Wavelet Network for Nonlinear Regression using Probabilistic Framework

Shu-Fai WONG and Kwan-Yee Kenneth WONG

Department of Computer Science and Information Systems
The University of Hong Kong, Hong Kong
{sfwong, kykwong}@csis.hku.hk

Abstract. Regression analysis is an essential tools in most research fields such as signal processing, economic forecasting etc. In this paper, an regression algorithm using probabilistic wavelet network is proposed. As in most neural network (NN) regression methods, the proposed method can model nonlinear functions. Unlike other NN approaches, the proposed method is much robust to noisy data and thus over-fitting may not occur easily. This is because the use of wavelet representation in the hidden nodes and the probabilistic inference on the value of weights such that the assumption of smooth curve can be encoded implicitly. Experimental results show that the proposed network have higher modeling and prediction power than other common NN regression methods.

1 Introduction

Regression have long been studied in statistics [1]. It receives attention from researchers because of its wide range of applications. These applications include signal processing, time series analysis and mathematical modeling etc.

Neural network is a useful tools in regression analysis [2]. Given any input signal $\{x_i, t_i\}_{i=1}^N$, neural network was found to be capable to estimate the nonlinear regression function $f(\cdot)$ such that the equation, $x_i = f(t_i) + \epsilon$, hold, where ϵ is the model noise. It outperforms many linear statistical regression approaches because it makes very few assumptions, such as linearity and normality assumptions, as other statistical approaches do [3]. It was also found that nonlinear neural network exhibits universal approximation property [4].

Researchers have proposed many neural networks for solving regression problem during past decades. For instance, multilayer network (MLP) [5], and Radial basis function networks (RBFN) [6] have been explored in previous research. Although the universal approximation property seems to be appealing, it may cause problem in regression especially when data contains heavy noise. Noisy data is quite common in applications such as financial analysis, signal processing etc.

Wavelet denoising and regression have been proposed to handle the noisy data [7]. The idea of using wavelet in noisy data regression is that the signal is firstly broken down into constituent wavelets, and those important wavelets are then chosen to reconstruct back the denoised signal. The signal without irrelevant wavelets or pulses is then used to approximate the nonlinear function

$f(\cdot)$. Such approach has been integrated with neural network to form wavelet network by some researchers such as [8]. However, the problem of over-fitting still remains because the number of hidden nodes is unknown before regression. If the number of hidden nodes is more than enough, over-fitting still occurs.

Inspired by the probabilistic framework of neural network [9] and bayesian model for wavelets [10] presented recently, probability framework for wavelet network (wavenet) is proposed in this paper. Under this framework, the weights in the wavenet will be updated according to the prior assumption of the model simplicity and smooth curve, instead of minimizing the square error as in other regression methods. Due to the above assumption, the final curve will be denoised in certain sense such that it is smooth and is also best fitted to all data points although the number of hidden nodes can be infinite initially.

2 Probabilistic Wavenet

As described in previous section, the proposed probabilistic wavenet aims at modeling the nonlinear function or the series of input data without the problem of over-fitting. In order to model the nonlinear function, simple wavelet analysis is performed to find out the constituent wavelets of the input series or signal. To be noise tolerant, wavelet denoising will be performed to remove those unimportant wavelets. By using the proposed probabilistic wavenet, these two steps can be performed automatically at once with high accuracy and in short time.

2.1 Wavelet Network

Wavelet network has been used for regression since its introduction [8]. The data set is in the form of time series data $D = \{x_i, t_i\}_{i=1}^{N}$, the wavenet will estimate the regression function $f(\cdot)$ for $X = \{x_i\}$ and $T = \{t_i\}$ by the regression model: $X = f(T) + \epsilon$.

In wavenet, the function is indeed represented by wavelet composition: $f(t_i; \omega) = \sum_{j=1}^{K} \omega_j \psi_j(t_i)$, where ω_j and ψ_j are the wavelet coefficients and the wavelet functions respectively.

To limit the number of wavelets to be used, dyadic wavelet is used. In other words, the wavelet functions are constructed by translating and scaling the mother wavelet as: $\psi_{m,n}(t) = 2^{-m/2}\psi(2^{-m}t - n)$, where m and n are the scaling and translating factor respectively. Given the signal size N, m is ranged from 1 to $log_2(N)$ and n is ranged from 0 to $2^{-m}N$.

In the proposed system, the mother wavelet function is set as the "Maxican Hat" function: $\psi(t_i) = (1 - (\frac{||t_i - \mu_\psi||}{\sigma_\psi})^2)exp(\frac{-||t_i - \mu_\psi||^2}{2{\sigma_\psi}^2})$, where μ_ψ and σ_ψ are the transition and scale factor of the mother wavelet.

Given a finite number of wavelet functions $\psi_{m,n}$, regression is done by finding optimal solution set of wavelet coefficients (ω_j). In most neural network applications, the value of such wavelet coefficients or weights are estimated by

minimizing the total error as shown:

$$\omega^* = min_\omega\{\sum_{i=1}^{N}(x_i - \sum_{j=1}^{K}(\omega_j\psi_j(t_i)))^2\} \tag{1}$$

Though it is usually possible to estimate the value of ω using common optimization methods, over-fitting may occur. To overcome such problem, probabilistic inference is proposed to estimate the value of the wavelet coefficients.

2.2 Probabilistic Framework

In order to handle the problem of over-fitting because of noisy data, probabilistic framework is adopted to estimate the value of the wavelet coefficients (or weights in wavenet) and thus estimate the original signal (or smoothened signal).

As described in previous subsection, wavenet perform regression analysis on time series data. Using the same set of notation as above, the regression model and the wavelet composition model can be combined in matrix form as
$X = \Psi\omega + \epsilon$, where Ψ is the $N \times K$ design matrix formed from the wavelet functions and ω is the weight vector.

With reference to the regression model, at any time t_i, the probability of having signal value x_i or the likelihood of the model is given by:

$$p(x_i|t_i) \sim N(f(t_i), \sigma^2) \tag{2}$$

where $\epsilon \sim N(0, \sigma^2)$ in the regression model.

To be more specific, the matrix form of the regression model can be used in expressing the probability of having data X given certain wavenet:

$$p(X|\omega, \sigma^2) = (2\pi\sigma^2)^{\frac{-N}{2}} exp\{-\frac{1}{2\sigma^2}||X - \Psi\omega||^2\} \tag{3}$$

In common neural network, the weights (ω) are found by error minimization and thus may cause over-fitting. In contrast, hyperparameter (α) is introduced in probabilistic wavenet to limit the number of wavelets with large weight value. The distribution of the value of weights is proposed as the following:

$$p(\omega|\alpha) = \Pi_{i=1}^{K} N(\omega_i|0, \alpha_i^{-1}) \tag{4}$$

From above, it is clear that the mean value of all weights is set to be zero. Thus, there is a preference to have fewer constituent wavelets. Those constituent wavelet should have large variance (α_i^{-1}). This conditional probability is served as the prior probability in estimating optimal wavelet coefficient such that the preference of fewer number of constituent wavelets is included in the estimation.

Given the weight (ω), the hyperparameter(α), the system noise (σ) and the time series data $D = (x_i, t_i)$ (or $X = \{x_i\}$), the prediction can also be done. Prediction is represented as the following through marginalization:

$$p(x^*|X) = \int\int\int p(x^*|\omega, \alpha, \sigma^2)p(\omega, \alpha, \sigma^2|X)d\omega d\alpha d\sigma^2 \tag{5}$$

Estimation of the value of weights, hyperparatmeters and system noise is indeed done by maximizing the prediction power of the regression model.

The second term can be expressed as:

$$p(\omega, \alpha, \sigma^2|X) = p(\omega|X, \alpha, \sigma^2)p(\alpha, \sigma^2|X) \tag{6}$$

From above, the posterior probability of having certain weight can be now represented as:

$$p(\omega|X, \alpha, \sigma^2) = \frac{p(X|\omega, \sigma^2)p(\omega|\alpha)}{p(X|\alpha, \sigma^2)} \tag{7}$$

Suppose the normalizing term above follows Gaussian distribution, the posterior probability can be simplified using Equation (3) and Equation (4):

$$p(\omega|X, \alpha, \sigma^2) = (2\pi)^{-\frac{N+1}{2}}|\Sigma|^{-\frac{1}{2}}exp\{-\frac{1}{2}(\omega-\mu)^T\Sigma^{-1}(\omega-\mu)\} \tag{8}$$

where $\Sigma = (\sigma^{-2}\Psi^T\Psi + A)^{-1}$, $\mu = \sigma^{-2}\Sigma\Psi^T X$ and $A = diag(\alpha_1, ..., \alpha_N)$. This gives expected value of weights, μ.

On the other hand, the second term in Equation (6) can be expressed as:

$$p(\alpha, \sigma^2|X) \propto p(X|\alpha, \sigma)p(\alpha)p(\sigma^2) \tag{9}$$

It is possible to consider the likelihood function alone to obtain the optimal values of α and σ. Here is the likelihood expression:

$$\begin{aligned} p(X|\alpha, \sigma^2) &= \int p(X|\omega, \sigma^2)p(\omega|\alpha)d\omega \\ &= (2\pi)^{-\frac{N}{2}}|\sigma^2 I + \Psi A^{-1}\Psi^T|^{-\frac{1}{2}}exp\{-\frac{1}{2}X^T(\sigma^2 I + \Psi A^{-1}\Psi^T)^{-1}X\} \end{aligned} \tag{10}$$

According to [9], the optimal value (α_{MP} and σ_{MP}) can be obtained by iteration:

$$\begin{aligned} \alpha_i^{new} &= \frac{\gamma_i}{\mu_i^2} \\ (\sigma^2)^{new} &= \frac{||X - \Psi\mu||^2}{N - \sum_i \gamma_i} \end{aligned} \tag{11}$$

where $\gamma_i = 1 - \alpha_i\Sigma_{ii}$.

After estimated the value of ω, α and σ at each stage, the prediction of the trend of the signal can be made as the following:

$$\begin{aligned} p(x^*|X, \alpha_{MP}, \sigma^2_{MP}) &= \int p(x^*|\omega, \sigma^2_{MP})p(\omega|X, \alpha_{MP}, \sigma^2_{MP})d\omega \\ &\sim N(x^*|\mu_*, \sigma_*^2) \end{aligned} \tag{12}$$

where $\mu_* = \mu^T\psi(t_{N+1})$ and $\sigma_*^2 = \sigma^2_{MP} + \psi(t_{N+1})^T\Sigma\psi(t_{N+1})$. The predicted value ($\mu_*$) and its variance ($\sigma_*^2$) is thus obtained.

In regression, the whole inference process repeats with the new values of wavelet coefficients, the new values of both hyperparameters and system noise are evaluated using Equation (8) and Equation (11) respectively until the equilibrium state is achieved. In making prediction, the predicted value and its variance can be obtained using Equation (12).

3 EXPERIMENTAL RESULT

The proposed system was implemented using Visual C++ under Microsoft Windows. Two experiments was performed to test the system. The experiments were done on a P4 2.26 GHz computer with 512M ram running Microsoft Windows.

In the first experiment, the denoising and modeling power of the proposed network was tested. A Doppler function ($f(t) = \sqrt{t(1-t)}sin(2\pi(1+a)/t+a)$) with $a = 0.05$) which contains additive noise (10dB) was used in this experiment. This function has been commonly used in research in wavelet denoising such as [10]. Noisy signal generated from such function which with signal size 1024 was analysed by the network. The result of Doppler function modeling is shown in Figure 1. The average relative square error ($((|x_{original} - x_{denoised}|^2)/(x^2_{original}))$ is 0.153. The processing time is around 4 seconds. Except the highly oscillated region in the begining of the signal (up to signal point 180), the relative square in later part is usually lower than 2. It shows that the modeling power of the proposed network is good without susceptible to noise.

In the second experiment, the prediction power of the proposed network was tested. Mackay-Glass chaotic time series were used in this experiment. The Mackay-Glass series ($x_{t+1} = (0.2x_{t-\Delta})/(1 + (x_{t-\Delta})^{10}) + 0.9x_t$ with $\Delta = 17$, $x_t = 0.9$ for $0 \leq t \leq 17$) has been used in prediction test for a long time and a comparative study in regression methods using such series can be found in [11]. In the experiment, a range of data of size 64 was analysed by the network each time and the prediction is made at the time slot 65. Prediction result is obtained by performing predictions using the appropriate range of data shifting along the input signal (with size 1024) generated from Mackey-Glass series stated above. The result is shown in Figure 2. The average relative square error is 0.317. The processing time is usually less 1 second in making each prediction. The normalized prediction error ($\varepsilon = (\sum_t (x_t - x_t^{predict})^2)/(\sum_t (x_t - x_{mean})^2)$) is 0.4777%. This result is relatively lower than the result given by other neural network approaches or linear approach as indicated in [11] (normalized error of MLP is 1.0%, those of RBF is 1.1%, those of polynomial fitting is 1.1%, those of local linear fitting is 3.3%). It shows that the prediction power of the proposed system is better than those of the other common neural network approaches.

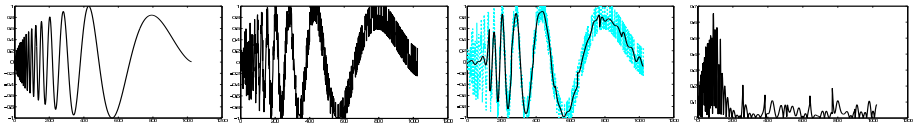
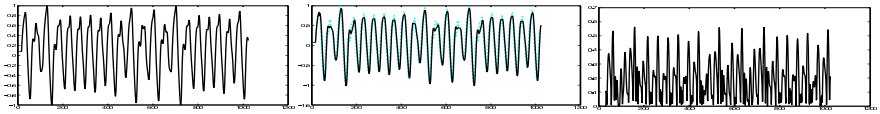


Fig. 1. *In (a), it shows the original Doppler function. In (b), it shows the Doppler function contains 10dB noise. In (c), it shows the denoised signal in black solid line compare with the noisy signal in cyan dotted line. In (d), it shows the relative square error with peak value 0.67 and average value 0.153.*



(a) (b) (c)

Fig. 2. *In (a), it shows the Mackay-Glass series. In (b), it shows the prediction curve in black solid line compare with the noisy signal in cyan dotted line. In (c), it shows the relative square error with peak with 0.56 and average value 0.317.*

4 Conclusions

In this paper, an regression algorithm using probabilistic wavelet network is proposed to perform nonlinear regression reliably such that it can be applied to real life applications such as economic forecasting. Experimental results show that the proposed network have relatively high modeling power and prediction power compare with common neural network regression methods. The proposed method can model nonlinear functions reliably without susceptible to data noise.

References

1. Fox, J.: Multiple and Generalized Nonparametric Regression. Sage Publications, Thousand Oaks CA (2000)
2. Stern, H.S.: Neural networks in applied statistics. Technometrics **38** (1996) 205–214
3. Bansal, A., Kauffmann, R., Weitz, R.: Comparing the modeling performance of regression and neural networks as data quality varies: A business value approach. Journal of Management Information Systems **10** (1993) 11–32
4. Cybenko, G.: Approximation by superposition of a sigmoidal function. Mathematics of Control, Signals and Systems **2** (1989) 303–314
5. Hornik, K., Stinchcombe, M., White, H.: Multilayer feedforward networks are universal approximators. Neural Networks **2** (1989) 359–366
6. Park, J., Sandberg, I.W.: Universal approximation using radial-basis function networks. Neural Computation **3** (1991) 246–257
7. Donoho, D., Johnstone, I.: Ideal spatial adaption by wavelet shrinkage. Biometrika **81** (1994) 425–455
8. Zhang, Q.: Using wavelet network in nonparametric estimation. IEEE Trans. Neural Networks **8** (1997) 227–236
9. MacKay, D.J.C.: Bayesian methods for backpropagation networks. In Domany, E., van Hemmen, J.L., Schulten, K., eds.: Models of Neural Networks III. Springer-Verlag, New York (1994)
10. Ray, S., Chan, A., Mallick, B.: Bayesian wavelet shrinkage in transformation based normal models. In: ICIP02. (2002) I: 876–879
11. Lillekjendlie, B., Kugiumtzis, D., Christophersen, N.: Chaotic time series - part ii: System identification and prediction. Modeling, Identification and Control **15** (1994) 225–243