Available online at www.sciencedirect.com

SCIENCE DIRECT®

Neural Networks 18 (2005) 297–306

Neural Networks

www.elsevier.com/locate/neunet

Predictive neural networks for gene expression data analysis

Ah-Hwee Tan[a,*], Hong Pan[b]

[a]School of Computer Engineering, Nanyang Technological University, Nanyang Avenue, Singapore 639798, Singapore
[b]Genome Institute of Singapore, 60 Biopolis Street #02-01, Genome, Singapore 138672, Singapore

Received 27 July 2004; revised 17 January 2005; accepted 17 January 2005

Abstract

Gene expression data generated by DNA microarray experiments have provided a vast resource for medical diagnosis and disease understanding. Most prior work in analyzing gene expression data, however, focuses on predictive performance but not so much on deriving human understandable knowledge. This paper presents a systematic approach for learning and extracting rule-based knowledge from gene expression data. A class of predictive self-organizing networks known as Adaptive Resonance Associative Map (ARAM) is used for modelling gene expression data, whose learned knowledge can be transformed into a set of symbolic IF-THEN rules for interpretation. For dimensionality reduction, we illustrate how the system can work with a variety of feature selection methods. Benchmark experiments conducted on two gene expression data sets from acute leukemia and colon tumor patients show that the proposed system consistently produces predictive performance comparable, if not superior, to all previously published results. More importantly, very simple rules can be discovered that have extremely high diagnostic power. The proposed methodology, consisting of dimensionality reduction, predictive modelling, and rule extraction, provides a promising approach to gene expression analysis and disease understanding.

Keywords: Knowledge discovery; Gene expression analysis; Predictive modelling; Rule extraction; Feature selection

1. Introduction

Measurements of gene expression activities have provided a vast resource for medical diagnosis and disease understanding. Specifically, gene expression may provide the additional information needed to improve cancer classification and diagnosis (Slonim, Tamayo, Mesirov, Golub, & Lander, 2000). Many machine learning methods, such as Support Vector Machines (SVMs) (Furey et al., 2000), clustering (Alon et al., 1999), Self-Organizing Map (SOM), and a weighted correlation method (Golub et al., 1999), have been successfully applied to gene expression data. Although fairly high predictive performance accuracy has been obtained, most methods focus on diagnostic accuracy rather than extracting comprehensible knowledge. More recently, a method called *Emerging Patterns* has been proposed to identify gene groups characterizing specific disease classes from gene expression data (Li & Wong, 2002). To tackle the high feature dimensionality issue, a feature discretization algorithm based on entropy was used to identify the most discriminative genes before pattern discovery.

The main motivation of our work, similar to that of Li and Wong (2002), is to extract accurate as well as comprehensible knowledge from gene expression data. Specifically, we present a systematic and robust three-stage procedure for learning and extracting diagnostic knowledge from gene expression data (Fig. 1). First, feature selection is applied to the raw expression data so as to reduce the feature dimensionality to a manageable scale in accord with the number of samples available. Next, a predictive model of the gene data is learned based on the expression data in the reduced feature space. Finally, comprehensible knowledge in the form of rules are extracted from the predictive model for interpretation.

To build predictive models, we explore a class of self-organizing neural networks, known as predictive Adaptive Resonance Theory (predictive ART) networks (Carpenter, Grossberg, & Reynolds, 1991; Tan, 1995), for learning the linkages between gene expression data and diseases. Predictive ART networks are designed for fast and incremental learning of multidimensional pattern mappings. Members of predictive ART networks, such as fuzzy

[*] Corresponding author. Tel.: +65 6790 4326; fax: +65 6792 6559.
E-mail address: asahtan@ntu.edu.sg (A.-H. Tan).

0893-6080/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2005.01.003

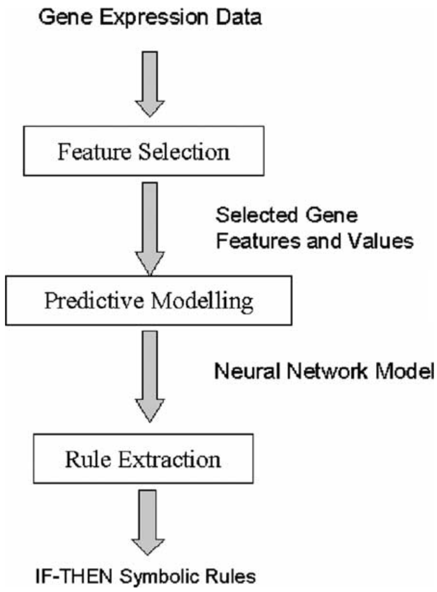

Fig. 1. The proposed methodology for gene expression analysis, consisting of feature selection, predictive modelling using neural networks, and rule extraction.

ARTMAP (Carpenter, Grossberg, Markuzon, Reynolds, & Rosen, 1992), ART-EMAP (Carpenter & Ross, 1993), and Gaussian ARTMAP (Williamson, 1996), have been successfully applied to a wide range of pattern analysis and recognition problems. However, to the best of our knowledge, there has been no attempt to date to use predictive ART networks for analyzing gene expression data.

In this paper, we adopt a simplified predictive ART architecture, known as fuzzy Adaptive Resonance Associative Map (fuzzy ARAM) (Tan, 1995), that produces classification performance equivalent to those of standard fuzzy ARTMAP. Fuzzy ARAM has been successfully applied to several machine learning tasks, including DNA promotor recognition (Tan, 1997), personal profiling (Tan & Soon, 2000), document classification (He, Tan, & Tan, 2003; Tan, 2001), and personalized content management (Tan, Ong, Pan, Ng, & Li, 2004). It has shown predictive performance comparable, if not superior, to those of many state-of-the-art learning-based systems, including C4.5 (Quinlan, 1993), Backpropagation Neural Network (Rumelhart, Hinton, & Williams, 1986), K Nearest Neighbour, and Support Vector Machines (Joachims, 1998). When performing classification tasks, fuzzy ARAM formulates recognition categories of input patterns and associates each category with a prediction. The knowledge that fuzzy ARAM discovers is compatible with IF-THEN rule-based representation. This enables the system architecture to be readily translated into a compact set of rules.

Two data sets, namely the ALL/AML data set (Golub et al., 1999) and the colon tumor data set (Alon et al., 1999), were used in our experiments. Identifying acute lymphoblastic leukemia (ALL) cases from acute myeloid leukemia (AML) cases is critical for the successful treatment of leukemia disease. Likewise, improvements in colon tumor classification have been central to advances in cancer treatment. One unique challenge of analyzing these gene expression data is the high feature dimensionality coupled with the small number of data samples. We illustrate fuzzy ARAM's predictive performance using features selected by two feature extraction methods, one statistical based (Furey et al., 2000) and the other entropy based (Fayyad & Irani, 1993). Our experiments show that fuzzy ARAM produces predictive performance comparable, if not superior, to those of all prior systems. More importantly, the rules extracted from fuzzy ARAM can be interpreted readily and used in disease understanding.

The rest of the paper is organized as follows. Section 2 presents two feature selection algorithms for reducing the dimensionality of the gene feature spaces. Section 3 presents the learning and prediction algorithms of the predictive model based on fuzzy ARAM. Section 4 illustrates how knowledge in the form of IF-THEN rules can be extracted from the predictive model. Section 5 reports our classification results and the knowledge extracted from the AML/ALL and the colon tumor data sets. The final section concludes and provides a discussion of our findings.

2. Dimensionality reduction

The first stage of our knowledge discovery process involves dimensionality reduction, in which the dimensionality of the gene expression data is reduced to a manageable number. We illustrate how predictive neural networks can work with two very distinct feature selection algorithms. The first algorithm, that computes a variant of the Fisher criterion scores, has been used by many statisticians and biologists. The Fisher method (Bishop, 1995) evaluates and selects each feature based on its own merits and preserves continuous gene expression values. The other algorithm, known as Entropy-based discretization (Fayyad & Irani, 1993), was proposed by computer scientists in the field of data mining. It employs a greedy method to select gene features, one at a time, which separate patterns into partitions with the minimum level of entropy. Both algorithms have been used in many prior experiments in extracting key features from gene expression data, including the two data sets that we investigate. Adopting the two algorithms enable us to compare the performance of the predictive neural networks in a more equal standing with those of alternative machine learning systems.

2.1. Fisher feature selection

The feature selection method based on a variant of the Fisher criterion (Furey et al., 2000; Golub et al., 1999) is summarized as follows. Consider a data set S with m expression vectors $\mathbf{x}^i=(x_1^i,\ldots,x_n^i)$, $1\le i\le m$ where m is