The Performance of Predictive Classification using Empirical Bayes

N. Deetae, S. Sukparungsee, Y. Areepong, and K. Jampachaisri

***Abstract*—**This research is aimed to compare the percentages of correct classification of Empirical Bayes method (EB) to Classical method when data are constructed as near normal, short-tailed and long-tailed symmetric, short-tailed and long-tailed asymmetric. The study is performed using conjugate prior, normal distribution with known mean and unknown variance. The estimated hyper-parameters obtained from EB method are replaced in the posterior predictive probability and used to predict new observations. Data are generated, consisting of training set and test set with the sample sizes 100, 200 and 500 for the binary classification. The results showed that EB method exhibited an improved performance over Classical method in all situations under study.

***Keywords*—**Classification, Empirical Bayes, Posterior predictive probability.

I. Introduction

DISCRIMINATION and classification are techniques often used together. The goal of discrimination is to search for distinct characteristics of observations in the training set of sample with known classes and used to construct a rule, called discriminant, to separate observations as much as possible [1], [2]. Classification is focused on allocating new observations in the test set of sample to labeled classes based on well-defined rule obtained from the training set [3]. Generally, classification can be performed with various methods, such as Bayesian, Nearest Neighbor, Classification Tree, Support Vector Machines and Neural Network etc.

Bayesian method is one of the most popular methods. This method is simple and effective for classification [4]. The Bayesian classification method, which classified observations into related classes using a decision rule, is known for its flexibility and accuracy [5]. Decision rule is defined by the posterior probability which class membership is indicated based on its highest posterior probability [6], [7]. Duarte-Mermoud and Beltran [8] proposed the Bayesian network in classification of Chilean wines and compared radial basis function neural network with support vector machine. The results disclosed that Bayesian network gave the best performance with 91% of correct classification in the test set. Williams and Barber [5] studied the problem of assigning an input vector into several classes using Gaussian prior in Bayesian classification and it was generalized to multiclass problem. Porwal and Carranza [9] had proposed classifiers using Bayesian network and applied to classify mineral deposit occurrence.

The Bayesian classification involves in hyper-parameters that should be known or be able to assess using previous knowledge prior to data collection [10]. However, little information is sometimes available in practice, causing the assessment of hyper-parameters impossible. As a result, Empirical Bayes (EB) can be utilized to estimate the unknown hyper-parameters using information in the observed data. Li [11] exhibited the use of EB to estimate unknown parameters and classify a set of unidentified input patterns into k separate classes using stochastic approximation. In addition, the results from Monte Carlo simulation study demonstrated the favorable estimation of unknown parameters in normal distribution with EB. Chang and Li [12] illustrated the use of EB to classify a new item or product using stochastic approximation with Weibull distribution into two classes (good or defective) or identified an item as produced from one of two production lines. Wei and Chen [13] studied EB estimation in two-way classification model. The results showed that EB yielded smaller mean square error matrix than the least sum of squares method. Ji, Tsui, and Kim [14] adopted EB to classify gene expression profiles, leading to the decrease of number of nuisance parameters in the Bayesian model.

With parametric classification, data are assumed to be normally distributed which rarely occurred in practice except for large sample. Sometimes, data may be distributed symmetrically or asymmetrically with long tail and short tail. EB, consequently, can be utilized to produce the posterior predictive probability of assigning new observations to known classes and it is believed to improve performance over Classical method. The objective of this research is to develop a classification technique to classify various types of data; near normal, short-tailed and long-tailed symmetric, short-tailed and long-tailed asymmetric, using EB method. The estimated hyper-parameters obtained from EB are substituted in the posterior predictive probability and used to classify new observations and then compared to Classical method. Data employed in this research are simulated and divided into two

N. Deetae is a PhD student in the Department of Applied Statistics at King Mongkut’s University of Technology North Bangkok, Bangkok, Thailand. (phone: +66870713312; fax: +6625856105; e-mail: natthineed@hotmail.com).

S. Sukparungsee is an Assistant Professor with the Department of Applied Statistics, King Mongkut’s University of Technology North Bangkok, Bangkok, Thailand (e-mail: swns@kmutnb.ac.th).

Y. Areepong is an Assistant Professor with the Department of Applied Statistics, King Mongkut’s University of Technology North Bangkok, Bangkok, Thailand (e-mail: yupaporn@ kmutnb.ac.th).

K. Jampachaisri is an Assistant Professor with the Department of Mathematics,Naresuan University, Phitsanulok (e-mail: katechanj@nu.ac.th).

sets: a set of sample used to create a rule, called training data, and a set of sample used to evaluate a rule derived from training data, called test data. In each situation, the percentage of correct classification is considered.

II. Empirical Bayes Method

EB can be performed using either parametric or nonparametric methods. With parametric EB, the prior distribution is assumed to be known, in contrast with nonparametric EB [10]. The estimation of hyper-parameters with EB method can be obtained from posterior marginal distribution function as follow

$$m(x\,|\,\delta)=\int f(x\,|\,\theta)\pi(\theta\,|\,\delta)d\theta \tag{1}$$

where θ is parameter which is continuous random variable in this case

δ is hyper-parameter

$m(x\,|\,\delta)$ is posterior marginal distribution function

In this research, the form of prior distribution, conjugate prior, is assumed to be known with known mean (θ_0) and unknown variance (σ^2**)**.

Informative prior: $\sigma^2 \sim IG(\alpha,\beta)$ that is,

$$\pi(\sigma^2)=\frac{\beta^{\alpha}(\sigma^2)^{-(\alpha+1)}\,\mathrm{e}^{-\beta/\sigma^2}}{\Gamma(\alpha)};\sigma^2>0,\alpha>0,\beta>0 \tag{2}$$

The steps of EB method are demonstrated below:

Step i: Find posterior marginal distribution function.

Consider probability distribution function of X .

$$f(x\,|\,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}\mathrm{e}^{-\frac{1}{2\sigma^2}(x-\theta_0)^2};-\infty<x<\infty,\sigma^2>0 \tag{3}$$

Posterior marginal distribution function is

$$m(x\,|\,\alpha,\beta)=\int_0^{\infty}f(x\,|\,\sigma^2)\pi(\sigma^2)d\sigma^2. \tag{4}$$

Therefore, posterior marginal distribution function of X is

$$m(x\,|\,\alpha,\beta)=\frac{\beta^{\alpha}}{\sqrt{2\pi}\cdot\Gamma(\alpha)}\cdot\frac{\Gamma\left(\frac{2\alpha+1}{2}\right)}{\left(\frac{(x-\theta)^2+2\beta}{2}\right)^{\left(\frac{2\alpha+1}{2}\right)}}. \tag{5}$$

Step ii: Find estimators of hyper-parameter using maximum likelihood method and use the Newton-Raphson method to solve a nonlinear equation.

Therefore, the estimators of $\hat{\alpha}$ and $\hat{\beta}$ are

$$\begin{bmatrix}\hat{\alpha}^{(r+1)}\\ \hat{\beta}^{(r+1)}\end{bmatrix}=\begin{bmatrix}\hat{\alpha}^{(r)}\\ \hat{\beta}^{(r)}\end{bmatrix}-\begin{bmatrix}\frac{\partial^2L^{(r)}}{\partial\alpha^2} & \frac{\partial^2L^{(r)}}{\partial\alpha\partial\beta}\\ \frac{\partial^2L^{(r)}}{\partial\beta\partial\alpha} & \frac{\partial^2L^{(r)}}{\partial\beta^2}\end{bmatrix}^{-1}\times\begin{bmatrix}\frac{\partial L^{(r)}}{\partial\alpha}\\ \frac{\partial L^{(r)}}{\partial\beta}\end{bmatrix} \tag{6}$$

where

$$L\left(X_i;\alpha,\beta\right)=\sum\nolimits_{i=1}^{n}\alpha\ln\beta+\ln\Gamma\left(\alpha+\frac{1}{2}\right)-\ln\sqrt{2\pi}$$
$$-\ln\Gamma\left(\alpha\right)-\left(\alpha+\frac{1}{2}\right)\cdot\ln\left(\frac{1}{2}\left(x_i-\theta\right)^2+\beta\right)$$

$$\frac{\partial^2L^{(r)}}{\partial\alpha^2}=\frac{n}{\alpha^{(r)}+\frac{1}{2}}+\frac{2n}{\left(2\alpha^{(r)}+1\right)^2}-\frac{n}{\alpha^{(r)}}-\frac{n}{2\alpha^{2(r)}}$$

$$\frac{\partial^2L^{(r)}}{\partial\alpha\partial\beta}=\frac{n}{\beta^{(r)}}-\sum\nolimits_{i=1}^{n}\frac{1}{\frac{1}{2}(x_i-\theta)^2+\beta^{(r)}}$$

$$\frac{\partial^2L^{(r)}}{\partial\beta^2}=-\frac{n\alpha^{(r)}}{\beta^{2(r)}}+\sum\nolimits_{i=1}^{n}\frac{\alpha^{(r)}+\frac{1}{2}}{\left[\frac{1}{2}(x_i-\theta)^2+\beta^{(r)}\right]^2}$$

$$\frac{\partial L^{(r)}}{\partial\alpha}=n\ln\beta^{(r)}+n\ln\left(\alpha^{(r)}+\frac{1}{2}\right)-\frac{n}{2\alpha^{(r)}+1}-n\ln\alpha^{(r)}$$
$$+\frac{n}{2\alpha^{(r)}}-\sum\nolimits_{i=1}^{n}\left[\ln\left(\frac{1}{2}(x_i-\theta)^2+\beta^{(r)}\right)\right]$$

$$\frac{\partial L^{(r)}}{\partial\beta}=\frac{n\alpha^{(r)}}{\beta^{(r)}}-\sum\nolimits_{i=1}^{n}\frac{\left(\alpha^{(r)}+\frac{1}{2}\right)}{\frac{1}{2}(x_i-\theta)^2+\beta^{(r)}}$$

and r represents the iteration number.

Step iii: Find the posterior distribution function.

$$\pi(\sigma^2\,|\,\underline{x})=\frac{f(\underline{x}\,|\,\sigma^2)\pi(\sigma^2)}{\int_0^{\infty}f(\underline{x}\,|\,\sigma^2)\pi(\sigma^2)d\sigma^2} \tag{7}$$

Thus, the posterior distribution function is

$$\sigma^2\,|\,\underline{X}\sim IG\left(\frac{2\alpha+n}{2},\frac{\sum_{i=1}^{n}\left(x_i-\theta_0\right)^2+2\beta}{2}\right). \tag{8}$$

Step iv: Take estimators of hyper-parameter from Step ii and replaced into the posterior distribution function.

$$\sigma^2\,|\,\underline{X}\sim IG\left(\frac{2\hat{\alpha}+n}{2},\frac{\sum_{i=1}^{n}\left(x_i-\theta_0\right)^2+2\hat{\beta}}{2}\right) \tag{9}$$

Step v: Compute the posterior predictive probability.

The posterior predictive probability is frequently used for prediction of new observation, y , in test data [15]. The posterior predictive probability of y conditionally on $\underline{x}$, denoted by

$$p(y \mid \underline{x}, \sigma^2) = \int_{\sigma^2} f(y \mid \sigma^2)\pi(\sigma^2 \mid \underline{x})d\sigma^2 \tag{10}$$

where

$$f(y \mid \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y-\theta_0)^2}$$

$$\pi(\sigma^2 \mid \underline{x}) =$$

$$\frac{\left(\frac{\sum_{i=1}^{n}(x_i-\theta_0)^2+2\beta}{2}\right)^{\frac{(2\alpha+n)}{2}} \left(\sigma^2\right)^{-\left(\frac{n}{2}+\alpha+1\right)} e^{-\frac{1}{\sigma^2}\left(\frac{\sum_{i=1}^{n}(x_i-\theta_0)^2+2\beta}{2}\right)}}{\Gamma\left(\frac{2\alpha+n}{2}\right)}$$

Sometimes the posterior predictive probability is not a tractable, so Markov Chain Monte Carlo (MCMC) technique [16] can be applied to estimate $p\left(y \mid \underline{x}, \sigma^2\right)$ as

$$\hat{p}(y \mid \underline{x}, \sigma^2) \approx \frac{1}{M}\sum_{t=1}^{M} p(y \mid \underline{x}, \sigma^{2(t)}) \tag{11}$$

where M is the generated MCMC samples size and $\sigma^{2(t)},\ t = 1, 2, ..., M$ is the generated MCMC samples.

Step vi: Classify the test data into classes based on the highest posterior predictive probability.

Step vii: Compute the percentages of correct classification.

III. Simulation results

Five characteristics of data are constructed by varying the values of skewness and kurtosis based on Shapiro, Wilk and Chen criteria [17] and Ramberg and Tadikamallla [18] as (skewness, kurtosis); (0.25, 2.80), (0.00, 2.40), (0.00, 6.00), (0.75, 2.80), and (0.75, 6.00) for near normal, short-tailed and long-tailed symmetric, short-tailed and long-tailed asymmetric, respectively. The percentages of correctly classified data in the case of known mean and unknown variance using EB and Classical method are shown in Table I. Table II illustrated the percentages difference of correctly classified data between EB and Classical method.

With three levels of sample sizes, both Classical and EB methods exhibited good classification when data were constructed as near normal, all the results were shown in Table I. In addition, EB method indicated an improved performance over Classical method in all situations under study, as displayed in Table II.

IV. Conclusion

Simulation results of this research suggested that classification using EB exhibited outperformance to Classical method in all cases.

TABLE I
Percentages of Correctly Classified Data using Classical And EB Method

Sample sizes (n)	Data distribution	Methods	
		Classical	EB
100	Near Normal	98.0700	98.0790
	Symmetric short -tailed	97.5030	97.5760
	Symmetric long -tailed	97.7020	97.7030
	Asymmetric short-tailed	97.5200	97.5880
	Asymmetric long -tailed	97.6130	97.6220
200	Near Normal	98.0965	98.1180
	Symmetric short -tailed	98.0415	98.0425
	Symmetric long -tailed	97.5325	97.5340
	Asymmetric short-tailed	97.5450	97.6000
	Asymmetric long -tailed	97.6495	97.6500
500	Near Normal	98.0712	98.0856
	Symmetric short -tailed	97.5268	97.5986
	Symmetric long -tailed	97.5910	97.5964
	Asymmetric short-tailed	97.5402	97.6120
	Asymmetric long -tailed	97.5542	97.5556

TABLE II
Percentages difference of correctly classified data between EB and Classical method

Data distribution	Sample sizes (n)		
	100	200	500
Near Normal	0.0092	0.0219	0.0147
Symmetric short -tailed	0.0749	0.0010	0.0736
Symmetric long -tailed	0.0010	0.0015	0.0055
Asymmetric short-tailed	0.0697	0.0564	0.0736
Asymmetric long -tailed	0.0092	0.0005	0.0014

Acknowledgment

The authors gratefully acknowledge Science Achievement Scholarship of Thailand for financial support during this research.

References

[1] L. Cucala, J.-M. Marin, C.P. Robert, and D.M. Titterington, “A Bayesian reassessment of nearest-neighbour classification”, University Paris-Sud, Project Select, unpublished.

[2] T. Damoulas, and M.A. Girolami, “Combining feature spaces for classification”, Pattern Recognition, Letters 42, pp. 2671-2683, 2009.

[3] R. Johnson, and D. Wichern, “Applied multivariate statistical analysis”, Prentice – Hall, 2002.

[4] M. Aci, C. Inan, and M. Avci, “A hybrid classification method of k nearest neighbor, Bayesian methods and genetic algorithm”, Expert Systems with Applications, 37, pp. 5061-5067, 2010.

[5] C.K.I. Williams, and D. Barber, “Bayesian Classification with Gaussian Processes”, IEEE Transactions On Paitern Analysis And Machine Intelligence, VOL. 20, NO. 12, 2008.

[6] R.O. Duda, P.E. Hart, and D.G. Stork, “Pattern Classification”, John Wiley & Sons, 2001.

[7] N.A. Samsudin, and A.P. Bradley, “Nearest Neighbour group-based classification”, Pattern Recognition, Letters 43, pp. 3458-3467, 2010.

[8] M.A. Duarte-Mermoud, and N.H. Beltran, “Classification of Chilean wines, Bayesian Network”, A Practical Guide to Applications, pp. 281-299, 2008.

[9] A. Porwal, and E.J.M. Carranza, “Classifiers for modeling of mineral potential”. Bayesian Network: A Practical Guide to Applications, pp. 149-171, 2008.

[10] B.P. Carlin, and T.A. Louis, “Bayesian Methods for Data Analysis”, Chamman & Hall, 2009.

[11] T.F. Li, “Bayes empirical Bayes approach to unsupervised learning of parameters in pattern recognition”, Pattern Recognition, Letters 33, pp. 333-340, 2000.

[12] S. Chang, and T.F. Li, “Empirical Bayes decision rule for classification on defective items in Weibull distribution”, Applied Mathematics and Computation, 182, pp. 425- 433, 2006.

[13] L. Wei, and J. Chen, “Empirical Bayes estimation and its superiority for two-way”, Statistics & Probability, Letters 63, pp. 165-175, 2003.

[14] Y. Ji, K.-W. Tsui, and K. Kim, “A novel means of using gene clusters in a two-step empirical Bayes method for predicting classes of samples”, Bioinformatics, Vol. 21, No. 7, pp. 1055-1061, 2005.

[15] T. Koski, “Bayesian Predictive Classification”, School of Swedish Statistical Association, Alternative Perspectives On Statistical Inference, unpublished.

[16] R. Guo, and S. Chakraborty, “Bayesian Adaptive Nearest Neighbor”, Statistical Analysis and Data Mining, DOI:10.1002/sam, pp. 92-105, 2009.

[17] S.S. Shapiro, M.B. Wilk, and H.J. Chen, “A Comparative Study of Various Tests for Normality, Journal of the American Statistical Association”, Vol. 63, No. 324, pp. 1343-1372, 1968.

[18] J.S. Ramberg, P.R. Tadikamalla, E.J. Dudewicz, and E.F. Mykytka, “A Probability Distribution and Its Uses in Fitting Data”, Journal of the American Statistical Association, Vol. 21, No. 2, pp. 201-214, 1979.