Information Sciences 316 (2015) 293–307

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Hidden Markov models for cancer classification using gene expression profiles

Thanh Nguyen *, Abbas Khosravi, Douglas Creighton, Saeid Nahavandi

Centre for Intelligent Systems Research (CISR), Deakin University, Waurn Ponds Campus, Victoria 3216, Australia

ARTICLE INFO

Article history:
Received 8 October 2014
Received in revised form 24 December 2014
Accepted 9 April 2015
Available online 16 April 2015

Keywords:
Analytic hierarchy process
Hidden Markov model
Cancer classification
DNA microarray
Gene expression profile
Gene selection

ABSTRACT

This paper introduces an approach to cancer classification through gene expression profiles by designing supervised learning hidden Markov models (HMMs). Gene expression of each tumor type is modelled by an HMM, which maximizes the likelihood of the data. Prominent discriminant genes are selected by a novel method based on a modification of the analytic hierarchy process (AHP). Unlike conventional AHP, the modified AHP allows to process quantitative factors that are ranking outcomes of individual gene selection methods including *t*-test, entropy, receiver operating characteristic curve, Wilcoxon test and signal to noise ratio. The modified AHP aggregates ranking results of individual gene selection methods to form stable and robust gene subsets. Experimental results demonstrate the performance dominance of the HMM approach against six comparable classifiers. Results also show that gene subsets generated by modified AHP lead to greater accuracy and stability compared to competing gene selection methods, i.e. information gain, symmetrical uncertainty, Bhattacharyya distance, and ReliefF. The modified AHP improves the classification performance not only of the HMM but also of all other classifiers. Accordingly, the proposed combination between the modified AHP and HMM is a powerful tool for cancer classification and useful as a real clinical decision support system for medical practitioners.

1. Introduction

DNA microarray is a collection of microscopic spots attached to a solid surface to measure the expression levels of genes. This technology enables researchers to study simultaneously a large number of genes (approximately 21,000 genes in the human genome). The cancer diagnosis using gene expression profiles therefore has been tremendously advanced. Cancer is basically a group of diseases when relevant genes stop functioning properly. In order to better diagnose, understand, and treat cancer, it is important to investigate which of the genes in cancer cells are working abnormally. Once subsets of differentially expressed genes are identified, classification techniques may be employed to distinguish cancer cells and normal cells.

Khan et al. [16] introduced a method for classifying cancers to specific diagnostic categories based on their gene expression signatures using artificial neural networks. The authors developed a stringent quality filter to include only the genes for which there were good measurements for all samples. Likewise, a procedure for multiclass cancer classification using

* Corresponding author. Tel.: +61 3 52278281; fax: +61 3 52271046.
E-mail address: thanh.nguyen@deakin.edu.au (T. Nguyen).

http://dx.doi.org/10.1016/j.ins.2015.04.012
0020-0255/© 2015 Elsevier Inc. All rights reserved.

multivariate partial least squares (PLS) dimension reduction combined with logistic discrimination or quadratic discriminant analysis classifiers was suggested in Nguyen and Rocke [19].

In another approach, a process that decomposes multiclass ranking statistics into class-specific statistics and uses Pareto-front analysis for gene selection was recommended in Rajapakse and Mundra [23]. You et al. [32] implemented a local dimension reduction algorithm TotalPLS based on PLS to select prominent genes for classification. Alternatively, a hybrid approach that embeds the Markov blanket with the harmony search algorithm for gene selection was suggested by Shreem et al. [25]. The procedure works well on selected genes with higher correlation coefficients based on symmetrical uncertainty.

More recently, Sun et al. [28] used kernel method to discover inherent nonlinear correlations among genes as well as between gene and target class. An iterative PLS algorithm based on backward variable elimination through the "variable influence on projection" statistic for gene selection and classification was initiated in Burguillo et al. [7]. Chen et al. [9] on the other hand utilized particle swarm optimization integrated with a decision tree to analyse gene expression data.

For evaluating a cancer classification approach, in addition to the predictive ability of gene subsets and classifiers, two other important aspects that need to be considered are the stability and computational costs. This paper introduces a hybrid method that combines a gene selector by modified analytic hierarchy process (AHP) and a classifier by hidden Markov models (HMMs). Accordingly, the contribution of this paper is twofold. First, it proposes a substantial modification to the conventional AHP to account for quantitative criteria that are statistical ranking results of five individual filter methods. Second, a supervised classifier is designed exploiting an underlying HMM that takes genes selected by AHP as inputs.

Traditional AHP often deals with qualitative factors that are derived from experts. Given that the number of genes in microarray data are at around tens of thousands and the gene knowledge available to experts is always limited, completion of assessments of various genes with respect to various criteria is not always a practical proposition. We therefore propose a substantial modification to AHP for gene selection. The modified AHP is able to quantitatively integrate statistical outcomes of individual gene ranking methods via an objective ranking procedure without consulting to possibly biased and inadequate expert knowledge. Through rigorous experiments, we show that the modified AHP yields gene subsets that lead to a classification stability at low computational cost without sacrificing the accuracy.

On the other hand, HMMs are designed following a supervised learning approach so that they are capable of realizing knowledge available from cancer training data. Cancer often develops through different stages. These stages resemble the state transition of HMMs. In addition, the modularity characteristic of HMMs allows them to be combined into larger ones where each HMM is individually trained for each cancer data class. Given a new sample, trained HMMs can predict whether it is from a cancer or normal cell. To our best knowledge, this is the first application of HMMs as a classifier for cancer classification using gene expression profiles. Through this study, we examine and compare performance of HMMs with classification methods frequently applied in literature. Experiments are conducted using four microarray datasets to make sure conclusions driven out of this study are valid and general.

Details of the HMM approach implemented as a classifier are described in Section 3. Before that, Section 2 presents a background of gene selection and the modified AHP method. Experimental results are presented and discussed in Section 4, followed by conclusions in Section 5.

2. Gene selection methods

Microarray data are commonly assembled with the number of genes much larger than the number of samples [5]. Standard techniques therefore find inappropriate or computationally infeasible in analysing such data. Not all of the thousands of genes are discriminative and needed for classification. Most genes are not relevant to the cancer development and do not affect the classification performance. Taking such genes into account enlarges the dimension of the problem, leads to computational burden, and presents unnecessary noise in the classification process. Therefore it is essential to select a small number of genes, called informative genes, which can suffice for good classification. However, the best subset of genes is usually unknown [31].

Common gene selection approaches are filter and wrapper methods. Filter methods rank all features in terms of their goodness using the relation of each single gene with the class label based on a univariate scoring metric. The top ranked genes are chosen before classification techniques are executed. In contrast, wrapper methods require the gene selection technique to combine with a classifier to evaluate classification performance of each gene subset. The optimal subset of genes is identified based on the ranking of performance derived from implementing the classifier on all found subsets. The filter procedure is unable to measure the relationship among genes whilst the wrapper approach requires a great computational expense.

In this paper, to enhance the robustness and stability of microarray data classification, we introduce a novel gene selection method based on a modification of the AHP. The idea behind this approach is to incorporate prominent discriminant genes from different gene selection ranking methods through a systematic hierarchy.

The next subsections scrutinize background of common gene selection filter methods, which are followed by our proposal. The following gene selection methods rank genes via scoring metrics, which are statistic tests based on two data samples in the binary classification problem. The sample means are denoted as μ_1 and μ_2, whereas σ_1 and σ_2 are the sample standard deviations, and n_1 and n_2 are the sample sizes.