Bulletin of the *Transilvania* University of Braşov • Vol 5(54), No. 2 - 2012
Series III: Mathematics, Informatics, Physics, 147-152

K-NEAREST NEIGHBOR ALGORITHM FOR UNIVARIATE TIME SERIES PREDICTION

Adela SASU[1]

Abstract

The paper contains a description of a k-nearest neighbor based method used for the univariate time series prediction problem. The method presented is inspired by machine learning techniques used for classification, here extended to perform regression. Experimental results and comparisons with the predicted values obtained by ARIMA show a good behavior of this approach.

2000 *Mathematics Subject Classification:* 91B70, 91B84, 60G25.
Key words: time series forecasting, ARIMA, k-nearest neighbor, prediction.

1 Introduction

The paper presents time series prediction using the k-nearest neighbor method (k-NN) and compares the results obtained by this approach to those given by ARIMA, a classical time series forecasting method. Despite its conceptual simplicity, it was proved to be effective for classification problems. Although initially designed as a classification model, it was further extended to predict time series – see [4], [5], [11].

2 K-Nearest Neighbor

k-NN is a nonparametric classification method, based on the measurement of a point's similarity to a training set containing patterns for which class labels are supplied. k-NN is a memory-based method and does not build a model through learning. The classification is made by aggregating the values provided by the training patterns in the vicinity of the current point.

k-NN can be also used for regression, when $y_i \in \mathbb{R}$. In this case, one computes the $\mathbf{x}$ associated value as an average of the ones associated to the k nearest neighbors of $\mathbf{x}$:

$$y = \sum_{i=1}^{k} \frac{y_{o(i)}}{k}. \tag{1}$$

[1] Faculty of Mathematics and Informatics, *Transilvania* University of Braşov, Romania, e-mail: asasu@unitbv.ro

A better accuracy of the prediction is reported to be obtained when one weights the neighbors, according to their closeness to $\mathbf{x}$:

$$y = \frac{\sum\limits_{i=1}^{k} w_{o(i)} y_{o(i)}}{\sum\limits_{i=1}^{k} w_{o(i)}} \tag{2}$$

where $w_{o(i)}$ is a non-decreasing function of the distance between $\mathbf{x}$ and $\mathbf{x_{o(i)}}$. This approach gives more importance to the closest neighbors. For example, one can use weights that are (approximately) inverse proportional to the distance:

$$w_{o(i)} = \frac{1}{\varepsilon + d\left(\mathbf{x}, \mathbf{x_{o(i)}}\right)},\ i \in \{1, 2, \ldots, k\}. \tag{3}$$

For constant $w_{o(i)}$, (2) becomes (1).

In order to use the k-NN method in a time series-prediction problem, one has to:

1. choose the concrete similarity functions used to find the closest neighbors

2. decide how to effectively produce the prediction.

For the former step, let $\{x_1, x_2, \ldots, x_n\}$ be the series to be analyzed. We search for k subsequences of length ws inside the given series, that are closest to the suffix vector $\{x_{n-ws+1}, \ldots, x_n\}$, with $ws > 1$. These k-subsequences are of the form:

$$\{x_{q_1}, x_{q_1+1}, \ldots, x_{q_1+ws-1}\}, \ldots, \{x_{q_k}, x_{q_k+1}, \ldots, x_{q_k+ws-1}\}$$

where $1 \leq q_i \leq n - ws$, $i = \{1, 2, \ldots, k\}$. The L_p metric can be used to compute the distances between series' suffix and a neighbor. The values k and ws are parameters of the prediction model.

There are two approaches for the latter step above. One of them is based on Giles, Lawrence, and Tsoi work in [1] and consists of predicting the direction of variation for the next value. There are three cases: increasing trend, decreasing trend and constant value. One considers that the trend is constant when the variation does not exceed a predefined percent. The second approach is based on effectively computing the predicted values ([5], [11]). For this case, suppose that we have to predict the next p values, *i.e.* to produce the values $\{x_{n+1}, \ldots, x_{n+p}\}$. The process is an iterative one, each step i predicts the next value x_{n+i} $(1 \leq i \leq p)$. This latter approach is the one we are further considering. The algorithm is given in Figure 1. A visual representation of this method is given in Figure 2.

Remark 1. *1. The initial series is successively extended with the predicted values;*

2. At every iteration the suffix has the same length ws, but is successively shifted right, i.e. *the suffix at the $i + 1$th step is obtained from the one at the previous step by removing its first element and appending the value predicted at step i;*

```
Input: univariate series $\{x_1, x_2, ..., x_n\}$
for $i = \overline{1, p}$
    look for the closest sequences
    $\{x_{q_1}, x_{q_1+1}, ..., x_{q_1+ws-1}\}, ...,$
    $\{x_{q_k}, x_{q_k+1}, ..., x_{q_k+ws-1}\}$
    to the suffix $\{x_{n+i-ws}, .., x_{n+i-1}\}$.
    Compute and estimation of $x_{n+i}$ based on
    $x_{q_1^i+ws}, ..., x_{q_k^i+ws}$
    Append $x_{n+i}$ to the end of $\{x_1, ..., x_{n+i-1}\}$
end for
Output: the predicted values are $\{x_{n+1}, ..., x_{n+p}\}$
```

Figure 1: The pseudocode for the k-nearest neighbor algorithm, which predicts p successive values for a univariate time series.

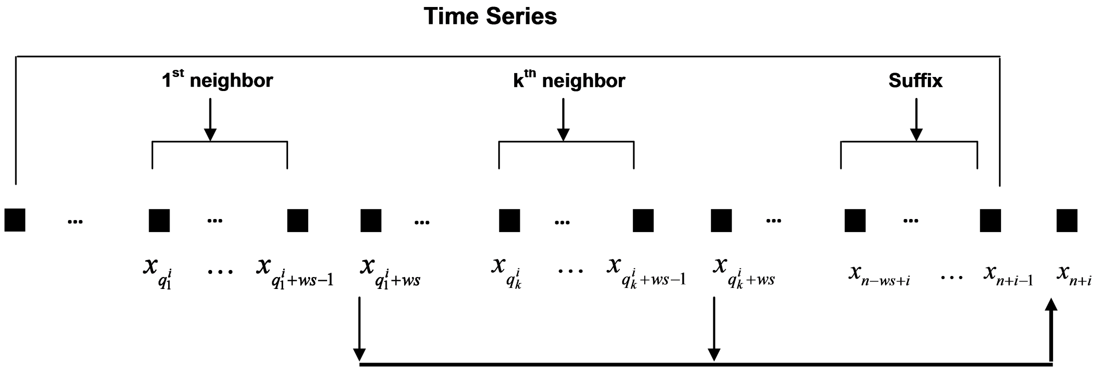

Figure 2: Computing the ith predicted output based on the current k nearest neighbors for the suffix.

3. *As the successive suffixes differ during the iterations, one might not obtain the same k closest neighbors for step i as for step $i+1$.*

Let

$$\mathbf{v_t^i} = (x_{q_t^i}, \ldots, x_{q_t^i+ws-1}) \tag{4}$$

be a sequence defining a neighbor and

$$\mathbf{s_i} = (x_{n-ws+i}, \ldots, x_{n+i-1}) \tag{5}$$

be the suffix of the time series, where $1 \leq t \leq k$. The predicted value is computed as:

$$x_{n+i} = f\left(\mathbf{s_i}, \mathbf{v_1^i}, g\left(x_{q_1^i+ws}, \mathbf{v_1^i}, \mathbf{s_i}\right), \mathbf{v_2^i}, g\left(x_{q_2^i+ws}, \mathbf{v_2^i}, \mathbf{s_i}\right), \ldots, \mathbf{v_k^i}, g\left(x_{q_k^i+ws}, \mathbf{v_k^i}, \mathbf{s_i}\right)\right) \tag{6}$$

where $g(x, \mathbf{v}, \mathbf{s})$ performs an adjustment of the predicted output x taking into account the relative position between the suffix $\mathbf{s}$ and the neighbor $\mathbf{v}$, and f is a function that performs the weighting of the values predicted by each considered neighbor. Various forms for the function g are given in [10], sections 4.2.3 and 4.2.4.

The necessity of the function g can be explained by the difference between the level of the suffix and the level of the neighbors. We considered two variants for function g: one that translates the first term of a considered neighbor onto the first term of the suffix; the later translates the average of the neighbor series onto the average of the suffix. After such a translation occurs, the translated value following the neighbor sequence can be used in estimation of the next value following the suffix. These translations are taken into account when computing the distance between the neighbor and the suffix.

More clearly, when we perform a translation that brings the first term of a neighbor $(x_a, \ldots, x_{a+ws-1})$, on the first term of the suffix $(x_{n-ws+i}, \ldots, x_{n+i-1})$, we compute:

$$\Delta_a = x_{n-ws+i} - x_a \tag{7}$$

and the distance between the two sequences is computed as:

$$d\left((x_a, \ldots, x_{a+ws-1}), (x_{n-ws+i}, \ldots, x_{n+i-1})\right) = \sum_{j=0}^{ws-1} |x_{n-ws+i+j} - x_{a+j} - \Delta_a| \tag{8}$$

After selecting the k nearest neighbors of the suffix, the predicted value corresponding to the neighbor $(x_{q_t^i}, \ldots, x_{q_t^i+ws-1})$, *i.e.* $x_{q_t^i+ws}$ will be $x_{q_t^i+ws} + \Delta_{q_t^i}$, $1 \le i \le p$, $1 \le t \le k$ and thus

$$g\left(x_{q_t^i+ws}, \mathbf{v_t^i}, \mathbf{s_i}\right) = x_{q_t^i+ws} + \Delta_{q_t^i} = x_{q_t^i+ws} + x_{n-ws+i} - x_{q_t^i}.$$

3 Experimental results and conclusions

Here we present some experimental results for 7 case studies, for monovariate time series. All of these are real data sets and they describe various economic processes, as found in [12]–[18]. The results were assessed by the accuracy function MAPE:

$$MAPE = \frac{1}{l} \sum_{t=1}^{l} \frac{|Y_t - \hat{Y}_t|}{|Y_t|} \cdot 100, \; (Y_t \neq 0), \tag{9}$$

where Y_t is the actual value, $\hat{Y}_t$ is the estimated value, and l is the number of predictions made. Lower values for MAPE means a better fit of the predicted data.

Similar experiments were performed by the author in papers [6], [7], [8], [10]. The current results are performed for actual data, up to 2012. The MAPE scores for ARIMA and k-NN for 7 data sets are given in Table 1.

The MAPE values for k-NN are $4 - 50\%$ lower than the corresponding values obtained by ARIMA, which is a classical method used for time series prediction [2]. From these results, we can conclude that k-NN is a good candidate method for time series prediction.

Series	**ARIMA MAPE**	**k-NN MAPE**
Unemployment [12]	8.39	6.07
Retail_NSA [13]	1.75	1.68
Retail_SA [14]	1.28	0.56
Construction [15]	4.97	4.05
Education [16]	0.72	0.56
Exports [17]	8.58	7.40
Gas [18]	18.45	10.13

Table 1: Experimental results obtained by ARIMA and the k-NN algorithm.

References

[1] Giles C.L., Lawrence S., Tsoi A.C., *Noisy Time Series Prediction using Recurrent Neural Networks and Grammatical Inference*, Machine Learning, Volume 44, Numbers 1-2, pp. 161-183(23), Springer, 2001

[2] Mills, T.C., *Time Series Techniques for Economists.* Cambridge University Press, 1990

[3] Mitchell, T.M., *The need for biases in learning generalizations*, CBM-TR 5-110, Rutgers University, New Brunswick, N.J., 1980

[4] Oswald, R.K., Scherer, W.T., Smith, B.L., *Traffic Flow Forecasting Using Approximate Nearest Neighbor Nonparametric Regression*, A Research Project Report No. UVACTS-15-13-7 For the National ITS Implementation Research Center, December 2001

[5] Plummer, E.A., *Time Series Forecasting with Feed-Forward Neural Networks: Guidelines and limitations*, Master Thesis, 2000

[6] Sasu, A., *An Application of ARIMA Models*, Bulletin of the Transilvania University of Braşov 11(46), Series B1. Transilvania Univ. Press-Brasov, 61-68, ISSN 1223-964X, 2005

[7] Sasu, A., *Trends in Economic Time Series*, 7-th European Conference E-COMM-LINE 2006, Bucureşti, C7-60-06, ISBN 973-88046-0-4, ISBN 978-973-88046-0-9, 2006

[8] Sasu, A., *Time Series Forecasting using ARIMA Models*, 8-th European Conference E-COMM-LINE 2007, 20-22 Sept., Bucureşti, ISBN-10: 973-88046-6-3, ISBN-13: 978-973-88046-6-1, 67e,4 pg., 2007

[9] Sasu, A., *A strategy for computing the parameters of k-NN*, Bulletin of the Transilvania University of Braşov, Vol 15(50), Series III: Mathematics, Informatics, Physics, 483-486, ISSN 1223-964X, 2008

[10] Sasu A., *Univariate time series analysis*, Ph.D. Thesis, 2009

[11] Yakowitz, S., *Nearest-neighbor methods for time-series analysis.* Journal of Time Series Analysis, Vol. 8, No. 2, 235-247, 1987

[12] http://research.stlouisfed.org/fred2/series/NYUR

[13] http://research.stlouisfed.org/fred2/series/RSAFSNA

[14] http://research.stlouisfed.org/fred2/series/RRSFS

[15] http://research.stlouisfed.org/fred2/series/CACONS

[16] http://research.stlouisfed.org/fred2/series/MIEDUHN

[17] http://research.stlouisfed.org/fred2/series/EXPGE

[18] http://research.stlouisfed.org/fred2/series/GASPRICE/