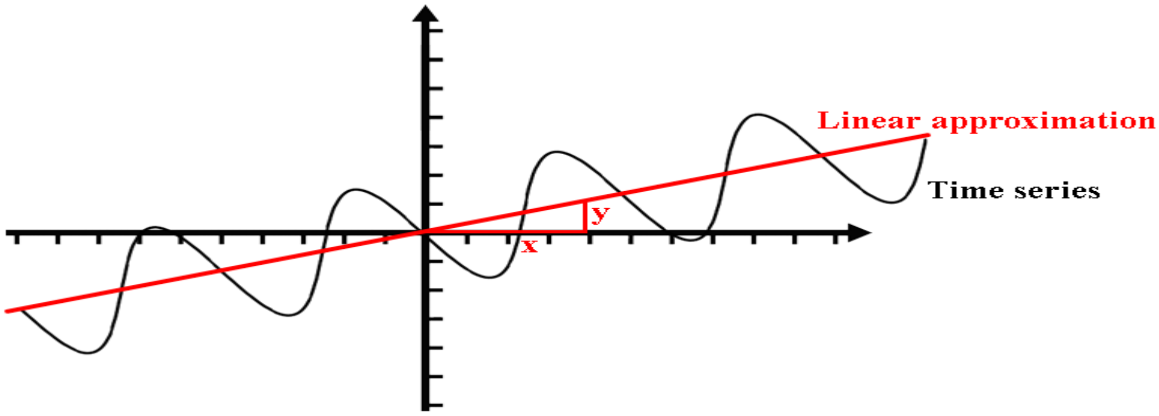

Fig. 1. Linear approximation of a time series

The slope of the approximated line is then calculated (2) and if this slope is within (-0.1, 0.1) the time series X is stored by its original values, otherwise the series Y is created from the original time series X as for every element the relative difference (3) to the previous element in X is calculated and stored.

$$m = \frac{y}{x} \tag{2}$$

$$y(t) = \frac{x(t) - x(t-1)}{x(t-1)} \tag{3}$$

If the time series is stored by its elements' relative differences, the initial values can be restored by a combination of multiplication and addition (4).

$$\mathrm{x(t) = y(t) * x(t-1) + x(t-1)} \quad (4)$$

This restoration of the original time series is possible because the first value in the time series has been stored and the differences between every other value and the previous one are known as well.

2. Model building stage

Most of the time series prediction methods use previous historical values in order to produce the next value (5):

$$y(t) = f(y(t-1), y(t-2), \dots, y(t-k)) \quad (5)$$

Thus, the mathematical model can be built analyzing all possible subseries of length k. Graphically this is shown in fig. 2 where the initial time series is separated into n-k+1 (n=23, k=5 in the example) subseries where n is the time series length and the k is model order. In one of the most commonly used autoregressive methods this subseries are considered as matrix rows and through these rows a set of parameters is calculated which is later used in the prediction stage to produce the future values. In our approach, the subseries are assigned into a given number of groups which can be considered as clusters. The assignments should be performed in such a way that the distance between the subseries within a group should be minimal while the distances between the subseries in different groups should be maximal. A distance between the subseries could be calculated in an arbitrary way, for example through the Minkowski distance algorithm. Various clustering algorithms could be used for the grouping of the subseries: k-means, ISODATA, hierarchical clustering, self-learning neural networks [4, 6, 10, 12].

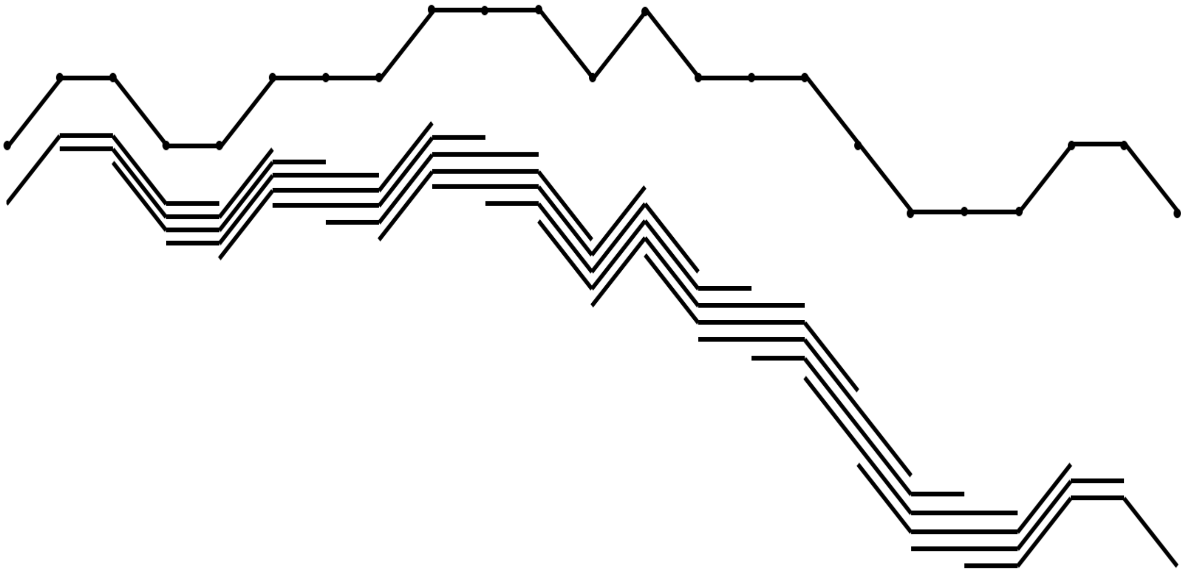

Fig. 2. Segment breakdown of a time series

For each segment, the relative change between each of its consecutive elements is calculated and stored. The resulting vectors are then grouped into clusters through one of the clustering algorithms. This achieves a clustering [2, 4] by segment shape as shown in fig. 3.

The subseries shown in fig. 2 are additionally separated in two parts. The first one is of length k-1 and the second one consists of only the last value. The clustering is performed considering only the first subseries part, but the second part remains attached to it. The clustering for the given time series in fig. 2 is shown in fig. 3.

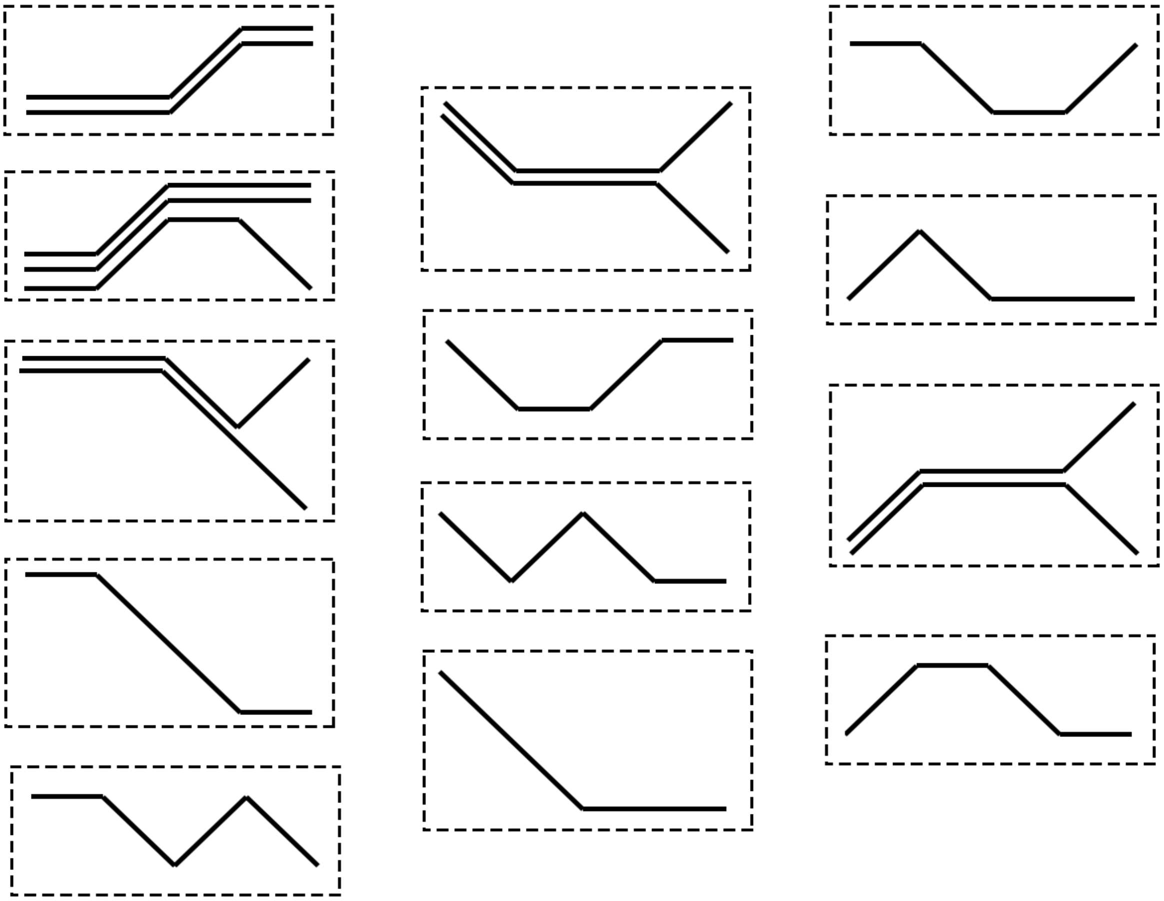

Fig. 3. Segment clusters formed from the initial time series

When the subseries groups are determined, each cluster's center for each of the groups is calculated. This is achieved through the calculation of the average value of all subseries' values at the specified index according to (6). The first k-1 values of the cluster's center are compared with an arbitrary segment while the last value is kept for the actual prediction of the next element. Various clusters and their calculated centers are shown in fig. 4.

$$C(i) = \frac{\sum_{j=1}^{p} Y(j,i)}{p}, where\ p\ is\ the\ number\ of\ subseries\ in\ the\ cluster \quad (6)$$

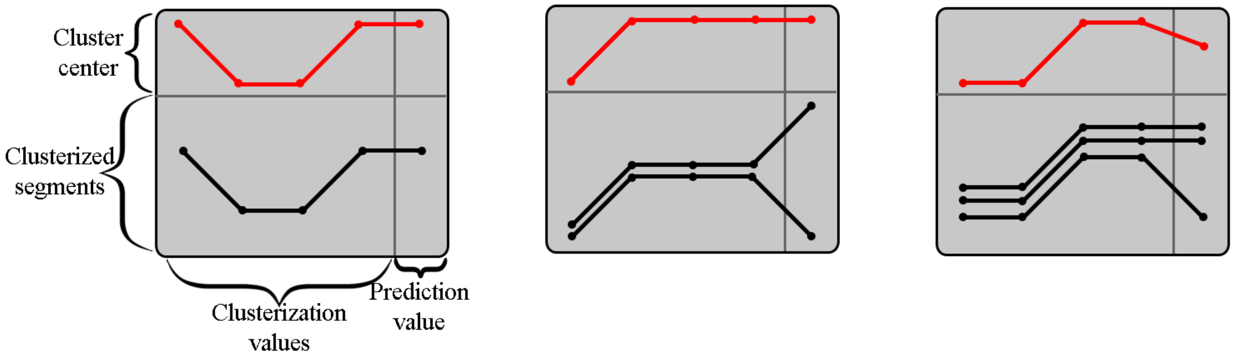

Fig. 4. Segment structure

3. Prediction stage

The prediction is an iterative multistep process. In every iteration a single value is generated by the model built in the previous stage. Fig. 5 illustrates the prediction process. The process consists of an arbitrary number of repetitions of the main prediction cycle. The cycle consists of the following operations:

a) A segment with length k-1 is extracted from the current end of the time series. This is not necessarily the original time series as predictions could have already been appended.

b) The extracted segment is compared to all the cluster centers in a consistent way with the comparison in the clustering phase. The cluster with the smallest distance to the segment is the cluster winner similarly to the competitive learning principle.

c) The cluster winner is the one that best matches the segment and it is chosen to provide the prediction value for the current iteration. The prediction value, formed as an average of the last values of all contained segments, is extracted from the cluster winner.

d) The value is calculated in accordance to the chosen cluster's prediction value (7)

e) The calculated value is appended to the end of the time series and process is repeated from step a) until the desired future values are obtained.

$$y(n+1) = \mathrm{y(n)} * \mathrm{c(n-1)} + \mathrm{c(n-1)}, \tag{7}$$

Here c is the cluster center of the best matching segment and n is the current length of the series

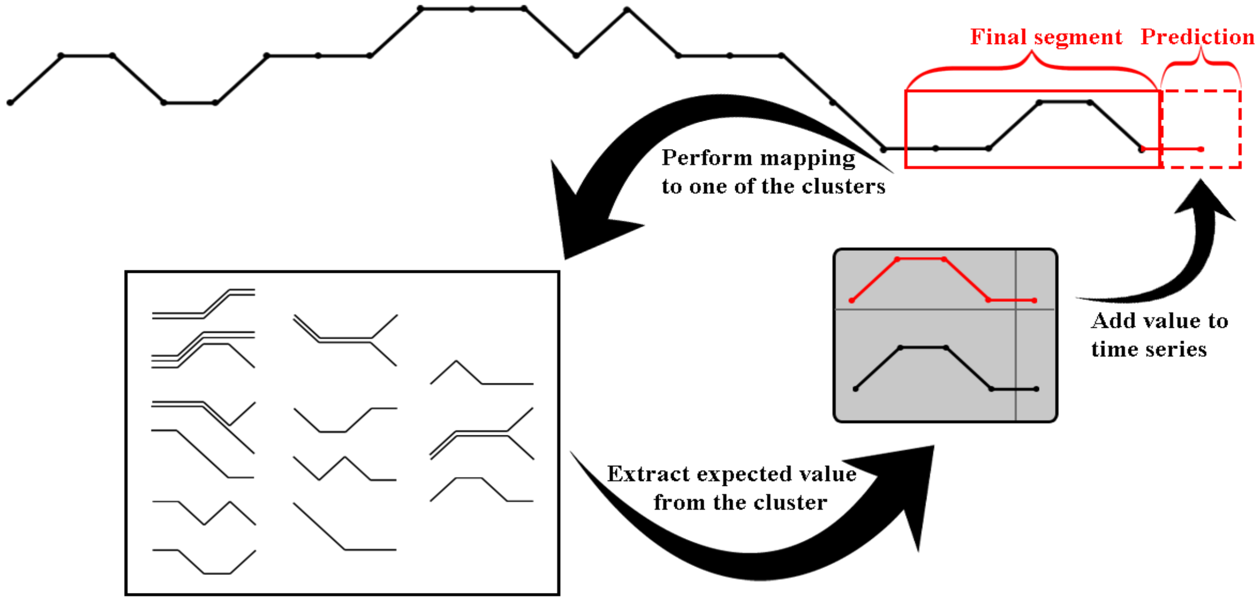

Fig. 5. The main prediction cycle

4. Conclusion and future work

Although the algorithm's performance is expected to be high because of its simplicity, the prediction capabilities of the proposed approach are yet to be evaluated. Moreover, the algorithm does not take into consideration the possibility of approaching the zero and going towards negative values which may be a valid concern in some application fields. In addition to that a comparison needs to be conducted between this approach and various other time series prediction algorithms to evaluate its benefits and drawbacks.

References

[1]. Box G. E. P., G. M. Jenkins. Time Series Analysis: Forecasting and Control, San Francisco, Holden-Day, 1970, 575 p.

[2]. Everitt, B. Cluster analysis, second edition. Social Science Research Council, London, 1980.

[3]. Hamilton, J. Time Series Analysis. Princeton University Press, ISBN: 0-691-04289-6, 1994, 799 p.

[4]. Touretzky, D., K. Laskowski. Neural Networls for Time Series Prediction. 15-486/782: Artificial Neural Networks, Lectures, Carnegie Mellon Univeristy, Fall 2006.

[5]. Xu, R., D. Wunsch. Clustering (IEEE Press Series on Computational Intelligence), 2009, 358 p.

[6]. Baretto, G. A. Time Series Prediction with the Self-Organizing Map: A Review. Studies in Computational Intelligence (SCI) 77, 2007, pp. 135-158.

[7]. Faraway, J. Time series forecasting with neural networks: a comparative study using the airline data. Appl. Statist., Vol. 47, No. 2, 1998. pp. 231-250.

[8]. Fu, Q., H. Fu, Y. Sun. Self-Exciting Threshold Auto-Regressive Model (SETAR) to Forecast the Well Irrigation Rice Water Requirement. Nature and Science, Vol. 2 no. 1, 2004, pp. 36-43.

[9]. Huang, W., Y. Nakamori, S. Wang, H. Zhang. Select the Size of Training Set for Financial Forecasting with Neural Networks. Lecture Notes in Computer Science, Vol. 3497/2005, Springer-Verlag, 2005, pp. 879-884.

[10]. Sanchez-Marono, N., O. Fontela-Romero, A. Alonso-Betanzos, B. Guijarro-Berdinas. Self-organizing maps and functional networks for local dynamic modeling. ESANN'2003 proceedings - European Symposium on Artificial Neural Networks, Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 39-44.

[11]. Simon, G., J. A. Lee, M. Verleysen. Unfolding preprocessing for meaningful time series clustering. Neural Networks Vol. 19, 2006, pp. 877-888.

[12]. Vesanto, J. Using the SOM and Local Models in Time-Series Prediction, Proceedings of Workshop on Self-Organizing Maps (WSOM'97), Espoo, Finland, 1997, pp.209-214.

[13]. Virili, F., B. Freisleben. Nonstationarity and Data Preprocessing for Neural Network Predictions of an Economic Time Series. International Joint Conference on Neural Networks, Vol.5, 2000, pp.129-134.

For contacts

Dr. Ventsislav Nikolov
Senior Software Developer
Eurorisk Systems Ltd.
31, General Kiselov Str., 9002 Varna, Bulgaria
E-mail: vnikolov at euroriskystems dot com

Sivo Daskalov
Eurorisk Systems Ltd.
31, General Kiselov Str., 9002 Varna, Bulgaria