International Journal of Pure and Applied Mathematics
Volume 81 No. 5 2012, 785-795
ISSN: 1311-8080 (printed version)
url: http://www.ijpam.eu

ijpam.eu

THE HYBRID CLASSIFICATION USING EMPIRICAL BAYES AND NEAREST NEIGHBOR

Natthinee Deetae[1], Saowanit Sukparungsee[2]
Yupaporn Areepong[3], Katechan Jampachaisri[4 §]

[1,2,3]Department of Applied Statistics
Faculty of Applied Science
King Mongkut's University of Technology
North Bangkok 10800, THAILAND
[4]Department of Mathematics
Faculty of Science
Naresuan University
Phitsanulok, 65000, THAILAND

Abstract: Classification is emphasized on allocating new observations in the test set of sample to labeled classes based on constructed rule from the training set. With the hybrid of several classification techniques has been developed and mostly exhibited results superior to a single classification technique. The aim of this study is to develop a new classification technique using Empirical Bayes in combination with Nearest Neighbor (EBNN) in the case of unknown mean and known variance. The realization of estimated hyper-parameters obtained from Empirical Bayes (EB) were adjusted using Nearest Neighbor method (NN), providing improved prediction of class membership when compared to that using single method. Data employed in this study are generated, consisting of training set and test set with the sample sizes 100, 200 and 500 for the binary classification. The results indicated EBNN method exhibited an improved performance over EB method in all situations under study.

AMS Subject Classification: 62H30, 62F15, 62C12
Key Words: classification, empirical bayes, nearest neighbor, posterior predictive probability, Markov chain Monte Carlo (MCMC)

Received: September 27, 2012

§Correspondence author

1. Introduction

Discrimination and classification are techniques usually used simultaneously. The goal of discrimination is to discover distinct attributes of observations in the training set of sample with known classes and construct a rule using information embedded in the training set, called discriminant, to separate observations as much as possible (see [12], [4], [5]). Classification is emphasized on allocating new observations in test set of sample to labeled classes based on well-defined rule obtained from the training set [12]. Generally, classification can be performed with various methods, such as Naive Bayes, Bayesian, Nearest Neighbor, Classification tree, Support vector machines and Neural network etc.

Nearest Neighbor is one of the most popular methods for classification. It is a well-known deterministic method used in supervised classification (see [4], [19]). Nearest Neighbor is performed depending on two choices: the number of the nearest neighbors, denoted as k (see [1], [4], [17]), and the form of distance function, which Euclidean distance is the most common use (see [21], [9]). Hastie and Tibshirani [10] developed a locally discriminant adaptive form of nearest neighbor classification that provided substantial improvements over standard Nearest Neighbors method in some problems. In addition, the Nearest Neighbor method may be obtained from the principle of Bayesian based on the posterior probability [17]. The Bayesian classification method, which classified observations into related classes using a decision rule, is known for its flexibility and accuracy [24]. Decision rule is defined by the posterior probability which class membership is indicated based on its highest posterior probability (see [7], [21]). Duarte-Mermoud and Beltran [6] 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 [24] studied the problem of assigning an input vector into several classes using Gaussian prior in Bayesian classification and it was generalized to multiclass problem.

The Bayesian classification involves in hyper-parameters that should be known or be able to assess using previous knowledge prior to data collection [2]. However, little information is sometimes available in practice, causing the assessment of hyper-parameters impossible. As a result, empirical Bayes can be utilized to estimate the unknown hyper-parameters using information in the observed data. Li [15] exhibited the use of Empirical Bayes to estimate unknown parameters and classify a set of unidentified input patterns into k

separate classes. In addition, the results from Monte Carlo simulation study demonstrated the favorable estimation of unknown parameters in normal distribution with Empirical Bayes. Chang and Li [3] illustrated the use of Empirical Bayes to classify a new item or product with Weibull distribution into two classes (good or defective) or identified an item as produced from one of two production lines. Wei and Chen [23] studied Empirical Bayes estimation in two-way classification model. The results showed that Empirical Bayes yielded smaller mean square error matrix than the least sum of squares method. Ji et al [11] adopted Empirical Bayes to classify gene expression profiles, leading to the decrease of number of nuisance parameters in the Bayesian model.

Recently, the hybrid of several classification techniques has been developed and mostly exhibited results superior to a single classification technique. Aci et al [1] studied the hybrid of three classification approaches; k-nearest neighbor, Bayesian and genetic algorithm, that are effective in machine learning. The hybrid approach provided better results of classification than EM algorithm. Kelly and Davis [13] studied the hybrid of genetic algorithm and the k-nearest neighbor in classification. Guo and Chakraborty [9] proposed Bayesian adaptive nearest neighbor method (BANN) which the shape of the neighborhood and the number of neighbors were chosen. BANN demonstrated substantial improvement over k-nearest neighbor and discriminant adaptive nearest neighbor (DANN) in all nine cases of the study. It also outperformed the probabilistic nearest neighbor (PNN) in most cases. Moreover, the results revealed that the performance of BANN method is the best in all cases. Ghosh and Godtliebsen [8] developed hybrid classification methods and compared their performance with parametric and nonparametric classifiers. When the underlying distributions were close to the parametric models, hybrid classifiers usually performed better than nonparametric classifiers. While the distributions were far from the assumed parametric models, hybrid classifiers performed substantially better than parametric classifiers.

There is a great attempt in the development of classification techniques. Many techniques have arisen from combinations of numerical and statistical points of view. The aim of this study is to develop a new classification technique using Empirical Bayes in combination with Nearest Neighbor(EBNN) in the case of unknown mean and known variance which has not been studied previously and it is believed to provide a good promise. The realization of estimated hyper-parameters obtained from Empirical Bayes were adjusted using Nearest Neighbor method, providing improved prediction of class membership when compared to that using single method. Data employed in this study are simulated and equally divided into two 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 under study, the percentage of correct classification is considered.

2. Methods

2.1. Empirical Bayes Method (EB)

The estimation of parameters with EB method can be classified into two types; parametric empirical Bayes and nonparametric empirical Bayes. Parametric empirical Bayes can be used when prior distribution is known, in contrast with nonparametric empirical Bayes [2]. The estimation of hyper-parameters with EB method can be obtained from posterior marginal distribution function as follow

$$m(x|\eta) = \int f(x|\theta)\pi(\theta|\eta)d\theta$$

where	θ	denotes	parameter which is continuous random variable in this case
	η	denotes	hyper-parameter
	$m(x\|\eta)$	denotes	posterior marginal distribution function.

The posterior marginal distribution function is used to estimate hyper-parameters using classical methods such as maximum likelihood method, moment method etc. Then the estimators of hyper-parameter are replaced in the posterior distribution function which can be used to estimate the parameters of interest.

2.2. Nearest Neighbor Method (NN)

The NN method is simple, quick, and often effective for classification [13]. It is a method that allocates new observations to the most common class in their neighborhood among the training set [7]. This method can be quite effective when the attributes of the data are equally important.

Given the training set $T = \{(x_i, y_j)\}; i = 1, 2, ..., n; j = 1, 2, ..., c$ where x_i is an instance and y_j is a class of label. So arbitrary instance x denotes by

$$\langle a_1(x), a_2(x), ..., a_m(x)\rangle$$

where $a_b(x)$ denotes the value of bth attribute of instance x; $b = 1, 2, ..., m$. Thus the distance between x_0, a new observation, and x_i is defined by

$$d(x_0, x_i) = \sqrt{\sum_{b=1}^{m} [a_b(x_0) - a_b(x_i)]^2}.$$

The steps of this method can be summarized as given below.

Step 1 Select the number of neighborhood (k). After some experiment a k value which gives the best result is selected.

Step 2 Define the distance measurement between a new observation and an instance.

Step 3 Sort the distances in ascending order and the minimum distance is taken.

Step 4 Classify a new observations into a class based on their neighborhood.

2.3. The EBNN Method

Based on EB method and NN method, we proposed a new classification technique; EBNN method which combines both methods together. Then, Metropolis-Hastings algorithm is adopted to construct a realization of parameters which are adjusted using NN method. Next, class membership is performed using posterior predictive probability. The steps of EBNN method are demonstrated below.

Step 1 Define the distribution function of X, $X \sim N(\theta, \sigma_0^2)$, where θ denotes unknown mean and σ_0^2 denotes known variance.

Step 2 Define the form of informative prior, $\theta \sim N(\mu, \tau^2)$.

Step 3 Compute the posterior distribution function by EB method, as follows

(3.1) Find the posterior marginal distribution function from

$$m(x|\mu, \tau^2) = \int_{-\infty}^{\infty} f(x|\theta)\pi(\theta)d\theta$$

which yields

$$x|\mu, \tau^2 \sim N(\mu, \tau^2 + \sigma_0^2)$$

(3.2) Find estimators of hyper-parameter using maximum likelihood method,

$$\hat{\mu}_{MLE} = \bar{X} \tag{1}$$

and

$$\hat{\tau}^2 = S^2 - \sigma_0^2; \quad S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2.$$

Since variance > 0,

$$\hat{\tau}^2_{MLE} = |S^2 - \sigma_0^2| \tag{2}$$

where $\hat{\mu}_{MLE}$ and $\hat{\tau}^2_{MLE}$ are the maximum likelihood estimators of μ and τ^2, respectively.

(3.3) Find posterior distribution function from

$$\pi(\theta|\underline{x}) \quad = \quad \frac{f(\underline{x}|\theta)\pi(\theta)}{\int_{-\infty}^{\infty} f(\underline{x}|\theta)\pi(\theta)d\theta}$$

Therefore, posterior distribution function is

$$\theta|\underline{X} \sim N\left(\frac{n\tau^2\bar{X} + \mu\sigma_0^2}{n\tau^2 + \sigma_0^2}, \frac{\tau^2\sigma_0^2}{n\tau^2 + \sigma_0^2}\right) \tag{3}$$

(3.4) Replaced the estimators of hyper-parameter from equation (1) and (2) in equation (3).

$$\theta|\underline{X} \sim N\left(\frac{n\hat{\tau}^2\bar{X} + \hat{\mu}\sigma_0^2}{n\hat{\tau}^2 + \sigma_0^2}, \frac{\hat{\tau}^2\sigma_0^2}{n\hat{\tau}^2 + \sigma_0^2}\right)$$

Step 4 Construct a realization of parameters from the posterior distribution function using Metropolis-Hasting techniques [18], as follows

(4.1) Initialize the chain to $\theta^{(0)}$ and set $t = 0$, where t is the number of iteration.

(4.2) Generate parameter $\theta^{(t)}$ from proposal distribution;

$$\theta^{(t)} \sim N(\hat{\mu}, \hat{\tau}^2)$$

where $t = 1, 2, ..., T$

(4.3) Compute α

$$\alpha = min\left[1, \frac{\pi(\theta^{(t)}|\underline{x})q(\theta^{(t)}|\theta^{(t-1)})}{\pi(\theta^{(t-1)}|\underline{x})q(\theta^{(t-1)}|\theta^{(t)})}\right]$$

where

$$q(\theta^{(t)}|\theta^{(t-1)}) \propto e^{-\frac{1}{2\hat{\tau}^2}(\theta^{(t)}-\hat{\mu})^2}$$
$$q(\theta^{(t-1)}|\theta^{(t)}) \propto e^{-\frac{1}{2\hat{\tau}^2}(\theta^{(t-1)}-\hat{\mu})^2}$$

(4.4) Generate U from a uniform $(0,1)$ distribution and
Accept $\theta^{(t)}$, if $u \leq \alpha$ and let $\theta^{(t+1)} = \theta^{(t)}$
Reject $\theta^{(t)}$, if $u > \alpha$ and let $\theta^{(t+1)} = \theta^{(t-1)}$

(4.5) Set $t = t + 1$ and repeat (4.2) through (4.4).

Step 5 Adjust the values of θ obtained from Step 4 using NN method according to the specified percentages, as follows

(5.1) Determine the number of nearest neighbor of θ by considering the Euclidean distance [16] between θ_i and $\bar{\theta}$, where θ_i is the ith value of θ in realization and $\bar{\theta}$ is an average value of all θs.

(5.2) Sort the distances in ascending order.

(5.3) Choose 80% of realization of θ with minimum distance and replace in the posterior distribution function.

Step 6 Compute the posterior predictive probability.

The posterior predictive probability is frequently used for prediction of new observation,y, in test data [14]. The posterior predictive probability of y conditionally on $\underline{x}$, denoted by

$$p(y|\underline{x},\theta) = \int_\theta f(y|\theta)\pi(\theta|\underline{x})d\theta$$

where

$$f(y|\theta) = \frac{1}{\sqrt{2\pi\sigma_0^2}}e^{-\frac{1}{2\sigma_0^2}(y-\theta)^2}$$

and

$$\pi(\theta|\underline{x}) = \frac{e^{-\frac{1}{2\tau^2\sigma_0^2}\left[(n\tau^2+\sigma_0^2)\left(\theta-\frac{n\tau^2\bar{x}+\mu\sigma_0^2}{n\tau^2+\sigma_0^2}\right)^2\right]}}{\left[2\pi\left(\frac{\tau^2\sigma_0^2}{n\tau^2+\sigma_0^2}\right)\right]^{1/2}}$$

Sometimes the posterior predictive probability is not tractable, so MCMC technique [9] can be applied to estimate $p(y|\underline{x},\theta)$ as

$$\hat{p}(y|\underline{x},\theta) \quad \approx \tfrac{1}{M}\textstyle\sum_{t=1}^{M} p(y|\underline{x},\theta^{(t)})$$

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

Step 7 Classify the test data into classes based on the highest posterior predictive probability.

Step 8 Compute the percentages of correct classification.

3. Simulation Study

Data employed in this study are generated, consisting of training set and test set with the sample sizes 100, 200 and 500 for the binary classification. The data are distributed as normal and constructed as short-tailed and long-tailed symmetric based on Shapiro, Wilk and Chen criteria [22], with the skewness of 0.0 and kurtosis of 2.4 and 6.0. The skewness of data is constructed using data transformation according to Ramberg and Tadikamalla [20] known as Generalized Lambda Distribution (GLD).

4. Results and Discussion

The simulation results indicated the percentages of correctly classified data in the case of unknown mean and known variance using EB method and EBNN method, as shown in Table 1 and Table 2, respestively. The percentage of relative efficiency of EBNN method to EB method is shown in Table 3.

With 3 levels of sample sizes, EB method showed good classification when data were distributed as normal, while EBNN method illustrated good classification when data had constructed as short-tailed symmetric. All the results were shown in Table 1 and 2. In addition, EBNN method indicated an improved performance over EB method in all situations under study, as displayed in Table 3.

Sample sizes (n)	EB method (%)		
	Long-Tailed symmetric	Short-Tailed symmetric	Normal
100	96.3180	96.6220	96.7880
200	96.3015	96.6790	96.7870
500	96.2978	96.6836	96.7794

Table 1: Percentages of correctly classified data using EB method

Sample sizes (n)	EBNN method (%)		
	Long-Tailed symmetric	Short-Tailed symmetric	Normal
100	96.5340	97.3030	97.2830
200	96.5275	97.3950	97.2610
500	96.5080	97.3726	97.2718

Table 2: Percentages of correctly classified data using EBNN method

Sample sizes (n)	Relative efficiency (%)		
	Long-Tailed symmetric	Short-Tailed symmetric	Normal
100	1.0022	1.0070	1.0051
200	1.0023	1.0074	1.0049
500	1.0022	1.0071	1.0051

Table 3: Percentages of relative efficiency classified of EBNN method to EB method

Acknowledgments

The authors would like to thank for the Graduate Colleges, King Mongkut's University of Technology North Bangkok for the Financial support during this research.

References

[1] M. Aci, C. Inan, M. Avci, A hybrid classification method of k nearest neighbor, Bayesian methods and genetic algorithm, *Expert Systems with Applications*, **37** (2010), 5061-5067.

[2] B.P. Carlin, T.A. Louis, *Bayesian Methods for Data Analysis*, Chamman & Hall (2009).

[3] S. Chang, T.F. Li, Empirical Bayes decision rule for classification on defective items in Weibull distribution, *Applied Mathematics and Computation*, **182** (2006), 425- 433.

[4] L. Cucala, J.-M. Marin, C.P. Robert, D.M. Titterington, A Bayesian reassessment of nearest-neighbour classification, University Paris-Sud, Project Select (2008).

[5] T. Damoulas, M.A. Girolami, Combining feature spaces for classification, *Pattern Recognition*, Letters **42** (2009), 2671-2683.

[6] M.A. Duarte-Mermoud, N.H. Beltran, Classification of Chilean wines, *Bayesian Network: A Practical Guide to Applications* (2008).

[7] R.O. Duda, P.E. Hart, D.G. Stork, *Pattern Classification*, John Wiley & Sons (2001).

[8] A.K. Ghosh, F. Godtliebsen, On hybrid classification using model assisted posterior estimates, *Pattern Recognition* (2011).

[9] R. Guo, S. Chakraborty, Bayesian Adaptive Nearest Neighbor, *Statistical Analysis and Data Mining*, DOI:10.1002/sam (2009).

[10] T. Hastie, R. Tibshirani, Discriminant Adaptive Nearest Neighbor Classification, *IEEE Transactions on Paitern Analysis and Machine Intelligence*, **18**, No. 6 (1996).

[11] Y. Ji, K.-W. Tsui, 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 (2005), pages 1055-1061.

[12] R. Johnson, D. Wichern, *Applied multivariate statistical analysis*, Prentice-Hall (2002).

[13] J.D. Kelly, L. Davis, A Hybrid Genetic Algorithm for Classification, *Learning and Knowledge Acquisition* (1991).

[14] T. Koski, Bayesian Predictive Classification, School of Swedish Statistical Association, Alternative Perspectives On Statistical Inference (2005).

[15] T.F. Li, Bayes empirical Bayes approach to unsupervised learning of parameters in pattern recognition, *Pattern Recognition*, Letters **33** (2000), 333-340.

[16] D.T. Larose, *Discovering knowledge in data*, John Wiley & Sons (2005).

[17] S. Manocha, A.M. Girolami, An empirical analysis of the probabilistic K-nearest neighbour classifier, *Pattern Recognition*, Letters **28** (2007), 1818-1824.

[18] W.L. Martinez, A.R. Martinez, *Computational Statistics Handbook with MATLAB*, Chamman & Hall (2002).

[19] P. Piro, R. Nock, F. Nielsen, M. Barlaud, Leveraging k-NN for generic classification boosting, *Neurocomputing*, **88** (2012), 3-9.

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

[21] N.A. Samsudin, A.P. Bradley, Nearest Neighbour group-based classification, *Pattern Recognition*, Letters **43** (2010), 3458-3467.

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

[23] L. Wei, J. Chen, Empirical Bayes estimation and its superiority for two-way, *Statistics & Probability*, Letters **63** (2003), 165-175.

[24] C.K.I. Williams, D. Barber, Bayesian classification with Gaussian processes, *IEEE Transactions on Paitern Analysis and Machine Intelligence*, **20**, No. 12 (2008).