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

Syed Rahat Abbas and Muhammad Arif

Jacqueline UFITIMANA

November 12, 2008

Outline

1. Introduction

- Definitions
- About Nearest Neighbor method
- Proposed method

2. Proposed algorithm for time series prediction

- Optimal window size
- Usual Pattern Matching Criterion
- Proposed Matching Criterion
  - Steps of the Algorthim
  - Avoidance of Biasing
  - Prediction on the base of best match pattern

3.Results and Discussions

4.Conlusion

Definitions

- Nearest neighbor

  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.
- Window size

  The number of values of the time series used for matching.

About Nearest Neighbor Method

- Nearest neighbor method was initially proposed by Cover and Hart.
- It has been used for classification and prediction problems.
- Modifications in it were carried out time to time.
- Examples:
  - Time series prediction using delay coordinate embedding.
  - the mixture of direct and iterated method for prediction using four nearest neighbors with interpolation

- The nearest neighbor method with upsampling and cross-correlation.
- Divide and conquer approach to develop pair-wise class.nearest neighbor method.
- Locally adaptive metric nearest neighbor classification method.(using updating of weighted distance for getting optimal nearest neighbors.)
- Discriminate adaptive nearest neighbor classification.

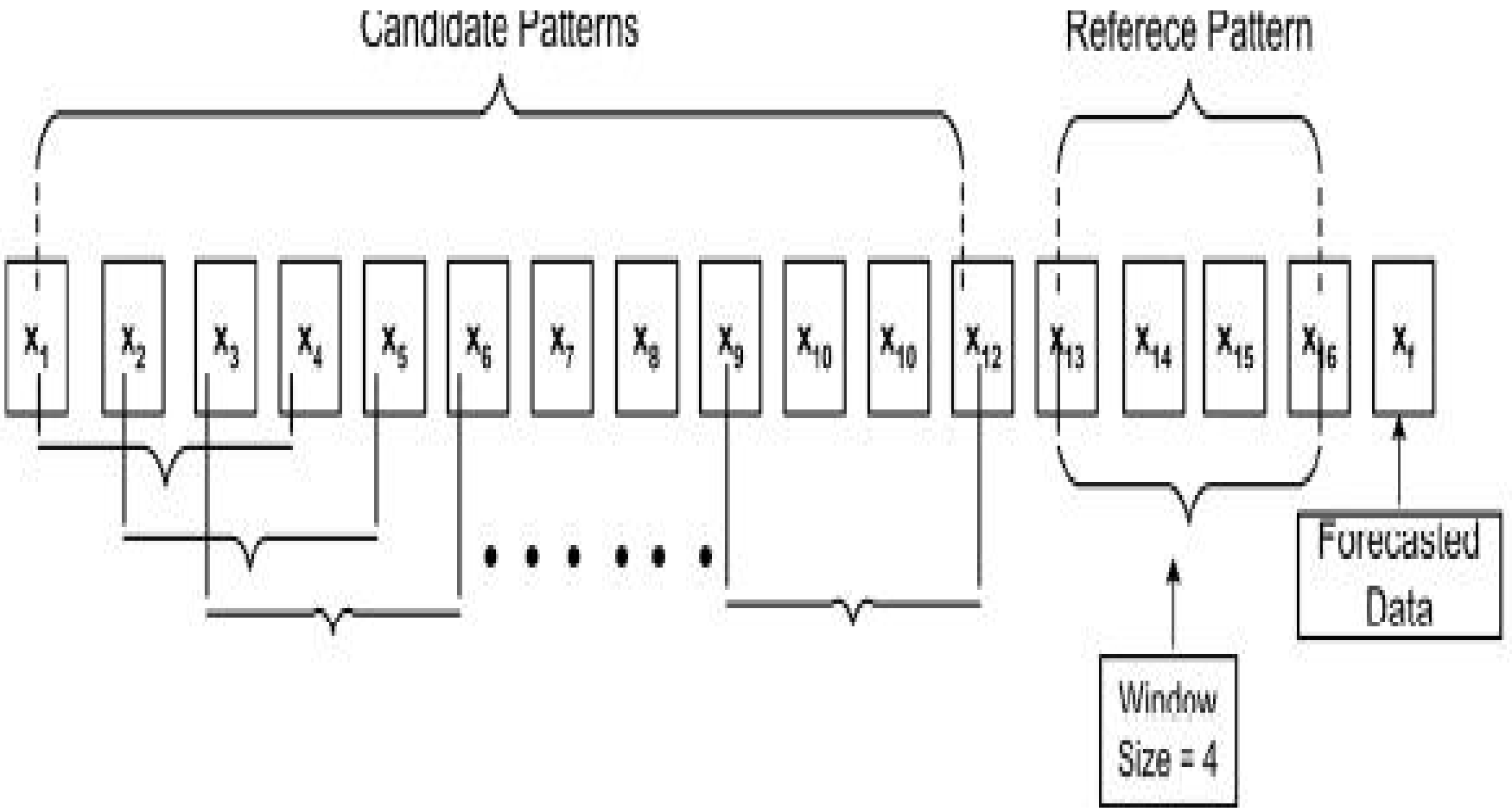
Proposed method

- The hybrid of Euclidean distance and normalized cross-correlation method.
- This method provides better forecasting than classical nearest neighbor method.
- 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.

Schematic For Nearest Neighbor Search



The main steps are:

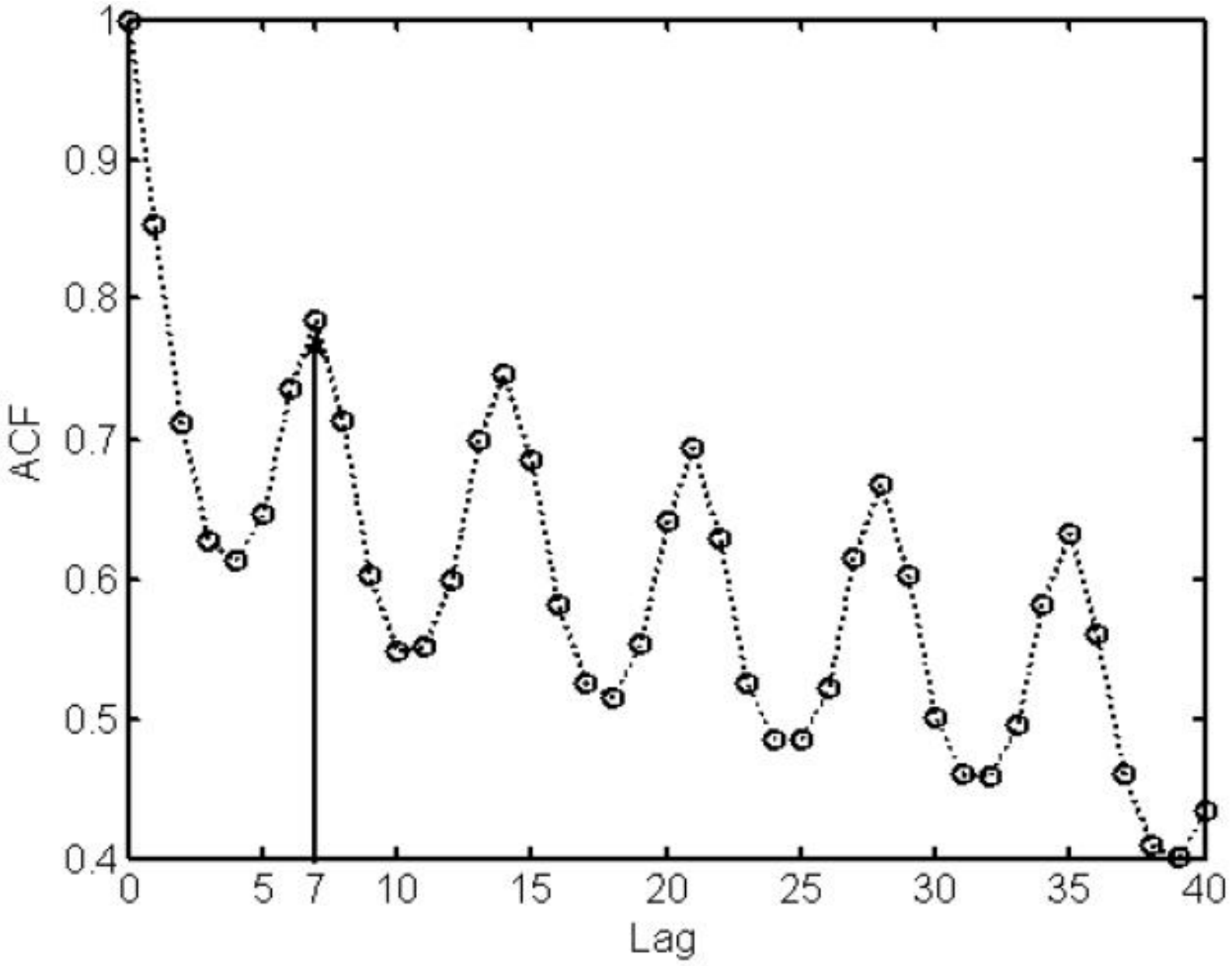
- window size selection
- pattern matching
- prediction procedure.

2.1. **Search for optimal window size**

- The first maximum after lag=0 of Auto-Correlation Function (ACF) plot gives the useful window size.
- The window sizes (‘*w*’) of six series studied in this paper are approximated by ACF plot and shown in the following Table.

Series	Window size (*w*)
Sunspot	10
IOWA Electricity Series	12
River Series	12
ESTSP08 First Series	12
ESTSP08 Second Series	7
ESTSP08 Third Series	24

The ACF plot for ESTSP Competition Series (2nd).



2.2 **Usual Pattern Matching Criterion**

- Usually in nearest neighbor algorithm the best match pattern is selected to be the one which has the least Euclidean distance from the reference pattern.

$$Ed = \sqrt{\sum_i (X(i) - Y(i))^2}$$

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 their shape.
- Maximum distance and Manhattan distance are also used for pattern matching.

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

$$r_{man} = \sum_{i}\left|X(i) - Y(i)\right|$$

The distances are amplitude dependent for example Sin(x) and 5Sin(x) 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.

If ‘ X’ and ‘Y’ are two vectors, the normalized cross-correlation with delay ‘TD’ is defined as

$$Xcorr_{norm}(TD) = \frac{\sum_i x(i) - Y(i - TD)}{\sqrt{\sum_i X^2(i) * Y^2(i)}}$$

•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 5Sin(x).

Example

- Let us consider the following set of patterns sampled with time step 0.1.

$$x = \sin t$$

$$y = 2.5 \sin t$$

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

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

$$p = 0.3 \cos t$$

Let ‘x’ is our reference pattern and other four are candidate patterns.

Distance and Cross-Correlation of Sin(t) with other series

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

Algorithm of the hybrid selection criteria

- **STEP1**:Take the zeroth order normalized cross-correlation of the reference pattern with the candidate patterns and arrange them in descending order.

- **STEP2**:Pick only those candidate vectors whose cross-correlation value with the reference pattern is greater than θ . We have tried different value of θ and found that it can be taken as 0.8. If no such candidate vector exists then only maximum/Manhattan distance will be used for pattern matching.

- **STEP3**: 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.

In our 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 previous table , it can be seen that minimum error in the forecasting of ‘x’ is achieved by using the pattern ‘v’.

Avoidance of Biasing

- Let for some i^{th} step ahead, the query vector be $\lfloor x_i \quad x_{i+1} \quad \text{L} \; x_{i+w-1} \rfloor$ and

the selected vector from the database is the r^{th} vector = $\lfloor x_r \quad x_{r+1} \quad \text{L} \; x_{r+w-1} \rfloor$ for $(i+1)^{th}$ step, the query vector will become $\lfloor x_i \quad x_{i+1} \quad \text{L} \; x_{r+w} \rfloor$.

The $(r+1)^{th}$ vector in the database is $\lfloor x_r \quad x_{r+1} \quad \text{L} \; x_{r+w} \rfloor$

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.

- Prediction on the base of best match pattern
  - The best match pattern is one which has the maximum correlation value
  - 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 are studied.
  - Wolfer Sunspot number time series where 200 values were used to forecast next 50.
  - Monthly electricity consumption in IOWA city US. 70 data values were used to forecast next 30 values.
  - River flow at fair oaks, California for the period October 1906 to September 1960. First 540 values were taken to forecast next 60 values.

using standard Euclidean
distance based nearest neighbors (SNN) algorithm
and proposed hybrid algorithms
are presented.

SR.No	Time Series	NMSE		
		SNN	Max Distance+ Xcorr	Manhatt +Xcorr
1	Sunspot	2.581	0.8143	2.5811
2	IOWA Elec	0.4915	0.1943	0.2641
3	River	0.9563	0.9158	1.4828

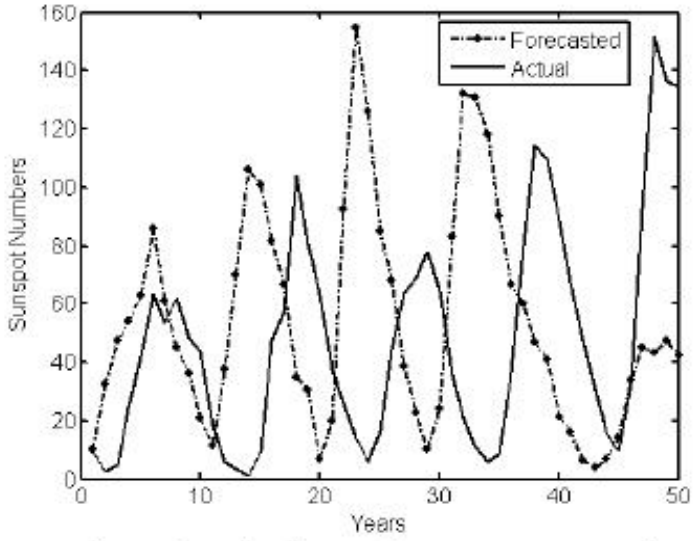
In case of hybrid algorithm of Euclidean distance and cross-correlation

the NMSE for Sunspot time series was 0.747,
For IOWA elec. time series was 0.4930

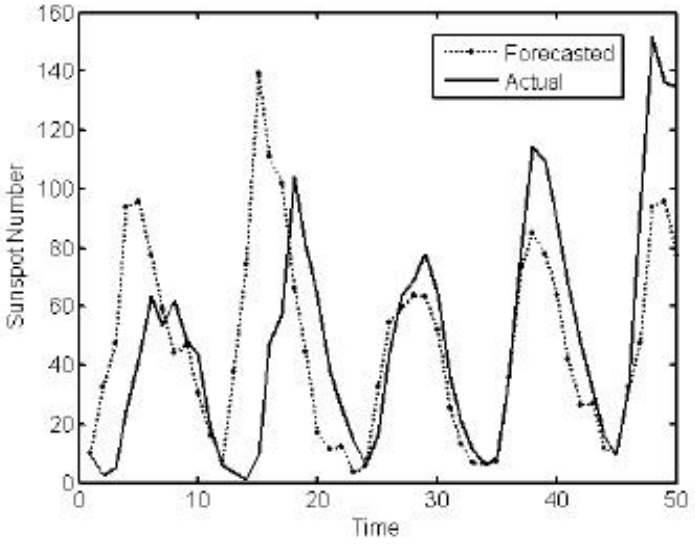
For River series it was 0.8956.

So for different time series different algorithm
performed well and there is no general conclusion.

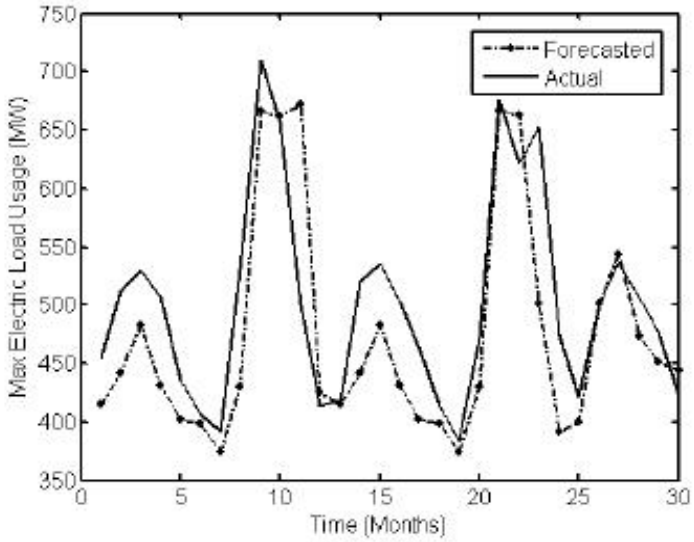
Comparison of classical nearest neighbor method and proposed algorithm



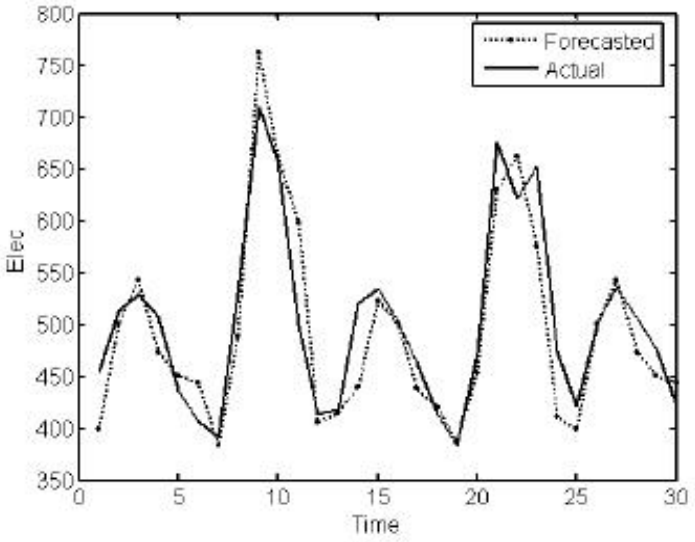
Using classical NN for Sunspot Series



Using proposed algorithm for Sunspot series

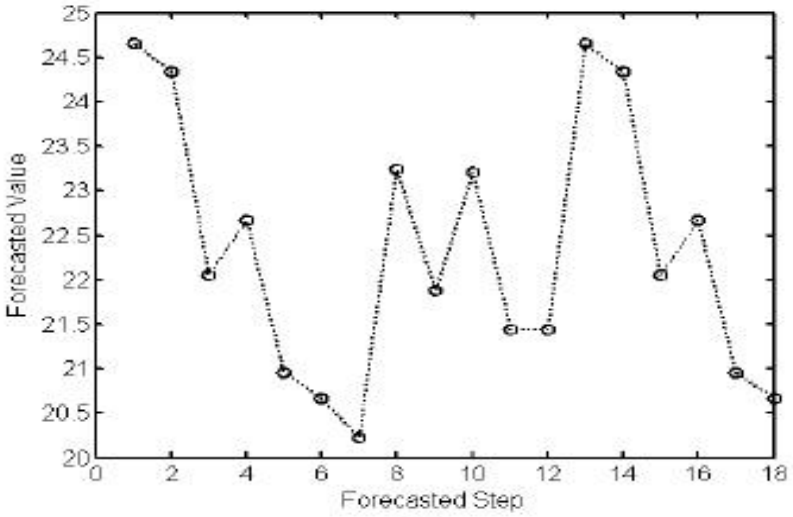


Using classical NN for IOWA Electric Series

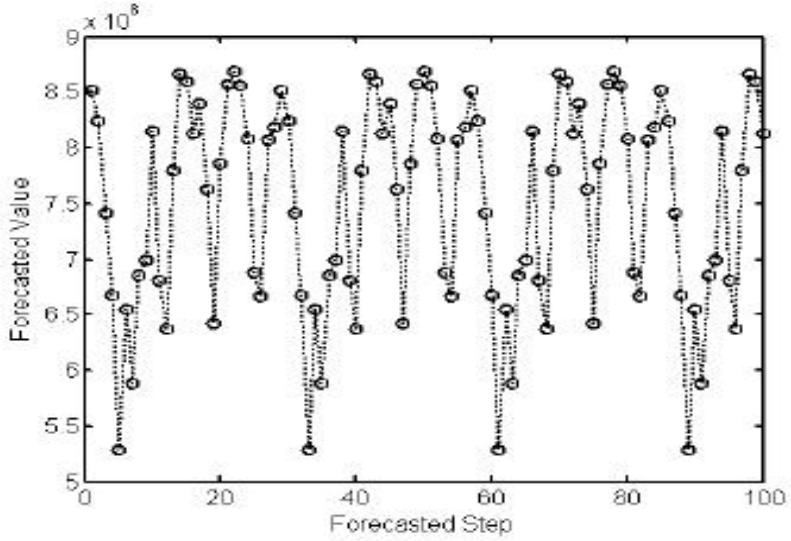


Using proposed algorithm for IOWA Electric Series

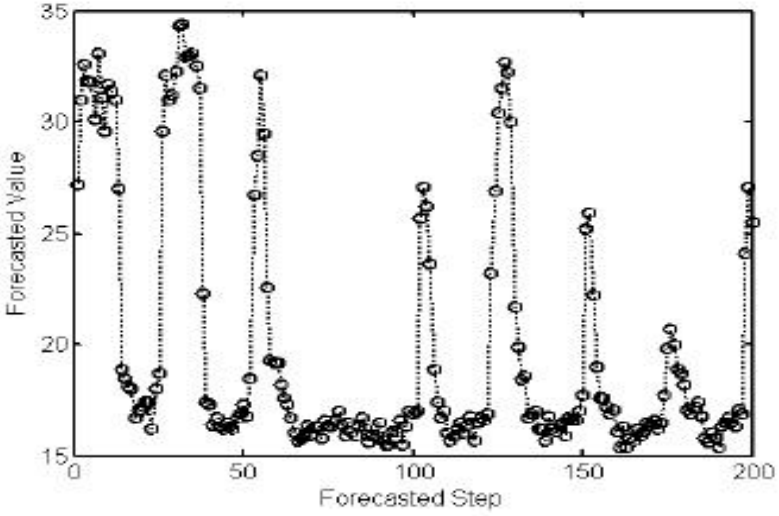
Forecasting Plots of ESTSP'08 Competition (a) Dataset 1 (b) Dataset 2 (c) Dataset 3



(a)



(b)



(c)

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

In future hybrid of other distances with cross-correlation can be studied.

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.

THANK YOU!