A Genetic-Based Input Variable Selection Algorithm Using Mutual Information and Wavelet Network for Time Series Prediction

Parviz Rashidi Khazaee
Computer Engineering Department
Iran University of Science and Technology (IUST)
Tehran, Iran
p_rashidi@comp.iust.ac.ir

N. Mozayani, M.R. Jahed Motlagh
Assistant Professor, Computer Engineering Department
Iran University of Science and Technology (IUST)
Tehran, Iran
{Mozayani, jahedmr}@iust.ac.ir

***Abstract*— in this paper we presented a genetic-based optimal input selection method. This method uses mutual information as similarity measure between variables and output. Based on mutual information the proper input variables, which describe the time series dynamics properly, will be selected. The selected inputs have a maximum relevance with output variable and there exists minimum redundancy between them. This algorithm prepares proper input for Wavelet Neural Network (WNN) prediction model. The WNN prediction model utilized for time series prediction benchmark in NN3 competition and sunspot data. Presented result shows that selected input with GA outperform other input selection method like correlation analysis, gamma test and greedy alg. prediction result indicates that proper inputs have a great impact on prediction efficiency.**

Keywords— feature selection, Genetic Algorithm, time series prediction, mutual information, wavelet network

I. Introduction

One way to have a good prediction model in high dimensional problem is to overcome data dimensionality [1], [2]. To do this, one can select optimal subset of features from feature set, which describes original dataset properly. There are different filters and wrappers feature selection methods for selecting optimal features [3]. Particle swarm optimization [4] and genetic algorithm [5] are two sophisticated method for optimal feature selection. In this paper we presented a Genetic-based feature (input) selection method to select relevant variables from input data for prediction model. Relevant input variables selection is one of the most important problems in modeling and prediction tasks. The objectives of input selection algorithms are finding a subset of inputs from original input dataset, which describe system dynamics properly [4][6], [7].

In the prediction tasks proper inputs means selecting the most valuable variables from high dimensional input variables which have maximum dependency with prediction variable and have minimum redundancy. Ding and Peng [8] presented a feature selection algorithm based on minimum redundancy. Peng et al. [9] proposed max-dependency, max-relevance, and min-redundancy (mRMR) criteria and used first order feature selection algorithm for variable Selection. Yousefi et al. [2] presented greedy input selection algorithm for feature selection and used LLNF (Local Linear Neuro Fuzzy) Model for time series prediction. Vahabei et al. [10] used greedy input selection algorithm and LLNF for load forecasting.

Most of recently published papers used Mutual Information (MI) as similarity measure between variables [2], [8]-[10]. Hao focused on MI-based input selection algorithms in her thesis [11] and presented efficient MI estimation algorithms and time series prediction models. Compared to the other similarity measures like correlation coefficient, partial correlation coefficient and Linear Regression, MI is very efficient method to estimating the dependency and redundancy between variables in nonlinear problems [2], [10]-[12].

In this paper we proposed a Genetic-based input selection algorithm, which uses MI as similarity measure, and Wavelet Neural Network (WNN) Model for time series prediction. By this algorithm most valuable input variable are selected for WNN Prediction Model. WNN are very efficient tools in nonparametric estimation [13] and they have been used for time series prediction successfully [2], [10], [11]. Chen et al. [14] used Local Linear WNN for time series prediction. Billings et al. [16] presented a new class of WNN structure. The WNN prediction model in this paper is based on Billings's model. The proposed prediction model will be used for prediction of the sunspot dataset [17] and 3rd time series on NN3 Forecasting Time Series [18]. Simulation results show that the proposed model outperforms other models with fewer nodes in hidden layer.

The paper is organized as follow: Mutual Information will be described in section 2. Section 3 deals with wavelet network structure. Proposed Genetic-based input selection algorithm will be presented in section 4. Prediction results are presented in section 5. Finally some conclusion is given.

II. Proposed Algorithm

In this Section Proposed Genetic-Based input variable selection algorithm are described. This algorithm uses Mutual information as similarity measure between inputs and target variables and selects most relevant variables from inputs.

1-4244-2384-2/08/$20.00 © 2008 IEEE