Gene expression

An ensemble approach to microarray data-based gene prioritization after missing value imputation

Dong Hua¹ and Yinglei Lai²,∗

¹Department of Computer Science, The George Washington University, 801 22nd Street, Suite 704 and ²Department of Statistics and Biostatistics Center, The George Washington University, 2140 Pennsylvania Avenue, N.W. Washington, DC 20052, USA

Received on August 31, 2006; revised on December 26, 2006; accepted on January 14, 2007

ABSTRACT

Motivation: Microarrays have been widely used to discover novel disease related genes. Some types of microarray, such as cDNA arrays, usually contain a considerable portion of missing values. When missing value imputation and gene prioritization are sequentially conducted, it is necessary to consider the distribution space of prioritization scores due to the existence of missing values. We propose an ensemble approach to address this issue. A bootstrap procedure enables us to generate a resample multivariate distribution of the prioritization scores and then to obtain the expected prioritization scores.

Results: We used a published microarray two-sample data set to illustrate our approach. We focused on the following issues after missing value imputation: (i) concordance of gene prioritization and (ii) control of true and false positives. We compared our approach with the traditional non-ensemble approach to missing value imputation. We also evaluated the performance of non-imputation approach when the theoretical test distribution was available. The results showed that the ensemble imputation approach provided clearly improved performances in the concordance of gene prioritization and the control of true/false positives, especially when sample sizes were about 5–10 per group and missing rates were about 10–20%, which was a common situation for cDNA microarray studies.

Availability: The Matlab codes are freely available at http://home.gwu.edu/~ylai/research/Missing.

Contact: ylai@gwu.edu

1 INTRODUCTION

Microarrays enable us to simultaneously monitor gene expressions at a genomic scale (Der et al., 1998). They provide the tool to extract biological significances such as the changes in expression profiling of genes under distinct types (e.g. normal versus cancer types), which shed the light on use of them in a number of studies over a broad range of biological disciplines including cancer classification (Golub et al., 1999), identification of the unknown effects of a specific therapy (Perou et al., 2000), identification of genes relevant to a certain diagnosis or therapy (Cho et al., 2003) and cancer prognosis (Shipp et al., 2002; van’t Veer et al., 2002). Due to their relatively high costs, the sample sizes of microarray studies are generally small, which may lead to considerable false positive rates. Since microarrays are widely used in pilot studies before the follow-up large sample validation studies, it is crucial to control the false positives in genes prioritized by microarray studies.

Some types of microarray, such as cDNA arrays, usually contain a considerable portion of missing values. These missing values exist due to various reasons including insufficient resolution, image corruption, dust or scratches on the slides or experimental error during the laboratory process (Kim et al., 2005). Effective imputation methods, which intend to recover these missing values, are important: on one hand, it is costly to repeat the experiments; and on the other hand, the repeat of experiments cannot guarantee data completeness.

Before data analysis, missing value imputation is generally required. Many algorithms for the gene expression data analysis, like support vector machines (Vapnik, 1995), and multivariate statistical analysis methods such as principal component analysis (Golub and van Loan., 1996), singular value decomposition (Alter et al., 2000) and generalized singular value decomposition (Alter et al., 2003), require a complete data set as the input. It can also be the case when microarrays are used in pilot studies for gene prioritization. Scores from a certain statistical test are generally used to rank genes. The sample sizes of different genes must be uniform so that the test scores of different genes can be comparable. One may consider using the corresponding P-values to rank genes, in which the sample sizes of different genes can be different. This approach is feasible when we know the theoretical test distribution. If such a distribution is unknown, which is usually the case in practice, we have to consider the permutation method for evaluating P-values. However, it is generally difficult to accurately evaluate the P-values for genes with missing observations when their sample sizes are small.

Recently, many methods have been developed for missing value imputations. These include a SVD-based method and a weighted k-nearest neighbors imputation (Troyanskaya et al., 2001), Bayesian approaches (Oba et al., 2003; Zhou et al., 2003), a fixed rank approximation algorithm (FRAA) (Friedland et al., 2003), a least squares method (Bo et al., 2004), a local least squares imputation (Kim et al., 2005),
a collateral imputation method (Seagal et al., 2005) and a SVM and orthogonal coding scheme based method (Wang et al., 2006). Furthermore, Kim et al. (2004) proposed to reuse the imputed data to improve missing value estimation, Tuikkala et al. (2006) proposed to consider gene ontology information for improving missing value estimation, and Gan et al. (2006) proposed to consider a set theoretic framework and biological knowledge to improve missing value estimation.

The impact of missing value imputation on differentially expressed gene identification has also been recently studied (Jornsten et al., 2005; Scheel et al., 2005). For multi-sample microarray data, gene prioritization is equivalent to detecting differentially expressed genes. When missing value imputation and gene prioritization are sequentially conducted, it is necessary to consider the distribution space of prioritization scores due to the existence of missing values. However, this issue has not been addressed since all the aforementioned missing value imputation methods only provide one estimate for each missing observation.

Ensemble methods, such as boosting (Freund and Schapire, 1997) and random forest (Breiman, 2001), have been widely used in the field of machine learning. When the number of predictor variables is relatively large, these methods can usually achieve satisfactory classification performance through combining a group of weak classifiers. In this study, we propose an ensemble approach to address the issue of microarray data matrix with missing value imputation after missing value imputation. We first describe some preliminaries. Then, we detail a bootstrapped procedure. A two-sample microarray data set is used to illustrate our approach.

2 METHODS

2.1 Preliminaries

Throughout the article, we will use $G \in \mathbb{R}^{m \times n}$ to represent a gene expression data matrix with $m$ genes (rows) and $n$ experiments (columns). $G$ may consist of the observed component $X$ and the missing component $Y$:

$$G = \{g_{ij} : i = 1, \ldots, m; j = 1, \ldots, n\} = X \oplus Y.$$ 

We will use $\hat{X}$ to denote the estimation of $X$ by an imputation method $f$:

$$\hat{X} = f(X \oplus Y).$$

The existing imputation methods designed for microarray missing value estimation focus on the once-for-all estimation of the missing values based on the observed gene expression values. The subsequent gene prioritization and other analyses are entirely separated from the missing value imputation. Typically, a complete matrix $G^*$ is obtained by replacing the missing component $Y$ with the estimated $\hat{Y}$. Then, the complete data matrix $G^*$ is used for the subsequent analyses. The subsequent analyses are no more relevant to the previous imputation process after $G^*$ is constructed. We will use $C$ to denote the operation of the subsequent analyses. The traditional way to the whole process can be represented by $C(X \oplus \hat{Y})$, which is a non-ensemble approach.

In this article, we introduce a novel ensemble approach, which is capable of incorporating any imputation method for missing value estimation, where the target outcome for the missing component $Y$ is featured to be random as $Y_a$ and to follow a multivariate distribution $\Omega_a$. We walk through $\Omega_a$, i.e. the instantiation of $Y_a$, through a bootstrap procedure. In this way, we can obtain an estimate for the operation $C$ through an ensemble over $C(X \oplus Y_a)$.

The whole process can be represented by $\tilde{C}(X \oplus Y_a)$. Generally, $C(X \oplus \hat{Y}) \neq \tilde{C}(X \oplus Y_a)$.

2.2 Bootstrapping imputation for individual samples

The bootstrap method, first proposed by Efron (1979), enables us to generate a resample distribution of estimates in a non-parametric manner. Since different samples are generally unrelated, we perform a bootstrap procedure for each column (sample) to generate a resample distribution of missing value estimates.

Without loss of generality, we describe the bootstrap procedure for the first column $g_1$. First, we sample $n-1$ number $(1,2,\ldots,n)$ from $(2,3,\ldots,n)$ with replacement; Then, we use these resample $n-1$ columns $(g_2,g_3,\ldots,g_n)$ to impute the missing values in the column $g_1$, (see Section 2.5 for the description of imputation method).

The above two steps are repeated $b$ times and we obtain $b$ resample columns $(g_{11},g_{12},\ldots,g_{1b})$. Notice that those non-missing values in the original column are not changed in these resample columns.

After performing the above procedure for all individual columns, we obtain $b \times n$ resample vectors $(g_1^{a1},g_1^{a2},\ldots,g_1^{ab})$ for each sample (column). $b$ will be referred to as the size of bootstrap in the rest of the article.

2.3 Constructing resample complete matrices

For each individual column $j, j = 1, 2, \ldots, n$, we randomly select a resample replicate $g_j^*$ from these $b$ resample replicates generated by the above procedure. In this way, we construct a resample matrix $(g_1^{a1},g_1^{a2},\ldots,g_1^{ab})$ (there are $b^n$ possible combinations.) We obtain $r$ resample matrices by repeating the above step $r$ times. (Notice that those non-missing values in the original matrix are not changed in these resample matrices.) $r$ will be referred to as the size of ensemble in the rest of the article.

2.4 Averaging resample prioritization vectors

Using the method for gene prioritization (see Section 2.6 for detail), we can obtain a resample score vector for each resample matrix. We consider the average of these $r$ resample vectors as the estimate of prioritization vector $C(X \oplus Y_a)$. When the size of ensemble $r$ is large, we consider averaging by mean. However, this is usually time-consuming. Generally, we set $r=100$ and consider averaging by median, which is less sensitive to outliers and is especially useful when $P$-values are used as prioritization scores. Notice that if there is no missing value in the original data, then this average vector will be the same as the prioritization vector calculated based on the original data.

2.5 Imputing missing values

The LLSimpute/L2 method proposed by Kim et al. (2005) is used for missing value imputation. We briefly describe it as follows. Without loss of generality, we consider the first gene and calculate the $L_2$-norm based similarity measures between this gene and the rest $m-1$ genes.
2.6 Prioritizing genes

For a two-sample data set, we first consider the Student’s t-test for prioritizing genes. We use $n_1$ and $n_2$ to denote the sample sizes of the first and the second groups, respectively, $n_1 + n_2 = n$. For each gene, we use \{x_{11}, x_{12}, \ldots, x_{1m}\} and \{x_{21}, x_{22}, \ldots, x_{2m}\} to denote its measurements (observed for non-missing or imputed for missing) in the first and the second groups, respectively. The Student’s t-test is given by $t = \Delta / s$, where $\Delta = (\bar{x}_1 - \bar{x}_2)$ and $s = \sqrt{s_0^2(1/n_1 + 1/n_2)}$; $\bar{x}_1 = \sum_{i=1}^{n_1} x_{1i}/n_1$, $\bar{x}_2 = \sum_{i=1}^{n_2} x_{2i}/n_2$, and $s_0 = \sqrt{\sum_{i=1}^{n_1} (x_{1i} - \bar{x}_1)^2 + \sum_{i=1}^{n_2} (x_{2i} - \bar{x}_2)^2} / (n_1 + n_2 - 2)$.

The above test is performed for all $m$ genes. For the data analyzed in this study, we observed through the Quantile-Quantile (Q-Q) plot that the theoretical t-distribution was consistent with the permutation distribution. Therefore, $P$-values from their corresponding t-distributions are used as the score vector for gene prioritization. This allows us to evaluate the performance of the non-imputation approach, which calculates the t-test only based on the observed data.

We also consider the SAM t-test (Tusher et al., 2001) for prioritizing genes. This test adds a fudge factor to the denominator of the Students t-test: $t = \Delta / (s + s_0)$. $s_0$ is numerically determined according to the given data. This test has been widely used in two-sample microarray data analyses since it can generally improve the control of false positives by excluding genes with relatively small variances. However, the theoretical distribution of this test is unknown, and we have to use the permutation method to evaluate the $P$-values of tests. Since it is difficult to evaluate the $P$-values for genes with missing observations (missing values are actually not allowed in the implemented R function sam), the performance of the non-imputation approach cannot be evaluated.

Remark: Although test scores and their corresponding $P$-values have equivalent effects in prioritizing genes when genes have uniform sample sizes, we recommend to use $P$-values since they can provide additional significance information of the tests.
3 RESULTS

We use the two-sample ZAP-70 dataset (Wiestner et al., 2003), which is publicly available at http://llmpp.nih.gov/cll/, to evaluate our ensemble imputation approach against the non-ensemble imputation approach as well as the non-imputation approach. We focus on the following two questions with the consideration of various sample sizes and missing rates:

- Will the ensemble imputation approach provide better concordant prioritization of genes?
- Will the ensemble imputation approach provide improved control of true and false positives?

3.1 ZAP-70 Dataset

There are 12447 genes and 107 cases involved in this two-sample data set, which was collected for the study of identification of a chronic lymphocytic leukemia subtype with unmutated immunoglobulin (Ig) genes, inferior clinical outcome and distinct gene expression profile. The sample sizes of the Ig-mutated and Ig-unmutated group are 79 (n1) and 28 (n2), respectively. The overall missing rate is 12.1%. There are 2508 genes with no missing value. We denote this subset as \( G \in \mathbb{R}^{2508 \times 107} \) and use it for our evaluation study with different sample sizes \( \{5, 10, 10\} \) and \( 15, 15\) and missing rates \( \{5, 10, 20\}\%\).

Without loss of generality, we briefly describe the procedure for data matrix generation of the sample size \(5+5\) coupled with \(5\%\) missing rate. First, we randomly choose five columns (samples) from each group to form a new complete data matrix \( G_c \in \mathbb{R}^{2508 \times 10} \). Then, we randomly knock out entries as missing with \(5\%\) probability. The newly constructed incomplete matrix \( G_m \in \mathbb{R}^{2508 \times 10} \), which contains missing values, is used for imputation. Based on these matrices, we consider both the ensemble and the non-ensemble approaches to gene prioritization after missing value imputation. We use \( v_c = (v_{c1}, v_{c2}, \ldots, v_{cm})^T \in \mathbb{R}^{m \times 1} \), \( v_E = (v_{E1}, v_{E2}, \ldots, v_{Em})^T \in \mathbb{R}^{m \times 1} \), and \( v_S = (v_{S1}, v_{S2}, \ldots, v_{Sm})^T \in \mathbb{R}^{m \times 1} \) to denote the prioritization score vectors generated by the complete data matrix \( G_c \), the ensemble and non-ensemble imputation approaches based on the incomplete data matrix \( G_m \), respectively. For the analysis based on the Student’s \( t \)-test, since theoretical \( P \)-values are used for gene prioritization, it is feasible to consider the non-imputation approach: prioritization scores are calculated only based on these observed data in \( G_m \). We use \( v_N = (v_{N1}, v_{N2}, \ldots, v_{Nm})^T \in \mathbb{R}^{m \times 1} \) to denote this score vector.

To answer the aforementioned two questions, we conduct the following two evaluations. We set the size of bootstrap \( b = 100 \), which enables us to generate a possibly huge ensemble. In order to determine the size of ensemble, for different sizes of ensemble, we calculated the Pearson correlation between the prioritization vectors \( v_c \) and \( v_E \). Both the Student’s and SAM \( t \)-tests were considered for prioritizing genes. Figure 2 shows the relationship between the correlation and size of ensemble: the correlation tends to be stable when the size of ensemble is close to \(100\) for all different configurations of sample size and missing rate. Therefore, we fix the size of ensemble \( r = 100 \) for the following evaluations.

When the SAM \( t \)-test is considered for prioritizing genes, it is difficult to evaluate the performance of non-imputation approach. Therefore, we only compare the ensemble imputation approach with the non-ensemble imputation approach.

3.2 Concordance of gene prioritization

One issue about gene prioritization after missing value imputation is whether the prioritization vector, which is used to select genes, is concordant with the one under the situation of no missing values. We use the Pearson correlation coefficient (Pearson, 1894) to measure the concordance. The improvement from the ensemble approach is measured by the Pearson Correlation Improvement Rate (PCIR) defined as follows. We first calculate the Pearson correlation coefficients \( \rho_N \) between \( v_E \) and \( v_c \), and \( \rho_S \) between \( v_S \) and \( v_c \). If \( v_N \) is available, then \( \rho_N \) between \( v_S \) and \( v_c \) is also calculated. Since these correlations are usually large, even small improvement will be considered significant. Therefore, we define the Pearson Correlation Improvement Rate (PCIR) between \( \rho_E \) and \( \rho_S \) as:

\[
PCIR = (\rho_E - \rho_S)/(1 - \rho_S).
\]

PCIR between \( \rho_E \) and \( \rho_S \) can also be similarly calculated.

As shown in Table 1, the proposed ensemble imputation approach consistently outperforms the non-ensemble imputation approach as well as the non-imputation approach: we always obtain positive PCIRs. Compared to the non-ensemble imputation approach, the ensemble imputation approach achieves 50% and higher PCIRs when the sample size is 5+5, and 20% and higher PCIRs when the sample sizes are 10+10 and 15+15. These are observed when either the Student’s or SAM \( t \)-test is used for prioritizing genes.

3.3 Control of true and false positives

When genes are selected after missing value imputation for the follow-up studies, it is necessary to understand the impact of missing value imputation on the control of true and false positives. Since only the selected genes will be used for the follow-up large sample validation studies, it is crucial to control the false positives in genes selected from microarray studies. Controlling the true positives is also important since it is undesirable to miss too many truly differentially expressed genes.

One difficulty to address this issue is that we do not know which genes are truly differentially or non-differentially expressed. In this study, for each configuration, we define gold standards based on the prioritization vector \( v_c \) from the complete data \( G_c \). Previous studies showed that an accurate estimate of the number \( N \) of differentially expressed genes can be obtained when the sample size is relatively large (Lai, 2006). Therefore, we first use the original large and complete data \( G \) to estimate \( N \) with a recently proposed method (Lai, 2006). Then, we define these genes with top \( N \) ranks as gold standard positives and the rest genes as gold standard negatives. With these gold standards defined, we can generate the widely used receiver operating characteristic (ROC) curves (true positive rate against false positive rate for different cutoff points) to compare different approaches. Since different results may be obtained when different test statistics are used for gene prioritization, the above procedure is performed separately for the Student’s and the SAM \( t \)-tests.
When the Student’s t-test is used for gene prioritization, we obtain the estimate $N = 1013 (40.4\%)$; when the SAM t-test is used for gene prioritization, we obtain the estimate $N = 984 (39.3\%)$. Figures 3 and 4 show the ROC curves for different configurations when the Student’s and the SAM t-test are used as test statistic. Although the curves become lower as missing rate increases, the ensemble imputation approach consistently outperforms the non-ensemble imputation and non-imputation approaches. The advantage is especially distinct when the sample sizes are about 5–10 per group and the missing rates
are about 10–20%, which is a common situation for cDNA microarray studies.

4 DISCUSSION
We proposed an ensemble approach to address the issue of gene prioritization after missing value imputation. Compared with the traditional missing value imputation methods, which only provide one estimate for each missing observation, our approach considers the distribution space of prioritization scores due to the existence of missing values. We simulated the distribution space through a bootstrap procedure. To compare different approaches, we evaluated their performances in the concordance of gene prioritization and the control of true/false positives. A published two-sample microarray gene expression data set was used for our evaluations. The results confirmed the
advantages of our proposed approach; the results also allowed us to compare the non-ensemble imputation approach with the non-imputation approach. From Table 1 and Figure 3, the non-imputation approach showed comparable or even better performances in many cases when it was compared with the non-ensemble imputation approach.

We also evaluated the classification performance of genes selected from a pilot study. Support vector machine (Vapnik, 1995) was used as the classifier and different numbers of genes with top ranks were selected. The results showed that a relatively low classification error rate could be achieved, regardless of the choice of approach to missing value imputation, when a certain number (30–100) of genes were included. This is not surprising since the classification performance depends on not only the differentiability of selected genes but also the combination of these genes.

Our proposed ensemble approach is novel. It is capable of incorporating any imputation method for missing value estimation and any statistical test for gene prioritization. In this study, for simplicity, we chose to use the Student’s or SAM test.

---

**Fig. 4.** Improvement of the ensemble method in receiver operating characteristic (ROC) curves when the SAM t-test is used for gene prioritization. Different sample sizes and missing rates are considered. x and y axes represent false and true positive rates, respectively.
\(t\)-test for gene prioritization. Because of its satisfactory performance (Kim et al., 2005), we chose the local least squares coupled with \(L_2\) norm based similarity measure for missing value imputation. Kim et al. (2005) also proposed an automatic estimator for the number \(k\) of nearest neighbor genes, and showed its satisfactory performances. However, this procedure is very time consuming. We actually performed this procedure for some of our evaluations and observed similar results. Since our purpose was to introduce an ensemble approach to gene prioritization after missing value imputation, we simply followed the convention and fixed \(k = 10\) to save our computation time.

It should be noted that the benefits of our proposed approach may be affected by different methods for imputing missing values. Many imputation methods have been proposed and each one has its advantage in a certain situation. It is necessary to conduct further studies so that the impacts of different imputation methods can be well understood. It should also be noted that the current approach may not be applicable to time series microarray data (Gan et al., 2006) since the dependence structure among observations from different time points. We are currently investigating a further development so that we can generate this ensemble idea to integrate the missing value imputation with gene prioritization for time series microarray data.

If the missing rate is almost zero, then there will be no clear difference between the ensemble and the non-ensemble approaches; if almost all the data are missing, then the missing value imputation will not work well and both approaches will have poor performances. If the sample size is extremely small, then the missing value imputation will not work well and there will be no clear difference between the two approaches; if the sample size is relatively large, then the result of gene prioritization will be relatively robust and insensitive to the missing values, and therefore there will be no clear difference between the two approaches. The above discussion implies that there are certain ranges of the sample size and the missing rate such that the ensemble approach will provide considerable improvements over the non-ensemble approach (as we observed in Table 1 and Figs 3 and 4). To better understand the performance of ensemble approach, it is necessary to conduct more theoretical and simulation studies.

ACKNOWLEDGEMENT

We thank the associate editor and two anonymous reviewers for their valuable comments. This work was supported by a NIH grant DK-75004.

Conflict of Interest: none declared.

REFERENCES


