Journal of Science and Technology, Vol. 9 No. 3 (2017) p. 140-144

Gene Selection for Colon Cancer Classification using Bayesian Model Averaging of Linear and Quadratic Discriminants

Oyebayo Ridwan Olaniran*, Mohd Asrul Affendi Abdullah

Department of Mathematics and Statistics, Faculty of Applied Sciences and Technology, Universiti Tun Hussein Onn Malaysia, Pagoh Educational Hub, 84600 Pagoh, Johor, Malaysia.

Received 30 September 2017; accepted 30 November 2017; available online 28 December 2017

Abstract: Recent findings reveal that various cancer types can be diagnosed using non-clinical approach which involves monitoring of the biological samples using their genes expression profiles. Two of the widely used methods are Linear and Quadratic discriminant analyses. In this paper, the behaviours of Linear and Quadratic Discriminants Analyses (LDA and QDA) were observed within the framework of Bayesian model averaging. We applied Bayesian model averaging to tackle model uncertain problem inherent in discriminant analysis. We calibrated the developed classifier on published real life microarray colon cancer data that contained 2000 gene expression profiles measured on 62 biological samples that comprised 40 tumorous tissue samples and 22 normal tissue samples. The data were also pre-processed using logarithmic transformation to base 10 and zero mean unit variance normalization as often done with the dataset. In addition, comparison with Random Forest (RF), Gradient Boosting Machine (GBM) and Bayesian Additive Regression Trees (BART) was also achieved. Various performance results established the supremacy of the proposed method.

Keyword: Cancer; Linear and Quadratic discriminant; Bayesian; Model Uncertainty; Model averaging.

1. Introduction

Recent findings reveal that various cancer types can be diagnosed using non-clinical approach which involves monitoring of the biological samples using their genes expression profiles. However, this advancement is possible due to the enhancement of microarray technology which made it possible to observe gene expression levels of several gene chips concurrently [1, 2]. Several authors have discussed the health benefits of the non-clinical diagnosis breakthrough, but the major problem that still exists is how to adequately identify the few subsets of thousands genes whose information can be used to reliably classify the mRNA samples into their respective biological groups. In addition, it has been observed that adequacy of any method strongly depends on the health problems [3].

Several types of machine learning algorithms have been proposed to perform the task of non-clinical diagnosis mRNA (messenger Ribonucleic acid). mRNA samples are usually collected one several biological features (genes). The resulting data structure are of the form $n \ll p$, where the number of patients n is far less than the number of biological features. This scenario is often termed as High-dimensional data [4]. High-dimensionality poses serious problem in statistical analysis and in-fact when building machine learning algorithms. This is because most statistical methods require the number of patients (equations) to be more than the number of attributes (parameters) for a unique solution to exist.

Bayesian procedures are the emerging solution to most applications of statistics in the recent time. In fact, it has the least error rate in theory [5,6]. LDA and QDA are often regarded as the Bayesian classifier [2, 7] because they are motivated by Bayes theorem.

Two of the important assumptions of LDA and QDA is normality assumption of the feature space (x) and also orthogonality of feature space [4]. The efficiency or accuracy of LDA and QDA classifier strongly relies on these two assumptions. The classifiers become unstable when the assumptions are not met.

High-dimensional data usually violates these assumptions as in the case of the data used in this research. The data do not satisfy the normality assumption as well as the dimensionality problem often lead to multicollinearity. In the light of this, robust

*Corresponding author: rid4stat@yahoo.com

penerbit.uthm.edu.my/ojs/index.php/jst

methods like Random Forests [8] stochastic gradient boosting [9], Bayesian additive regression trees [10], Bayesian additive regression trees using Bayesian model averaging [6]. have been developed. The methods are efficient but are computationally expensive than the simple LDA and QDA. Therefore, in this paper, we developed ensemble of LDA and QDA using Bayesian model averaging approach in order to increase the efficiency of LDA and QDA when analysing high-dimensional data.

2. Bayes Classifier

The foremost Bayesian classification methods are linear and quadratic discriminant analysis [11]. Specifically, Bayes classifier is defined by $f_C(x)$ over a random vector X and random vector Y where

$$f_C(x) \equiv Pr(X = x \mid Y = c) \qquad (1)$$

denoting the density function of X for an observation that comes from the Cth class. In other words, $f_C(x)$ is relatively large if there is a high probability that an observation in the Cth class has $X \approx x$, and $f_C(x)$ is small if it is very unlikely that an observation in the Cth class has $X \approx x$. Then Bayes' theorem states that;

$$Pr(X = x \mid Y = c) = \frac{\pi_C f_C(x)}{\sum_{l=1}^{C} \pi_l f_l(x)} \qquad (2)$$

If we assume x is normally distributed and estimate its associated location and scale parameter by maximizing the likelihood implies we are constructing a Linear Discriminant Analysis (LDA) or Quadratic Discriminant Analysis (QDA). The classifier constructed is LDA if we assuming equality of class variance if otherwise its QDA.

Formally, the discriminant $\delta_c(x)$ for a class c is define as;

$$\delta_c(x) = x^T \Sigma^{-1} \mu_C - \frac{1}{2} \mu_C^T \Sigma^{-1} \mu_C + log \pi_c \qquad (3)$$

where μ_C is the location mean for class c, and Σ is covariance matrix of p predictors for class c. (2) is often referred to as LDA since we assume class variance are the same across predictors. If otherwise we can obtain QDA as:

$$\delta_c(x) = x^T \Sigma^{-1} \mu_C - \frac{1}{2} \mu_C^T \Sigma_c^{-1} \mu_C + log \pi_c \qquad (4)$$

The estimates of the unknown parameters $\mu_1, \ldots, \mu_C$, $\pi_1, \ldots, \pi_C$, and Σ are usually estimated via Maximum Likelihood (MLE, [11]) as;

$$\hat{\mu}_C = \frac{1}{n_c} \sum_{i:y_i=C}^{n_c} x_i$$

$$\hat{\Sigma} = \frac{n_c - 1}{n - C} \sum_{c=1}^{C} \hat{\Sigma}_c$$

The estimates obtained are then plugged into (3) and (4) to determine the LDA and QDA classifier.

3. Bayesian Model Averaging of LDA and QDA

Given a classification rule $\delta(x)$ as earlier derived, and its associated prior probability $P(\delta)$, then we can define a posterior density

$$P(\delta_k|Y) = \frac{L(\delta_k(x))P(\delta_k)}{\sum_{k=1}^{m} L(\delta_k(x))P(\delta_k)} \qquad (5)$$

as density of all m possible classifiers. Usually, what determines classifier k is the subset of feature x in the matrix. For each k, there exist r subset of x space.

Following the earlier work of Clyde on Bayesian model averaging in [12]. It was derived that under regularized prior $P(\delta)$ can be approximated with the posterior $P(\delta_k|Y)$ as;

$$P(\delta_k|Y) = \frac{exp(-0.5BIC(\delta_k))P(\delta_k)}{\sum_{k=1}^{m} exp(-0.5BIC(\delta_k))P(\delta_k)} \qquad (6)$$

$$P(\delta_k|Y) = \frac{exp(-0.5BIC(\delta_k))}{\sum_{k=1}^{m} exp(-0.5BIC(\delta_k))} \qquad (7)$$

where;

$$L\big(\delta_k(x)\big) = exp(-0.5BIC(\delta_k))$$

$$BIC(\delta_k) = -2log\big[L\big(\delta_k(x)\big)\big] + plog(n)$$

After posterior estimation, we can then estimate parameter of interest by;

$$P[\Delta|Y] = \sum_{k=1}^{m} P(\delta_k|Y)P(\Delta|\delta_k, Y)$$

Specifically, for LDA and QDA, the posterior class probabilities are:

For LDA;

$$P[X = x|Y = c] = \sum_{k=1}^{m} P(LDA|Y)P(X = x|LDA, Y) \quad (8)$$

For QDA;

$$P[X = x|Y = c] = \sum_{k=1}^{m} P(QDA|Y)P(X = x|QDA, Y) \quad (9)$$

Equation (8) and (9) correspond to proposed methods used in this research.

4. Data Calibration

The data employed for this study were obtained from microarray Princeton repository (http://microarray.princeton.edu/oncology/affydata/index.html.) on colon cancer. The data contained 2000 gene expression profiles measured on 62 biological samples that comprised 40 tumorous tissue samples and 22 normal tissue samples. Details of the micro-array experiment that produces the data can be found in the works of Alon works in [13]. The data were also pre-processed using logarithmic transformation to base 10 and zero mean unit variance normalization as often done with the dataset. We compared the performance of the methods BMA-LDA and BMA-QDA with LDA, QDA, RF, BART and GBM using the class specific and overall performance metrics. The metrics used include overall accuracy, balance accuracy, sensitivity, specificity, negative predictive value, positive predictive value, false positive and false negative. The metrics are computed based on the confusion matrix between 10 folds cross validated test samples and the actual class. The confusion matrix is presented in Table 1 as observed in [14];

Table 1 Confusion matrix

True Class	Predicted Class 0 1		Total
0	TN	FP	N
1	FN	TP	P
Total	N*	P*	T

0: Normal, 1: Tumour

where TN represents True Negative, FP is the False Positive, FN represents False Negative and TP is the True Positive. Also, N* is the total predicted negative and P* represents total predicted positive. Similarly, N is the total actual negative while P is the total actual positive. T represents the total number of observation equivalent to;

$$T = TN + FP + TP + FN \quad (10)$$

Here negative means normal cells while positive means tumour cells.

Accuracy (ACC): $\%ACC = 100 \times \left(\frac{TN+TP}{T}\right)$

Sensitivity:$\%Sensitivity = 100 \times \left(\frac{TP}{P}\right)$

Specificity:$\%Specificity = 100 \times \left(\frac{TN}{N}\right)$

Balance Accuracy (BACC):

$$\%BACC = \frac{\%Sensitivity + \%Specificity}{2}$$

Positive Predictive Value (PPV):

$$\%PPV = 100 \times \left(\frac{TP}{P^*}\right)$$

Positive Predictive Value (PPV):

$$\%NPV = 100 \times \left(\frac{TN}{N^*}\right)$$

False Positive Rate (FPR):

$$\%FPR = 100 - \%Specificity$$

False Negative Rate (FNR):

$$\%FNR = 100 - \%Sensitivity$$

Misclassification Error Rate (MER):

$$\%MER = 100 - \%ACC$$

5. Results and Discussion

Table 2 and Table 3 show the results based on $v - folds$ cross validation where $v = 10$. The overall performance measure using accuracy revealed that the most accurate classifier for diagnosing colon cancer is either BMA-LDA or BMA-QDA. To control class imbalance, we computed the balance accuracy which also reveals that BMA-LDA or BMA-QDA are the most accurate classifiers.

Table 2 Performance measures in (%) for BMA-LDA, BMA-QDA, LDA and QDA based on average of 10 folds cross validation.

Metrics	Methods			
	BMA-LDA	BMA-QDA	LDA	QDA
sens	96.7	96.7	91.7	53.3
specs	95.0	95.0	87.5	92.5
PPV	93.3	93.3	83.3	85.4
NPV	97.5	97.5	95.5	80.5
FP	5.0	5.0	12.5	7.5
FN	3.3	3.3	8.3	46.7
mer	4.5	4.5	11.2	22.1
acc	95.5	95.5	88.8	77.9
BACC	95.8	95.8	89.6	72.9

Table 3 Performance measures in (%) for RF, BART and GBM based on average of 10 folds cross validation.

Metrics	Methods		
	RF	BART	GBM
sens	68.3	40.0	76.7
specs	87.5	90.0	90.0
PPV	81.7	76.2	85.2
NPV	85.0	75.2	90.2
FP	12.5	10.0	10.0
FN	31.7	60.0	23.3
mer	19.3	27.4	14.5
acc	80.7	72.6	85.5
BACC	77.9	65.0	83.3

The class specific metrics is very important when diagnosing cancer. The procedure must be highly sensitive for it to be able to detect its presence as early as possible. Sensitivity result in Table 1 shows that the most sensitive classifiers are BMA-LDA and BMA-QDA with about 97% sensitivity.

6. Conclusion

In this paper, we have presented Bayesian model averaging approach of LDA and QDA for non-clinical diagnosis of colon cancer. The results using the real life dataset indicated that BMA-LDA and BMA-QDA are the best in terms of overall diagnostic accuracy as well as class specific accuracy.

Acknowledgement

This work was supported by Universiti Tun Hussein Onn, Malaysia (grant number Vot, U607).

References

[1] Yahya W. B., Olaniran O. R. and Ige, S. O. (2014). “On Bayesian Conjugate Normal Linear Regression and Ordinary Least Square Regression Methods: A Monte Carlo Study” in Ilorin Journal of Science, Vol. 1 No. 1 pp. 216-227.

[2] Olaniran, O.R., Olaniran, S. F., Yahya, W. B., Banjoko, A. W., Garba, M. K., Amusa, L. B. and Gatta, N. F.(2016). “Improved Bayesian Feature Selection and Classification Methods Using Bootstrap Prior Techniques” in Anale. Seria Informatică. Vol. 14 No. 2 pp. 46-52.

[3] Sim, Adelene YL;Minary, Peter; Levitt, Michael (2012). "Modeling nucleic acids" in Current Opinion in Structural Biology Nucleic acids/Sequences and topology. Vol. 22 No. 3 pp. 273–278.

[4] Hastie, T., James, G., Witten, D., & Tibshirani, R. (2013). An Introduction to statistical learning. Springer, New York.

[5] Lesaffre E. and Lawson, A. B. (2013). Bayesian Biostatistics. John Wiley & Sons, Ltd, New Jersey.

[6] Hernández, B., Raftery, A. E., Pennington, S. R., & Parnell, A. C. (2015).“Bayesian Additive Regression Trees using Bayesian Model Averaging”.

[7] Olaniran, O. R., & Yahya, W. B. (2017). "Bayesian Hypothesis Testing of Two Normal Samples using Bootstrap Prior Technique" in Journal of Modern Applied Statistical Methods, Vol. 16 No. 2 pp. 618-638.

[8] Breiman, L. (2001). "Random forests". Machine Learning, Vol. 45, pp. 5–32.

[9] Friedman, J. H. (2002). "Greedy function approximation: A gradient boosting machine" in Annals Statistics., Vol. 29 No. 5 pp. 1189–1232.

[10] Chipman, H.A. George, E. I. and McCulloch R. E. (2010). "BART: Bayesian Additive Regression Trees" in Annals Applied Statistics, Vol. 4, pp. 266–298.

[11] Hastie T, Tibshirani R, Friedman J (2011). The Elements of Statistical Learning: Prediction, Inference and Data Mining. 2nd edition. Springer-Verlag, New York.

[12] Clyde, M. (1999). "Bayesian Model Averaging and Model Search Strategies (with discussion)." in Bayesian Statistics 6, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, 157{185. Oxford: Oxford University Press.

[13] Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D., Levine, A.J.(1999). "Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays". in Proceeding of National Academia of Science, Vol. 96 pp. 6745–6750.

[14] Banjoko, A. W., Yahya, W. B., Garba, M. K., Olaniran, O. R., Dauda, K. A., & Olorede, K. O. (2015). "Efficient Support Vector Machine Classification of Diffuse Large B-Cell Lymphoma and Follicular Lymphoma MRNA Tissue Samples". Annals. Computer Science Series, Vol. 13 No. 2 pp. 69-79.