Hybrid Criteria for Nearest Neighbor Selection
with Avoidance of Biasing for Long Term Time Series Prediction

Syed Rahat Abbas and Muhammad Arif

Pakistan Institute of Engineering and Applied Sciences
Department of Computer and Information Sciences
P. O. Nilore, Islamabad, Pakistan

Abstract. Nearest neighbor is pattern matching method for time series prediction in which most recent values of the time series are compared with previous available values and forecasting is achieved by finding the best match pattern (nearest neighbor). Usually Euclidean distance is used to check the similarity of pattern. In this paper two hybrid criteria of pattern matching are being proposed and evaluated for multistep-ahead time series prediction. The first selection criterion is hybrid of “Maximum distance and Cross-Correlation” and second is hybrid of “Manhattan distance and Cross-correlation”. Better forecasting has been achieved using these algorithms.

1 Introduction

Time series prediction plays an important role in management of many systems. A huge pyramid of prediction methods are available based upon simple regression to very complex machine learning methods. Each method has pros and cons.

Nearest neighbor method is a pattern matching method in which the most recent pattern of the time series (reference pattern) is matched with all the available past patterns (candidate patterns). The prediction is carried out by the next value of the best matched pattern.

Nearest neighbor method was initially proposed by Cover and Hart [1]. In different forms it has been used for classification and prediction problems. Modifications in the nearest neighbor method were carried out over time to time. Time series prediction using delay coordinate embedding was proposed in [2]; the mixture of direct and iterated method for prediction using four nearest neighbors with interpolation was carried out and method was applied for Santa Fe time series prediction competition in 1992. The nearest neighbor method with upsampling and cross-correlation was carried out [3]. The comparison of nearest neighbor method with other method for prediction of foreign exchange shows that results are data dependent [4]. Simultaneous nearest neighbor method performed marginally better than ARIMA and random walk methods as reported in [5]. The prediction of chaotic behavior of market response is carried out using multivariate nearest neighbor method for precise prediction [6]. Divide and conquer approach to develop pair-wise class nearest neighbor method was proposed in [7]. Locally adaptive metric nearest-neighbor classification method was also proposed in [8]; they have used updating of weighted distance for getting optimal nearest neighbors. It was proposed that advanced data structures significantly reduce the execution time of nearest neighbor
regression [9]. Subset features space was used by to improve nearest neighbor classification [10]. Subspace of candidate was used by [11] for fast search of nearest neighbors. Discriminate adaptive nearest neighbor classification was suggested by [12]. They used local linear discriminant analysis to estimate an effective metric for computing neighborhoods. The nearest neighbor method in economics is also used recently [13]. The hybrid of Euclidean distance and normalized cross-correlation method [14] is proposed by us; which provided better forecasting than classical nearest neighbor method.

In this paper the hybrid criterion of maximum distance with normalized cross-correlation and Manhattan distance with normalized cross-correlation is being proposed.

2 Proposed algorithm for time series prediction

In nearest neighbor method last few values of the available time series are taken which are considered as referenced pattern. The number of values of the time series used for matching are called window size (‘w’). The reference pattern is compared with all available patterns (candidate patterns) of same length. The forecasting is achieved as the next value of best matched pattern. The schematic of nearest neighbor algorithm is illustrated as Fig 1.

![Fig 1: Schematic For Nearest Neighbor Search](image)

The main steps are (a) window size selection (b) pattern matching (3) prediction procedure. Following is the detail of these steps.

2.1 Search for optimal window size

For nearest neighbor algorithm the first maximum after lag=0 of Auto-Correlation Function (ACF) plot gives the useful window size [3], [15]. The window sizes (‘w’) of six series studied in this paper are approximated by ACF plot and shown in Table 1. The description of time series and their sources are presented in section 3.
Table 1: Window Size for Time Series

<table>
<thead>
<tr>
<th>Series</th>
<th>Window Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sunspot</td>
<td>10</td>
</tr>
<tr>
<td>IOWA Electricity Series</td>
<td>12</td>
</tr>
<tr>
<td>River Series</td>
<td>12</td>
</tr>
<tr>
<td>ESTSP08 First Series</td>
<td>12</td>
</tr>
<tr>
<td>ESTSP08 Second Series</td>
<td>7</td>
</tr>
<tr>
<td>ESTSP08 Third Series</td>
<td>24</td>
</tr>
</tbody>
</table>

The ACF plot for ESTSP Competition Series (2nd) is shown in Fig 2.

![ACF plot of ESTSP08 Competition Series (2nd)](image)

**2.2 Usual Pattern Matching Criterion**

Usually in nearest neighbor algorithm the best match pattern is selected which has the least Euclidean distance from the reference pattern. The Euclidean distance ‘Ed’ between two vectors ‘X’ and ‘Y’ is given by Error! Reference source not found..
$E_d = \sqrt{\sum (X(i) - Y(i))^2}$

2.3 Proposed Matching Criterion

Euclidean distance based search in the standard nearest neighbor gives similarity in terms of the distance between the two patterns without considering the shape of two patterns. Two other distances i.e. Maximum distance and Manhattan distance are also used for pattern matching. The maximum and Manhattan distance between two vectors $X$ and $Y$ are given by

$$r_{\text{max}} = \max |X(i) - Y(i)|$$

$$r_{\text{man}} = \sum |X(i) - Y(i)|$$

The distances are amplitude dependent for example if $\sin(x)$ and $5\sin(x)$ is considered they will give high value of distance between them. We can also use zero order cross correlation to find the best nearest neighbor in terms of shape. Zero order cross correlation of two vectors can be described as follows, If ‘$X$’ and ‘$Y$’ are two vectors, the normalized cross-correlation $X_{\text{corr}}$ with delay ‘TD’ is defined as

$$X_{\text{corr}}(TD) = \frac{\sum X(i)Y(i + TD)}{\sqrt{\sum X^2(i) \cdot \sum Y^2(i)}}$$

For zeroth order cross-correlation $TD = 0$. The normalized cross-correlation is amplitude independent. It will give value ‘1’ (perfect match) for $\sin(x)$ and $5\sin(x)$.

An example of calculation is being presented to highlight the effect of these distances and cross correlation. Let us consider the following set of patterns sampled with time step 0.1.

$x = \sin t$

$y = 2.5\sin t$

$z = \cos(t + 0.3) \quad t \in [0, 2.5\pi]$

$v = 2\sin(t + 0.4)$

$p = 0.3\cos t$

Let ‘$x$’ is our reference pattern and other four are candidate patterns. Normalized zero order cross-correlation, Maximum and Manhattan distances of the reference pattern to the candidate patterns are given in Table 2. Maximum value of cross-correlation is considered as the measure of closest pattern and minimum value of distance is considered as the criterion for the closest pattern. If we consider only cross-correlation as selection criterion then the closest match of pattern ‘$x$’ is ‘$y$’ and
if we consider the maximum or Manhattan distance only, the closed pattern of ‘x’ is ‘p’. But Error column shows that closest pattern is ‘v’.

Cross-correlation based search can give the best nearest neighbor having similar shape but it is possible that the amplitudes of the two patterns may differ a lot. we have proposed hybrid selection criteria using normalized cross correlation and maximum distance and cross correlation with Manhattan distance. The normalized cross-correlation is invariant to the amplitude of the patterns but depends on the shape of the two patterns. Combining the two selection criteria can give us better pattern selection considering both shape and amplitude.

<table>
<thead>
<tr>
<th>Candidate Series</th>
<th>Series Name</th>
<th>Normalized Cross-correlation</th>
<th>Manhattan Distance</th>
<th>Maximum Distance</th>
<th>Error with Actual Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.5 Sin(t)</td>
<td>y</td>
<td>1.000</td>
<td>74.9352</td>
<td>1.4999</td>
<td>1.492</td>
</tr>
<tr>
<td>Cos(t+0.3)</td>
<td>z</td>
<td>-0.179</td>
<td>74.5138</td>
<td>1.6096</td>
<td>1.384</td>
</tr>
<tr>
<td>2Sin(t+0.4)</td>
<td>v</td>
<td>0.928</td>
<td>62.4379</td>
<td>1.1470</td>
<td>0.672</td>
</tr>
<tr>
<td>0.3 Cos(t)</td>
<td>p</td>
<td>0.126</td>
<td>49.7579</td>
<td>1.0440</td>
<td>1.025</td>
</tr>
</tbody>
</table>

Following is the algorithm of the hybrid selection criteria,

**Step 1:** Take the zeroth order normalized cross-correlation of the reference pattern with the candidate patterns and arrange them in descending order according to their cross-correlation values.

**Step 2:** Pick only those candidate vectors whose cross-correlation value with the reference pattern is greater than \( \theta \). We have tried different value of \( \theta \) and found that it can be taken as 0.8. If no such candidate vector exists then only maximum/Manhattan distance will be use for pattern matching.

**Step 3:** Calculate the maximum/Manhattan distance of all the patterns selected in step 2 with the reference vector and consider the best nearest neighbor having minimum maximum/Manhattan distance with the reference pattern.

Considering the above example, we will select vectors ‘y’ and ‘v’ only based on their cross correlation values with the reference pattern (Step 2). Considering the maximum or Manhattan distances of ‘y’ and ‘v’ from reference pattern ‘x’, pattern ‘v’ will be selected as the nearest neighbor. From the Table 2, it can be seen that minimum error in the forecasting of ‘x’ is achieved by using the pattern ‘v’.

In long-term forecasting, forecasted values are iterated to get multi-step ahead prediction. Hence any forecasting error occurred at a certain time will be propagated.
and increased in the later forecasting values. This makes the accuracy of the forecast value and the comparison strategies to find the nearest neighbor a critical issue for long range forecasting.

2.3.1 Avoidance of Biasing

Let for some ‘ith’ step ahead, the query vector is \([x_i \ x_{i+1} \ldots \ x_{i+n-1}]\) and the selected vector from the database is the ‘rth’ vector \([x_r \ x_{r+1} \ldots \ x_{r+w-1}]\). For ‘(i+1)th’ step, the query vector will become \([x_{i+1} \ x_{i+2} \ldots \ x_{i+n}]\). The ‘(r+1)th’ vector in the database is \([x_{r+1} \ x_{r+2} \ldots \ x_{r+w}]\). As the last value of both the query vector and ‘(r+1)th’ vector are exactly same so the search in ‘(i+1)th’ step will be biased towards this ‘(r+1)th’ vector in the database. To remove this biasing effect, it is proposed that the last value of the query vector will not participate in calculating the maximum or Manhattan distance.

2.3.2 Prediction on the base of best match pattern

The best match pattern is one which has the maximum correlation value, after finding it the prediction of one value is achieved as the next value in the time series of the best matched pattern.

For the multistep-ahead prediction, the reference vector is updated by dropping the oldest value in it and padding the forecasted value at the end so that the length of the reference pattern remains intact. The new reference vector is again matched with candidate patterns and this process is iterated for required prediction steps.

3 Results and Discussion

To evaluate the proposed algorithm, three time series from the forecasting literature are studied. First series is famous Wolfer Sunspot number time series which is chaotic and a benchmark for time series prediction. 200 values were used to forecast next 50 values. The second series is monthly electricity consumption in IOWA city US. 70 data values were used to forecast next 30 values. The third series is river flow at Fair oaks, California for the period October 1906 to September 1960. First 540 values were taken to forecast next 60 values. These series are taken from time series data library by R. J. Hyndman (web: http://wwwpersonal.buseco.monash.edu.au/~hyndman/tsdl/).

In Table 3 the comparison of forecasting error by using standard Euclidean distance based nearest neighbors (SNN) algorithm and proposed hybrid algorithms are presented. In case of sunspot series the normalized mean squared error (NMSE) was reduced from 2.581 to 0.8143 in case of hybrid of maximum distance. The hybrid of Manhattan distance did not improved results in this case. For electricity consumption time series the NSME with classical nearest neighbor method was 0.4915 which reduced to 0.2641 in case of Manhattan and further decreased to 0.1943 in case of maximum distance. In case of River flow series the NMSE reduced from
0.9563 to 0.9158 in case of hybrid of maximum distance. Manhattan distance in this case degrades the results.

Both of the proposed algorithms also used to forecast ESTSP’08 Competition time series. For dataset 1 only third column is used and exogenous inputs are ignored. Dataset 3 is very long; its subset is taken using visual guess. Three subsection of dataset 3 are taken for nearest neighbor search i.e. (13001:14500), (21501:23500), (30001:31614).

Table 3: Forecasting Results using Proposed Algorithm

<table>
<thead>
<tr>
<th>Sr. No</th>
<th>Time Series</th>
<th>NMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>SNN</td>
</tr>
<tr>
<td>1</td>
<td>Sunspot</td>
<td>2.581</td>
</tr>
<tr>
<td>2</td>
<td>IOWA Elec</td>
<td>0.4915</td>
</tr>
<tr>
<td>3</td>
<td>River</td>
<td>0.9563</td>
</tr>
</tbody>
</table>

In case of hybrid algorithm of Euclidean distance and cross-correlation [14] the NMSE for Sunspot time series was 0.747, for IOWA elec. time series was 0.4930 and for River series it was 0.8956. So for different time series different algorithm performed well and there is not general conclusion.
Using classical NN for Sunspot Series

Using proposed algorithm for Sunspot series

Using classical NN for IOWA Electric Series

Using proposed algorithm for IOWA Electric Series

Fig 3: Comparison of classical nearest neighbor method and proposed algorithm
It has been found that hybrid criterion of nearest neighbor selection based on maximum distance and cross-correlation performed better than that of Manhattan distance (Table 3, in Table 3 Xcorr is abbreviation of cross-correlation). In future hybrid of other distances with cross-correlation can be studied.

4 Conclusion

The hybrid criteria based on Maximum/Manhattan distances and zeroth order normalized cross-correlation are proposed. It is found that forecasting results for Sunspot series, IOWA electricity consumption series and River Flow series has been improved especially when maximum distance and cross-correlation is used. The forecasting results for ESTSP’08 competition series have been submitted.

References


