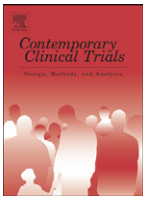
Contemporary Clinical Trials 36 (2013) 636–641

Contents lists available at ScienceDirect

Contemporary Clinical Trials

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



Overfitting in prediction models – Is it a problem only in high dimensions?

Jyothi Subramanian [a], Richard Simon [b,*]

[a] Emmes Corporation, USA
[b] Biometric Research Branch, National Cancer Institute, USA

ARTICLE INFO

Article history:
Received 25 February 2013
Revised 29 May 2013
Accepted 17 June 2013
Available online 27 June 2013

Keywords:
Classifiers
Prediction accuracy
Overfitting
Clinical trials
Patient selection

ABSTRACT

The growing recognition that human diseases are molecularly heterogeneous has stimulated interest in the development of prognostic and predictive classifiers for patient selection and stratification. In the process of classifier development, it has been repeatedly emphasized that in situations where the number of candidate predictor variables is much larger than the number of observations, the apparent (training set, resubstitution) accuracy of the classifiers can be highly optimistically biased and hence, classification accuracy should be reported based on evaluation of the classifier on a separate test set or using complete cross-validation. Such evaluation methods have however not been the norm in the case of low-dimensional, $p < n$ data that arise, for example, in clinical trials when a classifier is developed on a combination of clinico-pathological variables and a small number of genetic biomarkers selected from an understanding of the biology of the disease. We undertook simulation studies to investigate the existence and extent of the problem of overfitting with low-dimensional data. The results indicate that overfitting can be a serious problem even for low-dimensional data, especially if the relationship of outcome to the set of predictor variables is not strong. We hence encourage the adoption of either a separate test set or complete cross-validation to evaluate classifier accuracy, even when the number of candidate predictor variables is substantially smaller than the number of cases.

Published by Elsevier Inc.

1. Introduction

In many disease areas, and especially in oncology, recognition of the molecular heterogeneity of the disease has fueled the search for prognostic and predictive classifiers that identify patients who require new treatment regimens and who are likely to benefit from specific new regimens. Such classifiers can be used for selection and stratification of patients in clinical trials and for structuring the analysis plan of clinical trials. Advances in genomic technologies has moreover made it possible to measure gene expression levels for tens of thousands of genes, and these have been used in combination with traditional clinico-pathological variables to develop composite pharmacogenomic classifiers that could potentially be useful in the design and analysis of clinical trials [1]. The number of cases available for classifier development, however, remains much less, usually of the order of hundreds or less. This is commonly referred to as the high-dimensional, low sample size (HDLSS) (i.e., $p \gg n$) setting.

Overfitting, which is characterized by high accuracy for a classifier when evaluated on the training set but low accuracy when evaluated on a separate test set, has been recognized as a problem in $p \gg n$ settings [2]. In HDLSS settings, it has been repeatedly emphasized that the apparent (training set, resubstitution) accuracy of a classifier is highly optimistically biased and hence should never be reported and that accuracy should be estimated based on the evaluation of the classifier on separate test sets or through complete resampling in which the model is redeveloped for each resampling [2,3]. The use of resampling techniques or independent test sets for

[*] Corresponding author. Tel.: +1 2402766028.
E-mail address: rsimon@mail.nih.gov (R. Simon).

1551-7144/$ – see front matter. Published by Elsevier Inc.
http://dx.doi.org/10.1016/j.cct.2013.06.011

the evaluation of prediction accuracy are however not widespread in the traditional $p < n$ situations, even though overfitting is likely to be a problem in these settings also [4,5]. In the context of clinical trials, prediction problems with $p < n$ can arise, for example, when a classifier is developed on a combination of clinico-pathological and a small number of candidate genetic biomarker variables selected based on an understanding of the biology of the disease. When p is less than n, there exist rules of thumb, for example, specifying that the *effective*[1] sample size for training should be at least 10 times the number of candidate predictors [6,7]. However, these rules of thumb appear to have been developed for ensuring stability of regression coefficients [8,9] and it is not clear whether adoption of these rules also avoid overfitting.

We conducted simulation studies to investigate the existence and extent of the problem of overfitting under traditional low-dimensional settings. As p increases and starts exceeding n, traditional classification techniques like logistic regression or Fisher's linear discriminant analysis cannot be directly applied and some form of variable selection and/or shrinkage estimation becomes mandatory [4,5,10]. Shrinkage based approaches in fact are reported to be preferable in comparison to p-value based variable selection methods [5]. In our simulations, we study overfitting as a function of the ratio of p to the effective sample size, with and without feature selection. The results of these simulations and the significance of the results are reported in this paper.

2. Material and methods

2.1. Binary class prediction

In the binary class prediction problem we have a training set $\{X_i, Y_i\}$ of n observations where $Y_i \,\varepsilon\, \{0, 1\}$ is the outcome class label and $X_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ is a p-dimensional vector of predictor variables (features). The goal is to build a rule utilizing the information in X in order to predict Y. The rule is often known as a *classifier*. By developing the classifier on the training set of data, future unobserved outcomes can be predicted based on their corresponding measured predictor variables. Many methods exist for developing classifiers, including linear and quadratic discriminant analysis, logistic regression, decision trees, support vector machines, and others [11]. Additionally, variable selection may also be used in order to reduce the number of predictors in the classifier.

2.2. Simulations

For all our simulations, the number of candidate predictors, p was fixed at 10. Of the 10 predictors, 5 predictors were informative and the remaining 5 were non-informative. The number of samples in the training set, n, was varied from 20 to 1000. Half of the samples (i.e. $n/2$) were randomly assigned to class 0 ($Y = 0$) and the other half to class 1 ($Y = 1$). The effective sample size in our simulations was thus $n/2$. The informative predictors were generated from $N(0, I_5)$ for class 0 and $N(\mu, I_5)$ for class 1. The non-informative predictors were generated from $N(0, I_5)$ for both class 0 and class 1. Separate simulations were carried out for the values of μ in 0, 0.25 and 0.5 to represent the null signal and signals of increasing strength from moderate to high. Additionally, a simulation was conducted with two informative predictors with $\mu = 0.25$ and three informative predictors with $\mu = 0.5$.

To study the sensitivity of results to correlation among predictors, additional simulations were carried out with block diagonal correlation structures, where the informative and non-informative predictors were assumed to be correlated with pairwise common correlation coefficient r. Values of $r = 0.25$ and 0.75 were used.

Diagonal linear discriminant analysis (DLDA) was used as the classification method [10]. DLDA corresponds to Fisher's linear discriminant analysis where the class specific densities are assumed to have the same diagonal covariance matrix. In DLDA, a new sample with feature vector $X^* = (x^*_1, x^*_2, \ldots, x^*_p)$ is assigned to class 0 if

$$\sum_{j=1}^{p} \frac{\left(x_j^* - \overline{x}_j^{(0)}\right)^2}{s_j^2} \le \sum_{j=1}^{p} \frac{\left(x_j^* - \overline{x}_j^{(1)}\right)^2}{s_j^2}$$

and otherwise assigned to class 1. DLDA has the advantage that for p predictors, only p variances need to be estimated. In contrast to this, Fisher's linear discriminant analysis requires the estimation of $p(p+1)/2$ elements of the covariance matrix. DLDA is commonly used in $p > n$ settings as it is more robust to overfitting compared to Fisher's LDA and often results in greater predictive accuracy even when the features are correlated [11]. For $p > n$ problems, ordinary logistic regression too cannot be used because the design matrix is singular. Stepwise logistic regression tends to provide substantially overfit models in that setting and so penalized version of logistic regression are often used to shrink the regression coefficients.

DLDA was used in our simulations because of its stability in $p < n$ problems and its resistance to overfitting compared to Fisher's LDA and stepwise logistic regression in $p > n$ problems. When the class specific covariance matrices are equal and diagonal, DLDA is equivalent to logistic regression.

Since, typically, some form of variable selection is incorporated even in the low-dimensional case; simulations were conducted with and without variable selection to study the impact of variable selection on overfitting. The variable selection methods studied were:

(i) selecting variables with the largest k absolute value univariate t-test statistics with $k = 3$ or 5.

(ii) using cross-validation to select the optimal number of variables in the model.

[1] For the linear regression problem, the effective sample size is the actual sample size. In the case of proportional hazards regression, the effective sample size is the number of events, and in case of binary class prediction, the effective sample size is the number of observations in the smaller of the two classes [4].

Download English Version:

https://daneshyari.com/en/article/6151315

Download Persian Version:

https://daneshyari.com/article/6151315

Daneshyari.com