Kolmogorov-Smirnov and Continuous Ranked Probability Score Validation on the Bayesian Model Averaging for Microarray Data

Ani Budi Astuti

Department of Mathematics, Faculty of Mathematics and Natural Sciences
University of Brawijaya Malang Jl. Veteran Malang 65145 Indonesia and
Department of Statistics, Faculty of
Mathematics and Natural Sciences, Institut Teknologi Sepuluh Nopember Surabaya
Kampus ITS Sukolilo, Surabaya 60111 Indonesia

Nur Iriawan, Irhamah and Heri Kuswanto

Department of Statistics, Faculty of Mathematics and Natural Sciences
Institut Teknologi Sepuluh Nopember Surabaya
Kampus ITS Sukolilo, Surabaya 60111 Indonesia

Copyright © 2014 Ani Budi Astuti, Nur Iriawan, Irhamah and Heri Kuswanto. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

The Bayesian Model Averaging (BMA) required the validation step to determine the accuracy of BMA model. Kolmogorov-Smirnov (KS) and Continuous Ranked Probability Score (CRPS) are used to validate the BMA model. The absolute difference between the empirical cumulative distribution and the hypothesis cumulative distribution were the basic idea of these methods. The KS method uses the distance concept and CRPS method uses the area concept. The validation of BMA model on microarray data by KS and CRPS methods would be identified in this paper. The results have succeed to indentify the performance of KS and CRPS in the validation to BMA model on microarray data with an average value of KS=0.469 and CRPS=0.211 for n=10 and then the value of KS=0.403 and CRPS=0.11 for n=12.
Keywords: Kolmogorov-Smirnov, CRPS, Bayesian Model Averaging, Microarray

1. Introduction

Monitoring the activity of thousands of gene expressions simultaneously requires a certain technique, called microarray analysis [22]. Data as a result of microarray treatments called microarray data. Microarray data have some characteristics, limited availability of the samples, large number of variables would be measured, and the distribution of the data would be very complex (multimodal) ([18], [21], and [25]). These phenomena could be solved by Bayesian method. This method does not require the number of samples and for any distribution ([2] and [11]).

BMA is a Bayesian solution to form the best single model from all possible model. Completion of the BMA models based on the average of the posterior distribution from all possible best model ([5] and [10]). Various studies BMA had been done ([1], [4], [6], [9], [12], [19], [20], and [24]). The study of ([1] and [19]) using the microarray data. The importance stage in the modeling of BMA is the identification of BMA model validation.

KS and CRPS are several methods for BMA model validation. The basic concept of these methods is the absolute distance and the area between the empirical cumulative distribution and the hypothesis cumulative distribution. If the value of KS and CRPS smaller then the better model of BMA [17]. Various related studies had been done ([9], [12], [20] and [24]) using the CRPS as BMA model validation. In this paper would be demonstrated KS and CRPS methods for BMA model validation to microarray data.

2. Data and Methods

The data used in this paper is part of the data used by [13]. Data were collected for gene expression differences in diseased and healthy conditions to 10 genes ID with the number of samples n=10 and n=12.

2.1. Bayesian and BMA Analysis

Bayesian method is proposed as a method that combines two information, prior information and likelihood function, to establish the posterior probability distribution model ([2] and [7]). In the Bayesian methods, model parameter of $\theta$ viewed as a random variable. If there are observational data $x$ that have a likelihood function $f(x|\theta)$ then the prior $\theta$, $p(\theta)$, is an information about the unknown parameters before the observations were made. The posterior probability distribution of $\theta$, $p(\theta|x)$, determined by the rules of probability according to Bayes' theorem [11] as in Equation (1).
\[ p(\theta \mid x) = \frac{f(x \mid \theta) p(\theta)}{f(x)} \quad \text{where}, \quad (1) \]

\[ f(x) = E[f(x \mid \theta)] = \int_{\theta \in \Theta} f(x \mid \theta) f(\theta) d\theta \quad \text{if} \quad \theta \quad \text{continuous} \quad \text{and} \]

\[ f(x) = E[f(x \mid \theta)] = \sum_{\theta \in \Theta} f(x \mid \theta) p(\theta) \quad \text{if} \quad \theta \quad \text{discrete}, \quad \text{where} \quad f(x) \quad \text{is a normalized constant} \quad [2]. \quad \text{So that Equation (1) can be written as:} \]

\[ p(\theta \mid x) \propto f(x \mid \theta) p(\theta). \quad (2) \]

Based on the Equation (2), it can be shown the posterior probability will be used for decision making is proportional to the product of the likelihood function and the prior probability of the model parameters [17].

BMA is a Bayesian solution to model uncertainty in order to establish the best single model by considering all possible models. The completion of BMA models by averaging the posterior distribution of all the best models so that the BMA model combines the best of all possible models ([10] and [24]).

If \{M_{1}, M_{2}, ..., M_{q}\} is the set of possible models of \( M \) and \( \Delta \) is the value would be predicted, then the BMA prediction starts with determining the prior probability distribution of all the model parameters and the model \( M_{k} \) ([5] and [10]). Posterior distribution of \( \Delta \mid x \) is as in Equation (3).

\[ P(\Delta \mid x) = \sum_{k=1}^{q} P(\Delta \mid M_{k}, x) P(M_{k} \mid x), \quad (3) \]

where \( q \) is the sum of all the models that may be formed. Posterior distribution of \( \Delta \mid x \) is the average of the posterior distribution whilst the posterior probability of the model \( M_{k} \) is:

\[ P(M_{k} \mid x) = \frac{P(Y \mid M_{k}) P(M_{k})}{\sum_{j=1}^{q} P(Y \mid M_{j}) P(M_{j})}, \quad (4) \]

where \( P(x \mid M_{k}) = \int P(x \mid \theta_{k}, M_{k}) P(\theta_{k} \mid M_{k}) d\theta_{k} \).

\[ (5) \]

Equation (5) is the marginal likelihood of the model \( M_{k} \). The prior probability of \( \theta_{k} \mid M_{k} \) is \( p(\theta_{k} \mid M_{k}) \) and \( p(x \mid \theta_{k}, M_{k}) \) is likelihood function and \( p(M_{k}) \) is prior probability of the model \( M_{k} \). Implicitly, all probabilities depend on the model
$M$ so the expectation value of the coefficient $\Delta$ obtained by averaging the model $M$ as in Equation (6).

$$E(\Delta \mid x) = \sum_{k=1}^{d} P(M_k \mid x)E(\Delta \mid M_k, x).$$ \hspace{1cm} (6)$$

Value of $E(\Delta \mid x)$ in Equation (6) shows the weighted expected value of $\Delta$ in every possible combination models whilst the variance of $(\Delta \mid x)$ is as in Equation (7).

$$Var(\Delta \mid x) = \sum_{k=1}^{d} (\text{var}(\Delta \mid x, M_k) + [E(\Delta \mid M_k, x)]^2)P(M_k \mid x) - E(\Delta \mid x)^2).$$ \hspace{1cm} (7)$$

2.2. Markov Chain Monte Carlo (MCMC) Algorithm with Gibbs Sampler Approach

In the Bayesian analysis, the posterior distribution is often very complicated and requires a difficult process of integration so that takes a numerical approximation method of Markov Chain Monte Carlo (MCMC) [17]. MCMC approach will facilitate to quite complex model so this method is considered as a breakthrough in the use of Bayesian analysis [2]. MCMC method is a simulation method that combines Monte Carlo and Markov Chain nature to obtain the sample data based on specific sampling scenarios. Figure 1 is presented MCMC algorithm with Gibbs Sampler approach.

Step 1. Set initial values of $\theta^{(k)}$ at $k = 0$ so that $\theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_r^{(0)})$.

Step 2. The sampling process to obtain the value $\theta_j$ from conditional distribution by sampling for $\theta_j^{(k+1)}$ in $r$ steps as follows:
1. Sampling $\theta_1^{(k+1)}$ from $p(\theta_1 \mid X, \theta_2^{(k)}, \ldots, \theta_r^{(k)})$
2. Sampling $\theta_2^{(k+1)}$ from $p(\theta_2 \mid X, \theta_1^{(k)}, \theta_3^{(k)}, \ldots, \theta_r^{(k)})$
   .
   .
   .
   r. Sampling $\theta_r^{(k+1)}$ from $p(\theta_r \mid X, \theta_1^{(k)}, \theta_2^{(k)}, \ldots, \theta_{r-1}^{(k)})$.

Step 3. Doing iteration at step 2 as $M$ times with $M \to \infty$.

Figure 1. MCMC Algorithm with Gibbs Sampler Approach

2.3. Kolmogorov-Smirnov (KS)

According to [17], the theoretical concept of Kolmogorov-Smirnov test is to compare the empirical cumulative distribution function (CDF), $F_n(x_i)$, and the
hypothesis cumulative distribution function, \( \hat{F}(x) \). If \( X_{(1)}, X_{(2)}, \ldots, X_{(n)} \) is statistic order to the independent random variable with hypothesis distribution, \( \hat{F}(x) \) and the empirical distribution is as in Equation (8).

\[
F_n(x_i) = \frac{\text{the number of data } X_j \leq x_i}{n} \quad \text{for } i=1, 2, 3, \ldots, k < n,
\]

where \( F_n(x_i) \) is right continuous step function.

The formula of test statistic \( D_n \) can be written as in Equation (9).

\[
D_n = \sup \left| F_n(x) - \hat{F}(x) \right|.
\]

If the value of \( D_n \) smaller then the tested models is better. The concept of Kolmogorov-Smirnov test is shown in Figure 2.

![Figure 2. The Concept of Kolmogorov-Smirnov](image)

According to ([17] and [26]), requirements of Kolmogorov-Smirnov test are continuous distribution and for any sample size \( n \). Modification of the Kolmogorov-Smirnov test was performed to the exponential, normal, and Weibull distributions [3].

2.4. Continuous Ranked Probability Score (CRPS)

The theoretical concept of CRPS similar to the theoretical concept of Kolmogorov-Smirnov test. The difference is the use of area concept to CRPS whilst KS uses the distance concept. CRPS value is the value of Mean Square Error of the
hypothesis cumulative distribution [16]. According to ([8] and [14]), the formula of CRPS is shown in Equation (10).

\[ CRPS(P, x_{a}) = \int_{-\infty}^{\infty} \left( P(x) - P_{a}(x) \right)^{2} dx, \]  

(10)

where \( P(x) \) is CDF hypothesis and \( P_{a}(x) \) is empirical CDF.

If the value of CRPS smaller then the tested models is better [23]. According to [15], the concept of CRPS is shown in Figure 3.

![Figure 3. The Concept of CRPS](image)

According to ([14] and [25]), some of the advantages of using the evaluation of CRPS are good sensitivity of the calculation to the entire value of the parameters tested, easily interpreted, the sharpness of its value spreads on each outcome expectations, and can reduce the mean absolute error if the point estimators.

3. Results and Discussions

3.1. Identification of Distribution and Parameter Estimators from BMA Model

The identification results of distribution and parameter estimators for BMA model to gene expression differences data in desease and healthy conditions with \( n=10 \) and \( n=12 \) is shown in Table 1.
Based on Table 1 it can be seen that the distributions of 10 genes ID at n=10 and n=12 have a normal distribution.

### 3.2. BMA Model Validation by CRPS and KS

BMA model validation results based on the value of KS and CRPS to gene expression differences data in diseased and healthy conditions for 10 genes ID with n=10 and n=12 are shown in Table 2, Table 3, Figure 4, and Figure 5.
Figure 4. Plot of KS and CRPS Value, n=10

Table 3. BMA Model Validation by KS and CRPS for n=12

<table>
<thead>
<tr>
<th>No</th>
<th>Gene ID</th>
<th>Total Component Model in the BMA (%)</th>
<th>CRPS</th>
<th>KS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Normal Model of BMA</td>
<td>Univariate Normal Model</td>
</tr>
<tr>
<td>1</td>
<td>H55933</td>
<td>79</td>
<td>1.77E-01</td>
<td>2.00E-01</td>
</tr>
<tr>
<td>2</td>
<td>R39465-1</td>
<td>22.1</td>
<td>9.67E-02</td>
<td>0.127055134</td>
</tr>
<tr>
<td>3</td>
<td>R39465-2</td>
<td>32.3</td>
<td>7.70E-02</td>
<td>9.69E-02</td>
</tr>
<tr>
<td>4</td>
<td>R85482</td>
<td>65.9</td>
<td>4.56E-02</td>
<td>4.65E-02</td>
</tr>
<tr>
<td>5</td>
<td>U14973</td>
<td>15.7</td>
<td>5.72E-02</td>
<td>6.13E-02</td>
</tr>
<tr>
<td>6</td>
<td>R02593</td>
<td>1.3</td>
<td>4.94E-02</td>
<td>6.39E-02</td>
</tr>
<tr>
<td>7</td>
<td>T51496</td>
<td>69.3</td>
<td>5.87E-02</td>
<td>6.11E-02</td>
</tr>
<tr>
<td>8</td>
<td>H80240</td>
<td>75.2</td>
<td>0.334861488</td>
<td>0.366986407</td>
</tr>
<tr>
<td>9</td>
<td>T65938</td>
<td>3.6</td>
<td>0.133476163</td>
<td>0.135410537</td>
</tr>
<tr>
<td>10</td>
<td>T55131</td>
<td>4.7</td>
<td>7.13E-02</td>
<td>1.06E-01</td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td></td>
<td>1.10E-01</td>
<td>1.27E-01</td>
</tr>
</tbody>
</table>
Based on Table 2, Table 3, Figure 4 and Figure 5, it can be seen that the value of CRPS and KS from BMA model is smaller than the univariate model. This shows the model of the BMA has a very good level of model accuracy based on the value of KS and CRPS with an average value of KS=0.469 and CRPS=0.211 for n=10 and the value of KS=0.403 and CRPS=0.11 for n=12.

4. Conclusion

In this paper, the BMA model for microarray data is normally distribution with the average value of KS=0.469 and CRPS=0.211 for n=10 and the value of KS=0.403 and CRPS=0.11 for n=12. BMA model is better than univariate model. This is shown by the average value of KS and CRPS BMA model smaller than the univariate model. The performance of KS and CRPS is very good in the BMA model validation for microarray data.

Acknowledgements. This paper is part of the Doctoral Research at Department of Statistics, Institut Teknologi Sepuluh Nopember (ITS), Surabaya, Indonesia. We would like to thank to reseach group of Do, Muller and Tang to use his data and to anonymous refewer to this paper.

References


Received: August 29, 2014