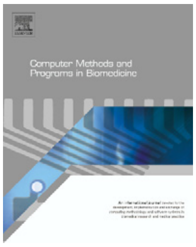
journal homepage: www.intl.elsevierhealth.com/journals/cmpb



The classification of cancer stage microarray data

Chi-Kan Chen*

Department of Applied Mathematics, National Chung Hosing University, Taiwan

ARTICLE INFO

Article history:
Received 30 March 2012
Received in revised form
20 June 2012
Accepted 17 July 2012

Keywords:
Classification
Ordinal response
Cancer stage
Microarray data

ABSTRACT

Correctly diagnosing the cancer stage is most important for selecting an appropriate cancer treatment option for a patient. Recent advances in microarray technology allow the cancer stage to be predicted using gene expression patterns. The cancer stage is in ordinal scale. In this paper, we employ strict ordinal regressions including cumulative logit model in traditional statistics with data dimensionality reduction, and distribution free approaches of large margin rank boundaries implemented by the support vector machine, as well as an ensemble ranking scheme to model the cancer stage using gene expression microarray data. Predictive genes included in models are selected by univariate feature ranking, and recursive feature elimination. We perform cross-validation experiments to assess and compare classification accuracies of ordinal and non-ordinal algorithms on five cancer stage microarray datasets. We conclude that a strict ordinal classifier trained by a validated approach can predict the cancer stage more accurately than traditional non-ordinal classifiers without considering the order of cancer stages.

1. Introduction

Microarray technology measures expression levels of thousands of genes in one experiment. Using broad gene expression patterns as signatures, subtypes of a single cancer were discovered, and predicted on test tissue samples [7]. Since then, numerous algorithms have been developed based on supervised learning techniques including linear discriminate analysis, logistic regression, artificial neural network, support vector machine, etc. to classify high-dimensional microarray data into predefined cancer classes using a relatively small number of examples (usually <100) [5,11,16,27].

Staging is a special case of cancer classification by the extent of growth and spread of a cancer. One most common cancer staging system, the TNM system, describes the cancer growth in a patient using the phenotypic variable T to represent the size of tumor, N the degree of spread to lymph node, and M the presence of metastasis. Other variables, such as the grade (G) that describes how poorly cancer cells differentiate observed under microscope, are also used by the system. The overall stage of a cancer can be defined by grouping these TNM variables. Like other health status variables, cancer stage variables can be in ordinal scale. For example, the primary tumor of a cancer can be classified into 5 stages T0, T1, . . ., T4. The stages are ordered in the sense that a lower (higher) stage indicates a smaller (larger) extent of tumor growth, and neighboring tissue invasion at diagnosis. Staging cancer is most important for determining a cancer treatment option, and predicting survival in patients. Recently, gene expression patterns have been utilized as signatures to predict cancer stage, histological tumor grade, and disease outcome [6,12,20,21,23,25].

A variety of methods for modeling ordinal response data were established using or redesigning regular classification methods [1,3,10,14,18]. Their applications in information retrieval have inspired much recent research works [13]. In this paper, we employ strict ordinal regressions including cumulative logit model in traditional statistics [14] with data dimensionality reduction, and two distribution free approaches of large margin rank boundaries implemented by the support vector machine [10,18], as well as an ensemble ranking model to model the cancer stage using gene

* Tel.: +886 4 22873181x612.
E-mail address: cchen@amath.nchu.edu.tw

0169-2607/$ – see front matter © 2012 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.cmpb.2012.07.001

expression microarray data. Predictive genes included in classifiers are selected by means of univariate feature ranking [5], and recursive feature elimination [8] to improve qualities of classifiers. We evaluate and compare prediction accuracies of ordinal classifiers and traditional non-ordinal classifiers without considering the order of cancer stage by performing external cross-validation experiments [19,26] on five publicly accessible cancer stage microarray datasets.

The paper is organized as follows. In Section 2, we describe ordinal response model, various training techniques of ordinal regression, together with feature selection methods for creating cancer stage classifiers using microarray data. Computational methods and settings are also included in the section. In Section 3, we describe performances of classifiers, both ordinal and non-ordinal, on cancer stage microarray datasets. Conclusions and comments are included in Sections 4 and 5.

2. Method

2.1. *Background*

Consider a cancer classification problem in which the covariate vector $\mathbf{x} = (x_1, \ldots, x_p)^T \in \mathbf{X} \subseteq \mathbb{R}^p$ contains the expression levels of p genes, and the response variable $y \in \mathbf{Y} = \{1, 2, \ldots, R\}$ labels the predefined cancer subgroup of tissue sample. Given a set $\{(\mathbf{x}_i, y_i)\}_{i=1}^N \subset \mathbf{X} \times \mathbf{Y}$ of N training examples, the primary goals are to classify new patient samples, and identify gene features responsible for classification using information provided by examples. Assume elements in $\mathbf{Y}$ represent ranked cancer stages such that $1 \prec 2 \prec \cdots \prec R$. The symbol "$\prec$" denotes the order relation between stages which represent "is a lower stage than". In this case, y is in ordinal scale, and can be modeled as a coarsely measured continuous latent variable $z \in \mathbb{R}$ such that the rank $y = r$ is corresponding to the interval $\theta_{r-1} < z < \theta_r$, where rank thresholds $-\infty = \theta_0 < \theta_1 < \cdots < \theta_{R-1} < \theta_R = +\infty$ divide the real line into R consecutive intervals. The decision function $z = f(\mathbf{x})$ in a predefined model space, and ordered thresholds θ_s are learned from training examples according to optimization principles. The above modeling is usually known as the (strict) ordinal regression. Let $\phi_l : \mathbf{X} \mapsto \mathbb{R}, \quad l = 1, \ldots, L$, with L probably $\gg N$, be fixed basis functions that perform the feature extraction from the gene expression vector. Assume $f(\mathbf{x})$ has the linear form $\langle \boldsymbol{\beta} \cdot \boldsymbol{\varphi}(\mathbf{x}) \rangle$, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_L)^T$ is the vector of predictor weights that maps the vector of predictive features $\boldsymbol{\varphi}(\mathbf{x}) = (\phi_1(\mathbf{x}), \ldots, \phi_L(\mathbf{x}))^T$ to a value by dot product in $\mathbb{R}^L$. We denote $\langle \boldsymbol{\varphi}(\mathbf{x}) \cdot \boldsymbol{\varphi}(\mathbf{x}') \rangle$ by the kernel function $k(\mathbf{x}, \mathbf{x}')$ for $\mathbf{x}, \mathbf{x}' \in \mathbf{X}$.

2.2. *Classifiers and solution methods*

2.2.1. *Cumulative logit model using kernel principle components*

The ordinal regression can be trained by different optimization schemes. As one most implemented method, the cumulative logit model (CLM) [14] assumes $z = f(\mathbf{x}) + \varepsilon$, where the random component ε distributes according to the standard logistic distribution. Under the assumption, the conditional probability $\Pr(y \preceq r|\mathbf{x})$ of rank no greater than r given $\mathbf{x}$ is $P_\varepsilon(\theta_r - f(\mathbf{x}))$, where $P_\varepsilon(\xi) = (1 + \exp(-\xi))^{-1}$ is the sigmoid function. The CLM, equivalent to the classic binary logistic regression when $R = 2$, is solved by the minimum of penalized negative log-likelihood of multinomial distribution

$$(\hat{\theta}_1, \ldots, \hat{\theta}_{R-1}, \hat{\boldsymbol{\beta}}) = \underset{\theta_1 < \cdots < \theta_{R-1}}{arg\ \min} \left\{ \sum_{i=1}^{N} - \ln \pi_\varepsilon(\theta_{y_i}, \langle \boldsymbol{\beta} \cdot \boldsymbol{\varphi}(\mathbf{x}_i) \rangle) + \frac{\lambda}{2} \langle \boldsymbol{\beta} \cdot \boldsymbol{\beta} \rangle \right\}, \quad (1)$$

where $\pi_\varepsilon(\theta_r, f(\mathbf{x})) = P_\varepsilon(\theta_r - f(\mathbf{x})) - P_\varepsilon(\theta_{r-1} - f(\mathbf{x})) > 0$ is the conditional probability of rank r given $\mathbf{x}$, and the coefficient $\lambda > 0$ of quadratic penalty controls the fitting accuracy. While (1) is uniquely solvable, the large L can make the numerical process computationally inefficient. To reduce the problem size, we let $\boldsymbol{\rho}(\mathbf{x}) = \mathbf{D}^{-1}\mathbf{U}^T(k(\mathbf{x}_1, \mathbf{x}), \ldots, k(\mathbf{x}_N, \mathbf{x}))^T \in \mathbb{R}^N$. The columns of $\mathbf{U} \in \mathbb{R}^{N \times N}$ are orthonormal eigenvectors associated with eigenvalues $d_1 \geq \cdots \geq d_N \geq 0$ of kernel matrix $\mathbf{K} = (k(\mathbf{x}_i, \mathbf{x}_{i'})) \in \mathbb{R}^{N \times N}$ of training data, $\mathbf{D}^{-1} \in \mathbb{R}^{N \times N}$ is a diagonal matrix containing diagonal elements $d_1^{-1}, \ldots, d_n^{-1}, \quad 0, \ldots, 0$, and the rank n of $\mathbf{K}$ can be considerably lower than L. Herein, $\boldsymbol{\rho}(\mathbf{x})$ contains principle components (PCs) of $\boldsymbol{\varphi}(\mathbf{x})$ (a.k.a. kernel principle components (KPCs) of $\mathbf{x}$) in the feature space if $\boldsymbol{\varphi}$ is centered, i.e., $\sum_{i=1}^N \boldsymbol{\varphi}(\mathbf{x}_i) = \mathbf{0}$ [17]. The 1st-order optimality condition implies that $\hat{\boldsymbol{\beta}}$ can be written as $\sum_{i=1}^N \hat{\alpha}_i \boldsymbol{\varphi}(\mathbf{x}_i)$ for some $\hat{\alpha}_i$. Substituting the form into (1), we arrive at an alternative problem

$$(\hat{\vartheta}_1, \ldots, \hat{\vartheta}_{R-1}, \hat{\boldsymbol{\delta}}_J) = \underset{\vartheta_1 < \cdots < \vartheta_{R-1}}{arg\ \min} \left\{ \sum_{i=1}^{N} - \ln \pi_\varepsilon(\vartheta_{y_i}, \langle \boldsymbol{\delta} \cdot \boldsymbol{\rho}(\mathbf{x}_i) \rangle) + \frac{\lambda}{2} \langle \boldsymbol{\delta} \cdot \boldsymbol{\delta} \rangle \right\}, \quad (2)$$

where $\boldsymbol{\delta} \in \mathbb{R}^J \times \{0\}^{N-J}$, and $J \in \{1, \ldots, n\}$. Once the solutions of (2) are obtained, $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_N)^T$ is approximated by (equal to if $J = n$) $\mathbf{U}\mathbf{D}^{-1}\hat{\boldsymbol{\delta}}_J$, and $\hat{\theta}_s$ by $\hat{\vartheta}_s$. The decision function $\hat{f}(\mathbf{x}) = \langle \hat{\boldsymbol{\beta}} \cdot \boldsymbol{\varphi}(\mathbf{x}) \rangle$ is given by the kernel function expansion $\sum_{i=1}^N \hat{\alpha}_i k(\mathbf{x}_i, \mathbf{x})$ using approximated $\hat{\alpha}_i$ as coefficients.

2.2.2. *Support vector ordinal regression (I)*

The standard two-class support vector machine (SVM) finds a predictor weight vector and a single threshold separating output values of decision function of training data from each class with largest distance between the threshold and a closest output value (margin) [22]. Similar to the way that the CLM extends the binary logistic regression model, a support vector ordinal regression (SVOR(I)) generalizes the binary SVM to find $R - 1$ ordered thresholds in the real line for R ranks using the large margin principle [4,18]. For each $s = 1, \ldots, R - 1$, let the training examples be divided into two classes and labeled as $y_i^s = -1$ if $y_i \preceq s$, and $y_i^s = +1$ if $s \prec y_i$. The mechanism of SVOR(I) is to solve

$$(\hat{\theta}_1, \ldots, \hat{\theta}_{R-1}, \hat{\boldsymbol{\beta}}, \hat{\xi}_1^1, \ldots, \hat{\xi}_N^{R-1}) = arg\ \min \left\{ \frac{1}{2} \langle \boldsymbol{\beta} \cdot \boldsymbol{\beta} \rangle + C \sum_{s=1}^{R-1} \sum_{i=1}^{N} \xi_i^s \right\}, \quad (3)$$

$$\text{subject to} \quad y_i^s(\langle \boldsymbol{\beta} \cdot \boldsymbol{\varphi}(\mathbf{x}_i) \rangle - \theta_s) \geq 1 - \xi_i^s, \ \ \xi_i^s \geq 0, \quad i = 1, \ldots, N, \quad s = 1, \ldots, R - 1,$$

where ξ_i^s is the error yielded by $\langle \boldsymbol{\beta} \cdot \boldsymbol{\varphi}(\mathbf{x}_i) \rangle$ being within the margin or on the wrong side of θ_s, and the coefficient $C > 0$ controls the total amount of training errors. This training model maximizes the smallest margin separating values of decision function of data points from a lower ($y_i^s = -1$), and an upper ($y_i^s = +1$) class with tolerated errors. Introducing the 1st-order Karush–Kuhn–Tucker (KKT) optimality conditions into