Journal of Biomedical Informatics 53 (2015) 355–362

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Practical approach to determine sample size for building logistic prediction models using high-throughput data

Dae-Soon Son [a,f,g,1], DongHyuk Lee [b,1], Kyusang Lee [c], Sin-Ho Jung [d], Taejin Ahn [a,f], Eunjin Lee [a,f], Insuk Sohn [e], Jongsuk Chung [a,f], Woongyang Park [a], Nam Huh [f,*], Jae Won Lee [g,*]

[a] Samsung Genome Institute, Samsung Medical Center, Seoul, Republic of Korea
[b] Department of Statistics, Texas A&M University, College Station, TX 77843, USA
[c] Clinomics, Inc., A-616 Gardenfive Works, Seoul, Republic of Korea
[d] Department of Biostatistics and Bioinformatics, Duke University, NC 27710, USA
[e] Samsung Cancer Research Institute, Samsung Medical Center, Seoul, Republic of Korea
[f] In vitro Diagnostics Research Lab, Bio Research Center, Samsung Advanced Institute of Technology, Gyeonggi-do, Republic of Korea
[g] Department of Statistics, Korea University, Seoul, Republic of Korea

ARTICLE INFO

Article history:
Received 11 July 2014
Accepted 18 December 2014
Available online 30 December 2014

Keywords:
Sample size
Statistical power
Prediction and validation
Permutation
Null distribution

ABSTRACT

An empirical method of sample size determination for building prediction models was proposed recently. Permutation method which is used in this procedure is a commonly used method to address the problem of overfitting during cross-validation while evaluating the performance of prediction models constructed from microarray data. But major drawback of such methods which include bootstrapping and full permutations is prohibitively high cost of computation required for calculating the sample size.

In this paper, we propose that a single representative null distribution can be used instead of a full permutation by using both simulated and real data sets. During simulation, we have used a dataset with zero effect size and confirmed that the empirical type I error approaches to 0.05. Hence this method can be confidently applied to reduce overfitting problem during cross-validation. We have observed that pilot data set generated by random sampling from real data could be successfully used for sample size determination. We present our results using an experiment that was repeated for 300 times while producing results comparable to that of full permutation method. Since we eliminate full permutation, sample size estimation time is not a function of pilot data size. In our experiment we have observed that this process takes around 30 min.

With the increasing number of clinical studies, developing efficient sample size determination methods for building prediction models is critical. But empirical methods using bootstrap and permutation usually involve high computing costs. In this study, we propose a method that can reduce required computing time drastically by using representative null distribution of permutations. We use data from pilot experiments to apply this method for designing clinical studies efficiently for high throughput data.

1. Introduction

One of the main reasons to utilize high dimensional data from microarrays in clinical research is to develop statistical models that predict clinical outcomes such as, time to recurrence, progression of disease and survival of patients. Finding high quality samples is costly and difficult but it constitutes a key task in performing clinical studies. The task of finding minimum number of samples for scientific study is very important to minimize the wastage of valuable resources and retain clinical utility of the experiment. Determination of sample size based on sound technical basis is a significant part of guidelines set by Institutional Review Board (IRB).

Several important methodologies were proposed to determine sample size for microarray experiments. Liu and Hwang report a formula suitable for comparison studies with multiple independent samples [1]. Methods which introduce the concept of controlling False Discovery Rate (FDR) in microarray analysis were further developed to estimate power and sample size [2–5]. These methods are aimed at discovering statistically valid

* Corresponding authors.
E-mail addresses: ds3.son@samsung.com (D.-S. Son), dhyuklee@tamu.edu (D. Lee), klee@clinomics.co.kr (K. Lee), sinho.jung@duke.edu (S.-H. Jung), taejin.ahn@samsung.com (T. Ahn), eunjin.lee@samsung.com (E. Lee), insuk.sohn@samsung.com (I. Sohn), doogie.chung@samsung.com (J. Chung), woongyang.park@samsung.com (W. Park), bio.stat@daum.net (N. Huh), jael@korea.ac.kr (J.W. Lee).
[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.jbi.2014.12.010
1532-0464/

biomarkers. However, because of the inherent complexities in the genetic make-up of diseases such as cancer, diabetes and other immune diseases these methods suffer from less-than-desirable accuracy for medical practice. This implies that one has to consider the necessity of multiple parameters in the prediction models, as well as the variations from experimental platform. Recent FDA clearance of Affymetrix™ system as a diagnostic platform presents an example of rapid upgrade in reliability of such platforms for use in clinical settings [6]. Statistical prediction models such as one used in OncotypeDX™ which employs multiple biomarkers validate the value of such predictive models. These models have formed a trend in many clinical trials in combination with co-diagnosis approach [7,8].

Recently, Pang and Jung provided an idea to rigorously determine the sample size required to construct such a predictive model [9]. It estimates empirical power of a suggested sample size using simulated data from bootstrapping based on a predictive model developed using pilot project data. A proof was given by Jung and Young [10] that demonstrates the structure of covariance from pilot data and bootstrapped data from the pilot data are approximately identical. They also suggest a method to estimate empirical power when the response variable is of survival type. Since this method constructs individual prediction models from numerous simulated data sets and performs cross-validation and permutation each time, it reduces the problem of over-fitting while adding expensive computation time for repeated calculations.

The concept of prediction–validation method is the first of its kind to determine the sample size of multi-dimensional data [9]. However, it remains a concern that it requires lot of time to determine proper sample size of a data set with many variables. This approach would be more practical if the computation complexity could be reduced. We were inspired by an observation that a set of simulated data sets from pilot data seem to generate similar non-centrality parameters when each set was estimated by maximum likelihood method. Thus, it seemed reasonable to assume that a carefully selected single null distribution could be re-used in other sets for adjusting p-values. Our current study provides a method to determine sample size for the case of binary response variables using this idea. We demonstrate empirical evidence by extensive simulation which supports the fact that sample size can be conveniently approximated.

2. Methods

It is known that statistical power can be estimated from a number of simulated data by bootstrapping based on a prediction model from pilot data. A prediction model is constructed for each simulated data. Validation of the models can estimate the empirical power from the ratio of valid models over the total set of simulated data [9]. We carry out a χ^2-test for each model from a simulated data and regard the model to be valid when the p-value is less than the significance level of 0.05. This procedure heavily depends on repeated CV with permutation on simulated data. Consequently, it results in immense computational cycles which prohibit practical applications of this method.

We have created a representative permutation null distribution from one randomly chosen simulated data among those of showing the highest marginal frequency within the group of whole simulated data. Details of each step to determine this null distribution is explained in the following section.

2.1. Bootstrapping data generation

In order to estimate empirical power, many simulated data sets are required. A small-sized pilot data is defined as $\mathcal{M} = \{w_i, (x_{i1}, \ldots, x_{ig}), i = 1, \ldots, m\}$, where w_i is a response variable for ith subject and x_{ig} is gth gene expression level for ith subject. That is, there are m individuals and the number of features is g. The sample mean and the standard deviation are denoted by $\bar{x}_j, s_j$ respectively for feature (or genes in our example) $j(= 1, \ldots, g)$. Let $\widetilde{\mathcal{M}} = \{y_i, (z_{i1}, \ldots, z_{ig}), i = 1, \ldots, N\}$ be a bootstrapped simulated data, where y_i is a response variable in the bootstrap sample, $z_{ij} = (x_{i'j} - \bar{x}_j)\varepsilon_i/s_j$ for random variables of $\varepsilon_1, \ldots, \varepsilon_N \sim i.i.d.N(0, 1)$, and i' is a randomly chosen number from $(1, \ldots, m)$. Note that bootstrapped sample size N can differ from pilot sample size m $(N > m)$. We repeat this process to generate many simulated data sets to use when estimating empirical power. It is known that $Cov(\widetilde{\mathcal{M}}|\mathcal{M}) \to Cov(\mathcal{M}),$ as $n \to \infty$ [10].

In order to construct a multiple regression prediction model, candidate markers are selected among thousands of genes. It is done by univariate logistic regression applied to the pilot data, and it selects t candidate markers. The ID of selected genes are represented as $(\tilde{1}, \ldots, \tilde{t})$, and their expression values, $Z_i = (Z_{i\tilde{1}}, \ldots, Z_{i\tilde{t}})$ for the individual i respectively.

Risk score of an individual $i(= 1, \ldots, N)$ can be represented as $\hat{p}_i = P(y_i = 1 \mid z) = 1/(1 + \exp\{-\beta^T Z_i\})$, the probability estimated from multiple logistic regressions. Therefore, Bernoulli trial of probability $\hat{p}_i$ allows evaluation of binary response variable, $y_i(i = 1, \ldots, N)$ which corresponds to the simulated data generated from bootstrapping. By repeating the procedure one can generate a data set of bootstrap microarray data, $\widetilde{\mathcal{M}}_b = \{y_i, (z_{i1}, \ldots, z_{ig})\}$, where $i = (1, \ldots, N), b = (1, \ldots, B)$, each of which are of sample size N.

2.2. Methods of prediction and evaluation

Since the statistical power is the conditional probability of rejecting the null hypothesis when it is really false, we can estimate by calculating the proportion of rejection, that is, the p-values less than the level of significance. Thus we need to clarify what p-value is in this situation. Here we defined p-value for the specific model based on the comparison between predictive values and original values.

Prediction: We used multiple logistic regression and 3-fold CV with permutation method to predict $\hat{y}_i$, the predicted response variables of y_i in $\widetilde{\mathcal{M}}_b$. The construction of prediction models for each simulated data sets, $\widetilde{\mathcal{M}}_b$ starts with selecting top t predictors through univariate logistic regression in each data set as explained in detail below. We tried to reduce the concern of overfitting of cross-validation procedure by using permutation method. Thus the vector of predicted values $\hat{\boldsymbol{y}} = (\hat{y}_1, \ldots, \hat{y}_N)$ is determined:

- Divide the dataset into K (nearly) equal-sized subsets and for fixed $k(k = 1, \ldots, K)$, remove kth subset.
- Perform a univariate logistic regression analysis on each of the genes using the remaining $(K - 1)$ subsets and find top t predictors.
- Build a multiple logistic regression model with top t predictors and find the predicted values using the remaining kth subset. Those predicted values $\hat{y}_i$, where i is the index of kth subset composes the predicted vector $\hat{\boldsymbol{y}}$.

Evaluation: The similarity of $\hat{y}_i$ and y_i can be calculated by homogeneity test based on χ^2-statistic in bth bootstrapped data. The null hypothesis is, $H_0 : P(\hat{y}_i|y_i = 0) = P(\hat{y}_i|y_i = 1)$, and alternative hypothesis is $H_1 : P(\hat{y}_i|y_i = 0) \neq P(\hat{y}_i|y_i = 1)$. Thus the performance of prediction model can be evaluated by the p-value of homogeneity test for 2×2 contingency table:

- Calculate the homogeneity chi-squared statistic (χ^2_b) of bth bootstrapping data, $\widetilde{\mathcal{M}}_b$ using two vectors $\hat{\boldsymbol{y}} = (\hat{y}_1, \ldots, \hat{y}_N)$ and $\boldsymbol{y} = (y_1, \ldots, y_N)$.