# Input Variables Selection Using Mutual Information for Neuro Fuzzy Modeling with the Application to Time Series Forecasting

M. M. Rezaei Yousefi, M. Mirmomeni, and C. Lucas

*Abstract*—**This paper presents a methodology to select input variables for time series prediction. A main motivation is to find some proper input variables which describe the time series dynamics properly. It is shown that even when the choice of input variables is confined to the lagged values of the process to be predicted, a nonlinear analysis of the most significant factors is crucial for improving the prediction quality. The proposed method is used to select the appropriate input variables for neuro fuzzy models utilized for time series prediction benchmark in NN3 competition as well as a second benchmark to show the generality of the claims. Results depict the effectiveness of the proposed method in proper input selection for neuro fuzzy models for prediction task.**

## I. Introduction

PREDICTING the future, which has been the goal of many research activities in the last century, is an important problem for human to prevent or to reduce loss of time and property, arising from the fear of unknown phenomena and calamities all around the infinitely large world with its many variables showing highly nonlinear and chaotic behavior [1], [2]. Time series is a collection of measurements or observations from processes and phenomena that made sequentially in time [2]. Purposes of time series analysis can be defined as identifying the nature of corresponding processes and forecasting their future values.

Input variables selection is one of the most important problems in modeling and prediction tasks. The objective is to find a subset of inputs from original input data set [3]. By proper input variables selection which yields faster and more cost-effective input variables with more generalization capability, a better perception of the system is provided [4]. The selected inputs with minimum redundancy have maximum relevance with the output variables [5].

In this paper, a methodology based on Mutual Information for input variables selection is introduced to improve the performance of prediction tasks even when lagged values of the time series are the only possible input signals. The method can be used to eliminate less useful lags, thereby reducing the number of parameters and improving generalization performance. After selecting proper input variables the Locally Linear Neuro Fuzzy (LLNF) model as general function approximator [6] can be used as a general framework to predict the main patterns of the time series due its great performance in prediction of nonlinear and chaotic time series [7], [8].

The paper consists of five sections. The problem of input variables selection and mutual information criterion along with the appropriate algorithm for choosing input variables is presented in Section II. Section III presents the main aspects of the locally linear neuro fuzzy model. Prediction of two case studies has been considered in Section IV. The first case study has been chosen from neural forecasting competition (NN3) and the second case study is the sunspot number as a natural chaotic time series which has been considered as a difficult real world case study. The last section contains the concluding remarks.

## II. Input Variables Selection Methodology

Building models with many irrelevant or unnecessary inputs may cause model works imperfectly. Generally, if the number of free parameters is small, complexity of the model is not enough to capture the dynamics of the real system and prediction will not be very accurate. Conversely, selecting too many free parameters will force model to capture also the noise contained in the data [9] which is also known as over-fitting phenomenon [6], [10]-[12]. The over-fitting problem results from model complexity. Handling this problem becomes more difficult when there are many input variables [13]. Therefore, choosing a set of most relevant and non-redundant input variables is necessary to build an appropriate model with high performance, and to improve the interpretability of the selected set of inputs [14].

The problem can be defined as considering a set of candidate input variables and selecting a subset that has the best performance (according to the predefined criteria) in a prediction model. Let $X_i$, $1 \le i \le N_l$ ($N_l$ is the number of lags) be lags of the time series as input variables. The objective is to find an optimal subset of $X_i$ containing $d$ variables ($d \le N_l$) that will be used to build an adequate model [15]. Some heuristic search strategies are needed to choose a set of suboptimal input variables. Strategies which are common in selecting regressors for linear models are Forward Selection,

Manuscript received May 21, 2007.

M. M. Rezaei Yousefi is with the Control and Intelligent Processing Center of Excellence, School of Electrical and Computer Eng., University of Tehran, Tehran, Iran (corresponding author to provide phone: +98-21-8802-7756; fax: +98-21-8877-8690; e-mail: m.rezaie@ece.ut.ac.ir).

M. Mirmomeni, is with the Control and Intelligent Processing Center of Excellence, School of Electrical and Computer Eng., University of Tehran, Tehran, Iran (e-mail: m.mirmomeni@ece.ut.ac.ir).

C. Lucas is with the Control and Intelligent Processing Center of Excellence, School of Electrical and Computer Eng., University of Tehran, Tehran, Iran and School of Cognitive Sciences, Institute for studies in theoretical Physics and Mathematics, Tehran, Iran (e-mail: lucas@ipm.ir).

Backward Elimination and Stepwise Regression [6], [16]. The same strategies can be applied to select inputs for nonlinear models. Two criteria are needed to choose a suitable subset of inputs [16]:

1. *Saliency criterion*: to direct the search by ranking input variables according to their relevance to the output.
2. *Selection criterion*: to evaluate the relevance for a subset of input variables.

In order to reduce the computational complexity in the estimation and selection procedure, model-independent approaches are needed [16]. Two simple methods are Correlation- and Partial Correlation Analysis, but these approaches show poorly when relations are nonlinear. Two other approaches are Gamma Test and mutual information [17]-[19]. Mutual information is very effective in evaluating the relevance or redundancy of each input variable, where methods based on linear relations (like the correlation analysis) may give misleading information [20]. In this paper, mutual information is used to select subset of lags of the time series to predict its future.

### *A. Mutual information: theoretical foundation*

In Probability Theory, especially in Information Theory the mutual information can be used for evaluating any arbitrary dependencies between random variables [19], [21]. In fact, the mutual information between two random variables $X$ and $Y$, is a quantity that measures the knowledge on $Y$ provided by $X$ (or conversely the amount of knowledge on $X$ shared by $Y$). If $X$ and $Y$ are independent, therefore $X$ contains no information about $Y$ and vice versa; thus the mutual information between them is zero.

The definition of mutual information originates from the Shannon Entropy [22] in the information theory. The mutual information of two random variables $X$ and $Y$ is defined as:

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) = H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X;Y) \end{aligned} \tag{1}$$

where $H(.)$ is the entropy, $H(X|Y)$, $H(Y|X)$ are the conditional entropies, and $H(X;Y)$ is the joint entropy of $X$ and $Y$ that are defined by:

$$H(X) = -\int_x p_X(x) \log p_X(x) dx \tag{2}$$

$$H(Y) = -\int_y p_Y(y) \log p_Y(y) dy \tag{3}$$

$$H(X;Y) = -\int_x \int_y p_{X,Y}(x,y) \log p_{X,Y}(x,y) dx dy \tag{4}$$

where $p_{X,Y}(x,y)$, $p_X(x)$ and $p_Y(y)$ are the joint probability density function and marginal density functions of $X$ and $Y$, respectively. The marginal density functions are given by:

$$p_X(x) = \int_y p_{X,Y}(x,y) dy \tag{5}$$

$$p_Y(y) = \int_x p_{X,Y}(x,y) dx \tag{6}$$

Mutual information is the Kullback-Leibler distance between the joint distribution $p_{X,Y}(x,y)$ and the product distribution $p_X(x).p_Y(y)$. By substituting (2), (3), (4) into (1) one has the mutual information equation:

$$I(X;Y) = \int_x \int_y p_{X,Y}(x,y) \log \frac{p_{X,Y}(x,y)}{p_X(x) p_Y(y)} dx dy \tag{7}$$

In discrete forms, the integrations are replaced by summation over all possible values that appear in the data. In order to estimate the mutual information between $X$ and $Y$ we only need to estimate $p_{X,Y}(x,y)$ by (5), (6) and (7). Histogram and Kernel methods are widespread to estimate probability density functions [23]. To overcome the curse of dimensionality in these estimation methods and to reduce the computational complexity, we use a recent estimator that estimates entropy from the average distance to the $k$-nearest neighbors, averaged over all data [19], [24]. Consider a set of $N$ input-output pairs, $z_i=(x_i,y_i)$, $i=1,\ldots,N$, which are assumed to be realizations of a random variable $Z=(X,Y)$ with density $p_{X,Y}(x,y)$. Either $X$ and $Y$ have values in $R$ or in $R^p$ and the algorithm will use the natural norm (Euclidean norm) in those spaces. Input–output pairs are compared through the maximum norm [24]:

$$\|z - z'\|_\infty = \max\{\|x - x'\|, \|y - y'\|\} \tag{8}$$

It can be considered that $k$ is a fixed positive integer, then $z_{k(i)}=(x_{k(i)},y_{k(i)})$ is the $k$-th nearest neighbor of $z_i$ (with maximum norm). It can be denoted that:

$$\varepsilon_i/2 = \|z_i - z_{k(i)}\|_\infty \tag{9}$$

$$\varepsilon_i^x/2 = \|x_i - x_{k(i)}\|, \quad \varepsilon_i^y/2 = \|y_i - y_{k(i)}\| \tag{10}$$

$\varepsilon_i/2$ is the distance from $z_i$ to its $k$-th neighbor and $\varepsilon_i^x/2$ and $\varepsilon_i^y/2$ are the distances between the same points projected into $X$ and $Y$ subspaces. Obviously, $\varepsilon_i = \max\{\varepsilon_i^x, \varepsilon_i^y\}$.

$n_i^x$ and $n_i^y$ are the numbers of sample points with $\|x_i - x_j\| \le \varepsilon_i^x/2$ and $\|y_i - y_j\| \le \varepsilon_i^y/2$. The estimation for mutual information is then:

$$\hat{I}(X;Y) = \psi(k) - \frac{1}{k} - \frac{1}{N}\sum_{i=1}^{N}\left[\psi(n_i^x) + \psi(n_i^y)\right] + \psi(N) \qquad (11)$$

where $\psi$ is the Digamma function. With a small value for $k$, this estimator has a large variance and a small bias, whereas a large value of $k$ leads to a small variance and a large bias [25]. In this paper, $k = 6$ is used.

### B. Input variables selection algorithm

This section is devoted to describe the input variables selection algorithm. This algorithm has been used beforehand for feature selection in classification and pattern recognition problems [18], [20], and [26] which is proposed by Battiti in 1994. The objective of this algorithm is to maximize relevance between inputs and output and minimizes the redundancy of selected inputs. This algorithm computes $I(T;l)$ and $I(l;l')$, where $l$ and $l'$ are individual inputs and $T$ is output. The goal of these two terms is to select relevant input with the output which has least dependency with other selected inputs [26]. The algorithm is as follows:

1) *Initialization*: Set $L$ to 'initial set of $n$ inputs'; $S$ to 'empty set'; and $T$ to 'output'.
2) *Computation of the mutual information with the output*: For each input $l \in L$ compute $I(T;l)$.
3) *Choice of the first input*: Find the input $l$ that maximizes $I(T;l)$; Set $L \leftarrow L - \{l\}$ , $S \leftarrow \{l\}$ .
4) *Greedy selection*: Repeat until desired number of input variables are selected:
   a) *Computation of the mutual information between variables*: For all couples of variables ($l$, $S$) with $l \in L$ , $s \in S$ ; compute $I(l; s)$, if it is not already available.
   b) *Selection of the next input*: Chose the input $l \in L$ as the one that maximizes $I(T;l) - \frac{\beta}{|S|}\sum_{s \in S} I(l;s)$ ; set $L \leftarrow L - \{l\}$ , $S \leftarrow S \cup \{l\}$ .
5) Output the set $S$ containing the selected inputs.

To consider redundancy between input variables, Battiti imports $\beta$ as a parameter to adjust the relative importance of mutual information between the candidate input and the already selected inputs with respect to the mutual information with the output. If $\beta = 0$ the algorithm only attempts to maximize mutual information with output, so the redundancy between input variables is never considered. If $\beta$ increases, the total mutual information between already selected inputs influences the selection procedure much and the redundancy is then reduced [18], [26].

## III. Neuro Fuzzy Modeling

The proposed method, mutual information based input selection, is classified as a model-independent approach. We use LLNF model in this paper due to good performance of this method in previous works. The fundamental approach with the LLNF model is dividing the input space into small linear subspaces with fuzzy validity functions. Any produced linear part with its validity function can be described as a fuzzy neuron. Thus the total model is a neuro fuzzy network with one hidden layer, and a linear neuron in the output layer which simply calculates the weighted sum of the outputs of locally linear neurons:

$$\hat{y}_i = \omega_{i_0} + \omega_{i_1} u_1 + \omega_{i_2} u_2 + \ldots + \omega_{i_p} u_p \qquad (12)$$

$$\hat{y} = \sum_{i=1}^{M} \hat{y}_i \phi_i(\boldsymbol{u}) \qquad (13)$$

where $\boldsymbol{u} = \begin{bmatrix} u_1 & u_2 & \cdots & u_p \end{bmatrix}^T$ is the model input, $M$ is the number of locally linear neurons, and $\omega_{ij}$ denotes the linear estimation parameters of the $i$-th neuron. The validity functions are chosen as normalized Gaussians:

$$\phi_i(\boldsymbol{u}) = \frac{\mu_i(\boldsymbol{u})}{\sum_{j=1}^{M}\mu_j(\boldsymbol{u})} \qquad (14)$$

$$\begin{aligned}\mu_i(\boldsymbol{u}) &= \exp\left(-\frac{1}{2}\left(\frac{(u_1 - c_{i1})^2}{\sigma_{i1}^2} + \ldots + \frac{(u_p - c_{ip})^2}{\sigma_{ip}^2}\right)\right) \\ &= \exp\left(-\frac{1}{2}\frac{(u_1 - c_{i1})^2}{\sigma_{i1}^2}\right) \times \ldots \times \exp\left(-\frac{1}{2}\frac{(u_p - c_{ip})^2}{\sigma_{ip}^2}\right)\end{aligned} \qquad (15)$$

The $Mp$ parameters of the nonlinear hidden layer are the parameters of Gaussian validity functions: center ($c_{ij}$) and standard deviation ($\sigma_{ij}$). Optimization or learning methods are used to adjust the two sets of parameters, the rule consequent parameters of the locally linear models ($\omega_{ij}$ s) and the rule premise parameters of validity functions ($c_{ij}$s and $\sigma_{ij}$ s). Global optimization of linear consequent parameters is simply obtained by least squares technique. The global parameter vector contains $M(p+1)$ elements:

$$\boldsymbol{\omega} = \begin{bmatrix} \omega_{10} & \omega_{11} & \ldots & \omega_{1p} & \omega_{20} & \omega_{21} & \ldots & \omega_{M0} & \ldots & \omega_{Mp} \end{bmatrix}^T \qquad (16)$$

and the associated regression matrix $\mathbf{X}$ for $N$ measured data samples is:

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{X}_2 & \ldots & \mathbf{X}_M \end{bmatrix} \qquad (17)$$

$$\mathbf{X}_i = \begin{bmatrix} \phi_i(\underline{u}(1)) & u_1(1)\phi_i(\underline{u}(1)) & \dots & u_p(1)\phi_i(\underline{u}(1)) \\ \phi_i(\underline{u}(2)) & u_1(2)\phi_i(\underline{u}(2)) & \dots & u_p(2)\phi_i(\underline{u}(2)) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_i(\underline{u}(N)) & u_1(N)\phi_i(\underline{u}(N)) & \dots & u_p(N)\phi_i(\underline{u}(N)) \end{bmatrix} \quad (18)$$

Therefore,

$$\hat{\boldsymbol{y}} = \mathbf{X}.\hat{\boldsymbol{\omega}} \quad ; \quad \hat{\boldsymbol{\omega}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{y} \quad (19)$$

The structure of LLNF is shown in Fig. 1. The remarkable properties of locally linear neuro fuzzy model, its transparency and intuitive construction, lead to the use of least squares technique for rule antecedent parameters and incremental learning procedures for rule consequent parameters. In this paper, Locally Linear Model Tree (LoLiMoT) algorithm as an incremental tree-based algorithm is used to tune the rule premise parameters, i.e. determining the validation hypercube for each locally linear model [6], [7]. In each iteration, the worst performing locally linear neuron is determined to be divided. All the possible divisions in the $p$ dimensional input space are checked and the best is performed. The fuzzy validity functions for the new structure are updated; their centers are the centers of the new hyper-cubes, and the standard deviations are usually set as 0.7. For more detail refer to [6].

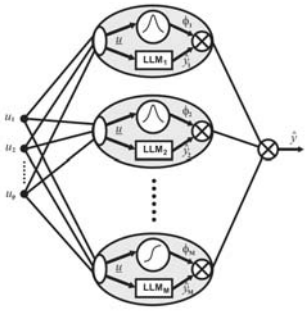

Fig. 1. Structure of locally linear neuro fuzzy model.

## IV. Case Studies

### *A. Prediction of 3rd times series of NN3 forecasting competition*

In this case study, it is tried to predict one of the NN3 forecasting competitions' time series (the 3rd time series from reduced data set) via proposed method. This time series could be downloaded from [27]. This time series includes 125 data. It is obvious that with this number of data it is difficult to train a LLNF model. Therefore in this paper bootstrap technique [28] is used to create a data set with some more samples which will be used to train the LLNF model.

First of all it needs to select proper input variables for LLNF model. The proposed input variables selection algorithm is applied to indicate the best input variables for prediction of this time series. 20 lags of the time series are considered as potential inputs. Table 1 shows the results of the proposed method compare to other methods such as correlation analysis, gamma test.

TABLE I
Order of Input Variables Selection According to The Applied Algorithm

| Input variables | Proposed algorithm | Correlation analysis | Gamma test |
|---|---|---|---|
| $x(t\text{-}1)$ | 1 | 2 | 1 |
| $x(t\text{-}2)$ | 3 | 16 | 4 |
| $x(t\text{-}3)$ | 8 | 18 | 9 |
| $x(t\text{-}4)$ | 13 | 13 | 6 |
| $x(t\text{-}5)$ | 16 | 8 | 3 |
| $x(t\text{-}6)$ | 20 | 7 | 13 |
| $x(t\text{-}7)$ | 18 | 9 | 11 |
| $x(t\text{-}8)$ | 14 | 12 | 17 |
| $x(t\text{-}9)$ | 9 | 17 | 16 |
| $x(t\text{-}10)$ | 4 | 20 | 20 |
| $x(t\text{-}11)$ | 2 | 4 | 19 |
| $x(t\text{-}12)$ | 5 | 1 | 18 |
| $x(t\text{-}13)$ | 6 | 3 | 2 |
| $x(t\text{-}14)$ | 7 | 19 | 8 |
| $x(t\text{-}15)$ | 11 | 15 | 10 |
| $x(t\text{-}16)$ | 15 | 11 | 15 |
| $x(t\text{-}17)$ | 17 | 6 | 14 |
| $x(t\text{-}18)$ | 19 | 5 | 7 |
| $x(t\text{-}19)$ | 12 | 10 | 5 |
| $x(t\text{-}20)$ | 10 | 14 | 12 |

According to the Table 1, the first five input variables are selected to create a five dimensional input vector. After creating the training and test data set, the LoLiMoT algorithm is applied to tune the parameters of the LLNF model. In this paper, Normalized Mean Square Error (NMSE) is used as error index with following definition:

$$NMSE = \left( \frac{\sum_{i=1}^{n}(y-\hat{y})^2}{\sum_{i=1}^{n}(y-\bar{y})^2} \right) \quad (20)$$

$y$, $\hat{y}$ and $\bar{y}$ are observed data, predicted data and average of observed data respectively. Note that just an average estimation of data gives a NMSE of 1. Using proposed method, the NMSE error for test data starts to increase after 4 iterations. Therefore, the optimal number of neurons is chosen to be four. Fig. 2 and Fig. 3 show the performance of the LLNF algorithm for training and test data set. It can be seen that the performance of the LLNF model with the proposed input variables selection algorithm is superior on test data set. Table 2 presents a comparison between this method and the predictions made by other methods. It is clear that proposed method, with a fewer number of neurons, achieves to a better generalization performance.

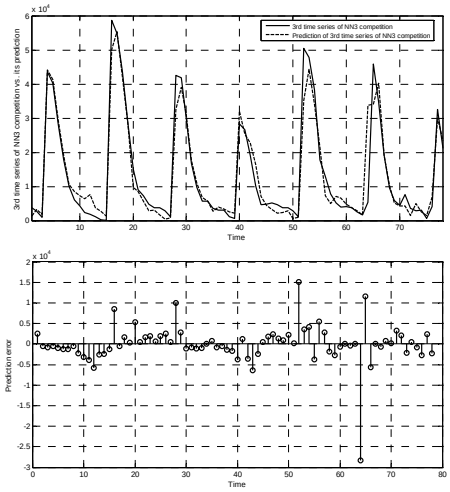


Fig. 2. Upper: 3rd time series of reduced data set from NN3 competition and its prediction with LLNF and proposed input variables selection algorithm for training set; Lower: Prediction error for training set.

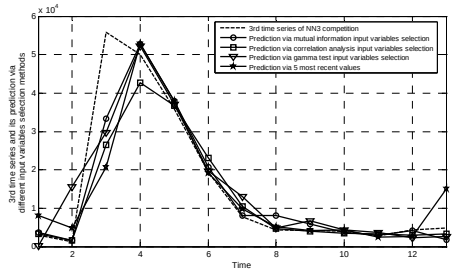


Fig. 3. 3rd time series of NN3 competition and its prediction with LLNF via different types of input variables selection algorithms for test set.

TABLE II
NMSE ERROR IN NN3 TIME SERIES PREDICTION VIA DIFFERENT INPUT VARIABLES SELECTION ALGORITHMS

| Input variables selection algorithm | Number of Neurons | NMSE error 'Train set' | NMSE error 'Test set' |
|---|---|---|---|
| *5 Most recent values* | 7 | 0.3053 | 0.3064 |
| *Correlation analysis* | 6 | 0.0105 | 0.2067 |
| *Gamma test* | 6 | 0.1995 | 0.2129 |
| *Proposed algorithm* | 4 | 0.0916 | 0.1151 |

### *B. Solar activity forecasting*

A second benchmark, prediction of sunspot numbers, will also demonstrate that good non-parametric nonlinear prediction methodologies are not the only important factors in achieving good results and good nonlinear input selection techniques are at least as important. The sunspot number is a good measure of solar activity and is computed according to the Wolf formulation:

$$R=k\,(10g+s) \tag{21}$$

Where $g$ is the number of sunspot groups, $s$ is the total number of spots in all groups and $k$ is a variable scaling factor which is related to the conditions of observation. The monthly and yearly averaged number of sunspots is accessible through several web sites from the sunspot Index Data Center in Belgium or US National Oceanic and Atmospheric Administration. To compare with the previous results the yearly sunspot number has been used in this paper, however the proposed method is capable of predicting the monthly values as well. The first 231 years, from 1700 to 1930, is used as training set and remaining data is used as test set. The proposed input variables selection algorithm is applied to indicate the best input variables for prediction of this time series. 15 lags of the time series are considered as potential inputs. Table 3 shows the results of the three input variables selection methods.

TABLE III
ORDER OF INPUT VARIABLES SELECTION ACCORDING TO THE APPLIED ALGORITHM

| Input variables | Proposed algorithm | Correlation analysis | Gamma test |
|---|---|---|---|
| $x(t\text{-}1)$ | 1 | 1 | 1 |
| $x(t\text{-}2)$ | 5 | 7 | 2 |
| $x(t\text{-}3)$ | 14 | 15 | 3 |
| $x(t\text{-}4)$ | 12 | 10 | 9 |
| $x(t\text{-}5)$ | 3 | 6 | 5 |
| $x(t\text{-}6)$ | 6 | 8 | 13 |
| $x(t\text{-}7)$ | 8 | 12 | 12 |
| $x(t\text{-}8)$ | 11 | 11 | 15 |
| $x(t\text{-}9)$ | 7 | 4 | 14 |
| $x(t\text{-}10)$ | 4 | 2 | 4 |
| $x(t\text{-}11)$ | 2 | 3 | 7 |
| $x(t\text{-}12)$ | 9 | 5 | 6 |
| $x(t\text{-}13)$ | 15 | 13 | 8 |
| $x(t\text{-}14)$ | 13 | 14 | 10 |
| $x(t\text{-}15)$ | 10 | 9 | 11 |

The NMSE error for test data starts to increase after 3 iterations. Fig. 4 and Fig. 5 show the performance of LLNF algorithm for training and test data set. It can be seen that the performance of the LLNF model with the proposed input variables selection algorithm with mutual information algorithm is acceptable according to the other well known methods in this domain. Table 4 presents a comparison between this method and the predictions made by other methods.

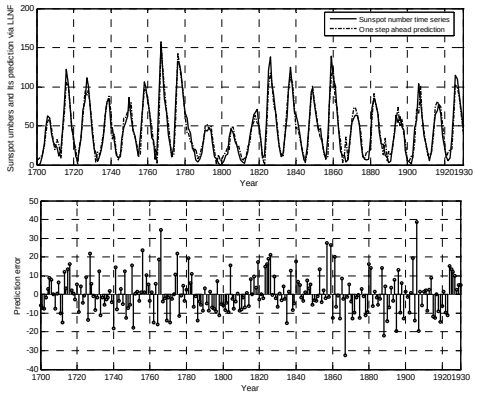


Fig. 4. Upper: Sunspot number and its prediction with LLNF and proposed input variables selection algorithm for training set; Lower: Prediction error for training set.

TABLE IV
NMSE ERROR IN SSN PREDICTION VIA DIFFERENT INPUT VARIABLES SELECTION ALGORITHM

| Input variables selection algorithm | Number of Neurons | NMSE error 'Train set' | NMSE error 'Test set' |
|---|---|---|---|
| *5 Most recent values* | 6 | 0.1095 | 0.1196 |
| *Correlation analysis* | 8 | 0.1420 | 0.2200 |
| *Gamma test* | 6 | 0.0981 | 0.1083 |
| *Proposed algorithm* | 3 | 0.1136 | 0.1159 |

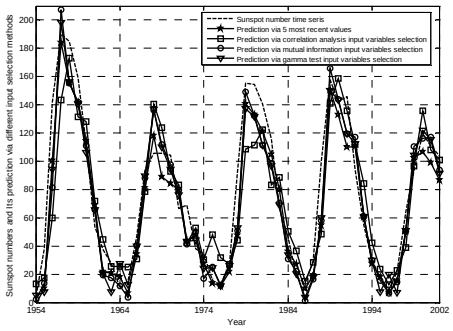


Fig. 5. Sunspot number and its prediction with LLNF via different types of input variables selection algorithm for test

## V. CONCLUSION

In this paper, information theoretic criterion is used to select a subset of input variables which have the richest information about the output to have a reliable prediction. The proposed algorithm is applied to select proper input variables for well known LLNF model to predict some time series and its results is compared with common input selection methods. Simulation results clarify the ability of mutual information to find the best subset of inputs, where relations are nonlinear and linear analysis fails to have satisfactory results. Although the concept of information theoretical approach to input selection has been discussed in several classification tasks, its importance in black box approach to nonlinear time series prediction is still not generally acknowledged. It can be argued that as more efficient nonlinear prediction techniques are progressively being developed, the importance of good nonlinear input selection routines is going to increase in future.

## REFERENCES


[1] M. Mirmomeni, C. Lucas, and E. Kamaliha, "Predicting chaotic time series using co-evolution of models and tests," *26th Int. Symp. on Forecasting,* Santander, Spain, Jun. 2006.

[2] T. Koskela, M. Varsta, J. Heikkonen, and K. Kaski, "Time series prediction using RSOM with local linear models," Research reports B15, Laboratory of Computational Engineering, Helsinki University of Technology, 1997, ISBN 951-22-3788-1.

[3] H. Yoon, K. Yang, and Cy. Shahabi, "Feature subset selection and feature ranking for multivariate time series," *IEEE Trans. on Knowledge and Data Engineering*, vol. 17, no. 9, pp. 1186-1198, Sep. 2005.

[4] I. Guyon, A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157-1182, Mar. 2003.

[5] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226-1238, Aug. 2005.

[6] O. Nelles, *Nonlinear system identification,* Springer Verlag, Berlin, 2001.

[7] A. Gholipour, C. Lucas, B. N. Araabi, M. Mirmomeni, and M. Shafiee, "Extracting the main patterns of natural time series for long-term neuro fuzzy prediction," *Neural Computing and Applications*, doi: 10.1007/s00521-006-0062-x, Aug. 2006.

[8] M. Mirmomeni, M. Shafiee, C. Lucas, and B. N. Araabi, "Introducing a new learning method for fuzzy descriptor systems with the aid of spectral analysis to forecast solar activity," *Journal of Atmospheric and Solar-Terrestrial Physics*, vol. 68, pp. 2061-2074, 2006.

[9] A. Lendasse, J. Lee , V. Wertz , and M. Verleysen, "Forecasting electricity consumption using nonlinear projection and self-organizing maps," *Neurocomputing*, vol. 48 , pp. 299–311, 2002.

[10] L. Ljung, *system identification: theory for the user*, New Jersey: Prentice-Hall, 1987.

[11] X. Hong, C. J. Harris, "Variable selection algorithm for the construction of MIMO operating point dependent neurofuzzy networks," *IEEE Trans. on Fuzzy Systems*, vol. 9, pp. 88–101, 2001.

[12] H. Leung, T. Lo, and S. Wang, "Prediction of noisy chaotic time series using an optimal radial basis function neural network," *IEEE Trans. on Neural Networks*, vol. 12, pp. 1163–1172, 2001.

[13] A. Lendasse, E. De Bodt, V. Wertz And M. Verleysen, "Non-linear financial time series forecasting – application to the Bel 20 stock market index," *European Journal of Economic and Social Systems*, vol. 14, no. 1, pp. 81-91, 2000.

[14] J. Hao, "Input selection using mutual information – applications to time series prediction," Helsinki University of Technology, M. S. thesis, Dep. of Computer Science and Engineering, Sep. 2005.

[15] N. Benoudjit, E. Cools, M. Meurens, and M. Verleysen, "Chemometric calibration of infrared spectrometers: selection and validation of variables by non-linear models," *Chemometrics and Intelligent Laboratory Systems*, vol. 70, pp. 47– 53, 2004.

[16] H. H. Yang, and J. Moody, "Feature selection based on joint mutual information," *in Advances in Intelligent Data Analysis (AIDA), Computational Intelligence Methods and Applications (CIMA), International Computer Science Conventions*, Rochester, New York, USA, 1999.

[17] A. J. Jones, "New tools in non-linear modeling and prediction," *Computational Management Science*, vol. 1, pp. 109–149, 2004.

[18] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Trans. On Neural Networks*, vol. 5, pp. 537–550, 1994.

[19] N. Reyhani, J. Hao, Y. Ji, and A. Lendasse, "Mutual information and gamma test for input selection," *in European Symposium on Artificial Neural Networks*, Bruges, Belgium, 2005.

[20] A. Al-Ani, M. Deriche, "An optimal feature selection technique using the concept of mutual information," *Int. Symposium on Signal Processing and its Applications (ISSPA)*, Kuala Lumpur, Malaysia, Aug. 2001, pp. 477-480.

[21] T. Cover, J. Thomas, *Elements of information theory*, John Wiley, 1990.

[22] C. E. Shannon, "A Mathematical theory of communication," *The Bell System Technical*, Vol. 27, pp. 379–423, 623–656, 1948.

[23] D.W. Scott, *Multivariable Density Estimation: Theory, Practice, and Visualization*, New York: John Wiley, 1992.

[24] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physics. Review*, E. 69, 066138, 2004.

[25] F. Rossi, A. Lendasse, D. François, V. Wertz, and M. Verleysen, "Mutual information for the selection of relevant variables in spectrometric nonlinear modeling," *Chemometrics and Intelligent Laboratory Systems*, vol. 80, pp. 215–226, 2006.

[26] N. Kwak, C. H. Choi, "Input feature selection for classification problems," *IEEE Trans. on Neural Networks*, vol. 13, pp.143–159, 2002.

[27] http://www.neural-forecasting- competition.com/datasets.htm.

[28] S. Theodoridis, K. Koutroumbas, *Pattern recognition*, Elsevier Academic Press, 2003.