Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 96 (2016) 418 – 427

Procedia Computer Science

Wavelet Neural Networks for DNA Sequence Classification Using the Genetic Algorithms and the Least Trimmed Square

Abdesselem Dakhli[a]*, Wajdi Bellil[b], Chokri Ben amar[c]

[a]Department of Computer Science, REGIM, University of Gabes 6002 Gabes, Tunisia
[b]Department of Electrical Engineering, REGIM, University of Gafsa 2110 Gafsa, Tunisia
[c]Department of Electrical Engineering, REGIM, University Sfax 3018 Sfax, Tunisia

Abstract

This paper presents a structure of the Wavelet Neural Networks used to classify the DNA sequences. The satisfying performance of the Wavelet Neural Networks (WNN) depends on an appropriate determination of the WNN structure optimization problem. In this paper we present a new method to solve this problem based on Genetic Algorithm (GA) and the Least Trimmed Square (LTS). The GA is used to solve the structure and the learning of the WNN and the LTS algorithm is applied to select the important wavelets. First, the scale of the WNN is managed by using the time-frequency locality of wavelet. Furthermore, this optimization problem can be solved efficiently by Genetic Algorithm as well as the LTS method to improve the robustness. The performance of the Wavelet Networks is investigated by detecting the simulating and the real signals in white noise. The main advantage of this method can guarantee the optimal structure of the WNN. The experimental results have indicated that the proposed method (WNN-GA) with the k-means algorithm is more precise than other methods. The proposed method has been able to optimize the wavelet neural network and classify the DNA sequences. Our goal is to construct a predictive approach that is highly accurate results. In fact, our approach allows avoiding the complex problem of form and structure in different groups of organisms. The experimental results are showed that the WNN-GA model outperformed the other models in terms of both the clustering results and the running time. In this study, we present our system which consists of three phases. The first one is the transformation, is composed of two sub steps; the binary codification of the DNA sequences and the Power Spectrum Signal Processing. The second step is the approximation; it is empowered by the use of the Multi Library Wavelet Neural Networks (MLWNN). Finally, the third one is the clustering of the DNA sequences, is realized by applying the algorithm of the k-means algorithm.

Peer-review under responsibility of KES International

Keywords: Wavelet Neural Networks, GA, DNA sequences, LTS, MLWNN, K-means ;

1. Introduction

The WNN have recently attracted extensive attention for its ability to identify effetely nonlinear dynamic systems with incomplete information [1-2-3-4-5]. The WNN were introduced by Zhang and Benveniste in 1992 as combination of the artificial neural network wavelet decomposition [1-2-3]. The generalization performance of the WNN is trained by least-square approach deteriorates when outliers are present. This training approach involves estimating parameters in the network by minimizing some function costs, a measure reflecting the approximation quality is performed by the network over the parameter space in the network.

However, the studies of the WNN have often concentrated on small dimension [6]. The reason is that the complexity of network construction will be raised exponentially with the input dimension, i.e. the curse of dimensionality to improve the performance of the WNN in high dimension application. It is a key problem that how to appropriate determinate the network

* Corresponding author. Tel.: +21628002213; fax:+ 216 75 270 686
E-mail address: abdesselemdakhli@gmail.com

Peer-review under responsibility of KES International
doi:10.1016/j.procs.2016.08.088

structure. This structure has been studied by several researchers. The research effort has been made to deal with this problem over the last decades [6-7-8-9]. The method, which is referred to as Matching Pursuit (MP), was first introduced by Mallat [10]. The Neural Network was used to solve the classification system, such as classification of the DNA sequences using the artificial neural networks [11]. Agnieszka E. et al used method to classify the genomic sequences. This method is combined a wavelet analysis and a self-organizing map algorithm [12]. It is used to extract feature of the oligonucleotide patterns of a sequence. The variation is quantified by the estimated wavelet variance, which yields a feature vector. This method is allowed the results to be visualized. When only thousands of nucleotides are available, wavelet-based feature vectors of short oligonucleotide patterns are more efficient in discrimination than frequency-based feature vectors of long patterns

The Wavelet analysis is applied to extract the features of the DNA sequences [13]. The classification of two types of DNA sequence is studied. As well as, 20 samples of the artificial DNA sequences whose their types are known, are given in order to recognize the other types of the DNA sequences. The way of wavelet denoising keeps more main formation in the original curve but it cannot predict accurately the gas zones.

In addition, the Wavelet analysis of frequency chaos game signal has been used to classify the DNA sequences. The results stemming from the complex Morlet wavelet analysis of the frequency chaos game signals have presented its accuracy in detection of variable DNA sequences structures. Moreover, this could serve in discovering unknown domains with potential biological significance in genomes [14-15-19].

This paper is organized as follow: in section II, we present an overview of the proposed method. In this section we present the transformation of the DNA sequences and the theory of wavelet neural networks. Section III deals with the simulation results of the proposed approach, which is used to classify the DNA sequences and Section IV ends up with a conclusion and a discussion.

2. Methods

This paper presents a new method of clustering of the DNA sequences based on wavelet network using the Multi Library Wavelet Neural Networks (MLWNN) to approximate f(x) of the DNA sequence. The genetic algorithm is used to solve the structure and the learning of the WNN. This approach is divided into two stages: approximation of the input signal and classification of feature extraction of the DNA sequences using the Wavelet Neural Network (WNN) and the k-means algorithm.

2.1. Conversion of the DNA sequence into a genomic signal

Our system is used the DNA sequences components to classify the species. These sequences are composed of four basic nucleotides (A, G, C and T) are called adenine, guanine, cytosine and thymine respectively, where each organism is identified by its DNA sequence [20-15-16].

The feature extraction of the DNA sequences can be viewed as finding a set of vectors which represents effectively information [17]. The method of the indicator translates the data into a digital format, which indicate the presence or absence of four nucleotides [18]. The binary indicator sequence is constructed to replace the individual nucleotides with values either 1 or 0. 0 stands for absence and 1 for presence of a particular nucleotide in specified location in DNA signal [19-20].

For example, if x[n] =[T T A T G TA ...], we obtain: x[n] =[0001 0001 1000 0001 0010 0001 1000 . ..]

2.2. Fourier Transform and Power Spectrum Signal Processing

After the DNA sequence data have been constructed into these indicator sequences, they can be manipulated with mathematical function. The Fourier Transform is used to each indicator sequence x(n) and a new sequence of complex numbers, named f (x) , is obtained:

$$f(x) = \sum_{n=0}^{N-1} X_e(n) e^{-j\pi n/N}, k = 0, 1, 2, \ldots N-1 \quad (1)$$

It is easier to work with sequence Power Spectrum, rather than original discrete Fourier Transform. The Power Spectrum Se[k] is defined as,

$$Se[k] = \left| f(x) \right|^2 \quad (2)$$

Where the frequencies k= 0, 1, 2, …, N-1.

2.3. Wavelet Neural Network and Time-frequency Analysis

The combination of the wavelet transform and the artificial neuron networks defines the concept of the wavelet networks, which applied the mother wavelet functions instead of the traditional sigmoid function as a transfer function of the each neuron. It is composed of three layers, called the input layer, the hidden layer, and the output layer. It has the same structure as the architecture radial function. The salaries of the weighted outputs are added. Each neuron is joined to the other following layer.