Constructive Neural Networks to Predict Breast Cancer Outcome by Using Gene Expression Profiles

Daniel Urda, José Luis Subirats, Leo Franco, and José Manuel Jerez

Departament of computer science,
University of Málaga. Spain
{durda,jlsubirats,lfranco,jja}@lcc.uma.es
http://www.lcc.uma.es

Abstract. Gene expression profiling strategies have attracted considerable interest from biologist due to the potential for high throughput analysis of hundreds of thousands of gene transcripts. Methods using artifical neural networks (ANNs) were developed to identify an optimal subset of predictive gene transcripts from highly dimensional microarray data. The problematic of using a stepwise forward selection ANN method is that it needs many different parameters depending on the complexity of the problem and choosing the proper neural network architecture for a given classification problem is not a trivial problem. A novel constructive neural networks algorithm (CMantec) is applied in order to predict estrogen receptor status by using data from microarrays experiments. The obtained results show that CMantec model clearly outperforms the ANN model both in process execution time as in the final prognosis accuracy. Therefore, CMantec appears as a powerful tool to identify gene signatures that predict the ER status for a given patient.

Keywords: Artificial neural networks; Constructive neural networks; Predictive modelling; Gene expression profiles; Breast cancer.

1 Introduction

Gene expression profiling strategies have been widely used in cancer studies for prediction of disease outcome [24,22] by using data from microarrays experiments. Microarrays technology is a powerful platform successfully used for the analysis of gene expression in a wide variety of experimental studies [13]. Nevertheless, this platform provides data that are in origin vulnerable to the curse of dimensionality (each sample is characterized by several thousand gene transcripts) and the curse of data set sparsity (only few examples are available for the analysis). In this sense, the estimation of prognosis models by using gene expression profiling requires the application of an additional procedure to select the most significant gene transcripts.

Methods using artifical neural networks (ANNs) were developed to identify an optimal subset of predictive gene transcripts from highly dimensional microarray data [22,14]. Recently, Lancashire et al. [7] proposed a stepwise forward

N. García-Pedrajas et al. (Eds.): IEA/AIE 2010, Part I, LNAI 6096, pp. 317–326, 2010.

selection artificial neural network approach, in which architecture size evolves sequentially by selecting and adding input neurons to the network, in order to identify an optimum gene subset based on predictive performance. This method was applied to a gene microarray dataset to identify and validate gene signatures corresponding with estrogen receptor in breast cancer.

Nevertheless, in spite of the demonstrated efficiency of ANNs in prediction and classification problems in general, the estimation of prognosis models by using stepwise forward selection procedures has, on one hand, a high computational cost, and on the other hand, a high complexity related to the design of the optimal architecture size for the neural network models. Several approaches [16,2,8,5] were proposed in this sense to solve, or alleviate, the problem of choosing the proper neural network architecture for a given problem. However, there is no general agreement on the strategy to follow in order to select an optimal neural network architecture, and the computationally inefficient "trial and error" method is still much used in applications using ANNs.

Constructive neural networks (CNN) algorithms constitute an alternative in designing artificial neural networks models [9,3,6,1,11,21,10,12,19]. This family of algorithms generate the network topology online by adding neurons during the training phase. This work presents the application of a novel constructive algorithm (CMantec) that incorporates competition and cooperation among neurons to obtain smaller in size architectures and better generalization capabilities, and analyzes its application to the problem of predicting estrogen receptor status by using data from microarrays experiments.

2 Materials and Methods

2.1 Materials

The dataset that has been used in this work was downloaded from[1] and consists of 49 samples each with 7129 corresponding variables specifying the intensity of the probe sets targeting each transcript. These data were previosly published by West et al [23]. This used microarray technology to analyse primary breast tumours in relation to estrogen receptor (ER) status. The dataset consists of 25 patients with ER+ and 24 with ER-.

2.2 Methods

For the ANN analysis, the code was written using the neural network toolbox of MATLAB and was launched on a computational cluster with 30 nodes. The CMantec algorithm was developed in Visual Studio 2008 and was launched on a personal computer. In order to compare CPU time of these algorithms, both of them were launched on the same single computer. In section 3 the reader could find time results for these algorithms.

[1] `http://data.cgt.duke.edu/west.php`