

0. 12/12/12 (2/

0.25

**PROCEEDINGS
THE 4TH INTERNATIONAL CONFERENCE ON SCIENCE &
TECHNOLOGI 2012 (ICSTIE2012)
Application in Industry and Education**

GREENER TECHNOLOGY FOR A HEALTHY ENVIRONMENT

The Editors

Associate Professor Sarina Md Jam

Ch'ng Pei Eng

Cheng Siak Peng

Ng Set Foong

Norshuhada Samsudin

Rozita Kadar

Rafizah Kechil

Wan Nur Shaziayani Wan Mohd Rosly

e-ISBN : 978-983-42204-9-5

K-nearest neighbour and Empirical Bayes for Predictive Classification

Natthinee Deetae¹, Saowanit Sukparungsee², Yupaporn Areepong³ & Katechan Jampachaisri⁴

^{1,2,3}Department of Applied Statistics, Faculty of Applied Science, King Mongkut's University of Technology
North Bangkok, Bangkok, Thailand.

⁴Department of Mathematics, Faculty of Science, Naresuan University, Phitsanulok, Thailand.

¹natthineed@hotmail.com, ²swns@kmutnb.ac.th, ³yupaporn@kmutnb.ac.th,
⁴katechanj@nu.ac.th

ABSTRACT

Classification is emphasized on allocating new observations in the test sets of sample to labelled classes based on constructed rule from the training set. With parametric classification, data are assumed to be normally distributed which rarely occurred in practice. The aim of this study is to develop a classification technique to classify various types of data; normal, near normal, asymmetric short-tailed and asymmetric long-tailed, using empirical Bayes method (EB) and K-nearest neighbour method (K-NN). The estimated hyper-parameters obtained from empirical Bayes are substituted in the posterior predictive probability and used to classify new observations and then compared to K-NN method. Data employed in this study are generated, consisting of training set and test sets with the sample size 100, 200 and 500 for the binary classification. In each situation under study, the percentage of correct classification is considered. With normal data, the EB method produces higher percentages of correct classification than K-NN method for all sample sizes, while K-NN method mostly produces higher percentages of correct classification than EB method when data are distributed as near normal, asymmetric short-tailed and asymmetric long-tailed. In addition, the percentage of correctly classified data tends to be higher as the number of neighbourhood (k) increases.

Keywords: Empirical Bayes, K-nearest neighbour, Classification, The posterior predictive probability, Markov chain Monte Carlo

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 (Cucala et al. 2008; Damoulas & Girolami 2009). Classification is emphasized on allocating new observations in test sets of sample to labeled classes based on well-defined rule obtained from the training set (Johnson & Wichern 2002). Generally, classification methods can be performed in parametric and non-parametric. With parametric classification, data are assumed to be normally distributed which rarely occurred in practice. As a result, classification usually carries out using non-parametric method such as the support vector machines, neural network, K-nearest neighbour, including Bayesian method.

K-nearest neighbour is one of the most popular methods for classification. This method is simple, quick, and often effective. It is a well-known deterministic method used in supervised classification (Cucala et al. 2008; Piro et al. 2012). K-nearest neighbour is performed depending on two choices: the number of the nearest neighbours, denoted as k (Aci et al. 2010; Cucala et al. 2008; Manocha & Girolami 2007), and the form of distance function, which Euclidean distance is the most common used (Samsudin & Bradley 2010; Guo & Chakraborty 2009). Hastie and Tibshirani (1996) developed a locally discriminant adaptive form of nearest neighbour classification that provided substantial improvements over standard nearest neighbours method in some problems. Guo and Chakraborty (2009) proposed Bayesian adaptive nearest neighbour method which the shape of the neighborhood and the number of neighbours were chosen. In addition the K-nearest neighbour method may be obtained from the principle of Bayesian based on the posterior probability (Manocha & Girolami 2007). The Bayesian classification method, which classified observations into related classes using a decision rule, is known for its flexibility and accuracy (Williams & Barber 2008). Decision rule is defined by the posterior probability which class membership is

indicated based on its highest posterior probability (Duda et al. 2001; Samsudin & Bradley 2010). Duarte-Mermoud and Beltran (2008) 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 (2008) 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 (2008) had proposed classifiers for modeling of mineral potential. The aim was to explain and display applications of Bayesian networks as another tool classifying potential for mineral deposit occurrence. This result found that Bayesian network classifiers are efficient tools for modeling of mineral potential mapping.

The Bayesian classification involves in hyper-parameters that should be known or be able to assess using previous knowledge prior to data collection (Carlin & Louis 2009). 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 (2000) exhibited the use of Empirical Bayes 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 Empirical Bayes. Chang and Li (2006) illustrated the use of Empirical Bayes 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 (2003) 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. (2005) adopted Empirical Bayes to classify gene expression profiles, leading to the decrease of number of nuisance parameters in the Bayesian model.

Non-parametric based classification method does not require assumption on the distribution of population which is rather motivated and attractive. The aim of this study is to develop a classification technique to classify various types of data; normal, near normal, asymmetric short-tailed and asymmetric long-tailed, using empirical Bayes and K-nearest neighbour method. The estimated hyper-parameters obtained from empirical Bayes are substituted in the posterior predictive probability and used to classify new observations and then compared to K-nearest neighbour method. Data employed in this study are simulated and 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. K-nearest neighbour method (K-NN)

The method of K-NN, an important approach to non-parametric classification, is simple, quick, and often effective (Kelly & Davis 1991). It is a method that allocates new observations to the most common class in their neighborhood among the training set (Duda et al. 2001). This method can be quite effective when the attributes of the data are equally important.

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

$$\langle a_1(x), a_2(x), \dots, a_m(x) \rangle \quad (1)$$

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

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

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

- Step I Select the number of neighborhood (k).
- Step II Define the distance measurement between a new observation and an instance.
- Step III Sort the distances in ascending order and the minimum distance is taken.
- Step IV Classify a new observations into a class based on their neighborhood.

3. Empirical Bayes method (EB)

The estimation of parameters with EB method can be classified into two types; parametric EB and nonparametric EB. Parametric EB can be used when prior distribution is known, in contrast with nonparametric EB (Carlin & Louis 2009). The estimation of hyper-parameters with EB method can be obtained from posterior marginal distribution function as follow.

$$m(x|\eta) = \int f(x|\theta)\pi(\theta|\eta)d\theta \quad (3)$$

Where θ is parameter which is continuous random variable in this case
 η is hyper-parameter

$m(x|\eta)$ is posterior marginal distribution function.

In this study, conjugate prior is assumed with unknown mean (θ) and known variance (σ_0^2).

Informative prior: $\theta \sim N(\mu, \tau^2)$ that is,

$$\pi(\theta) = \frac{1}{\sqrt{2\pi\tau^2}} e^{-\frac{1}{2\tau^2}(\theta-\mu)^2} \quad (4)$$

The steps of EB method are demonstrated below:

Step I: Find posterior marginal distribution function.

Consider probability distribution function of X .

$$f(x|\theta) = \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{1}{2\sigma_0^2}(x-\theta)^2}, -\infty < x < \infty, -\infty < \theta < \infty, \sigma_0^2 > 0 \quad (5)$$

Posterior marginal distribution function is

$$m(x|\mu, \tau^2) = \int_{-\infty}^{\infty} f(x|\theta)\pi(\theta)d\theta \quad (6)$$

Therefore, posterior marginal distribution function of X is

$$x|\mu, \tau^2 \sim N(\mu, \tau^2 + \sigma_0^2) \quad (7)$$

Step II: Find estimator of hyper-parameter by maximum likelihood method,

$$\hat{\mu}_{MLE} = \bar{X} \quad (8)$$

and

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

Since variance > 0 ;

$$\hat{\tau}_{MLE}^2 = |S^2 - \sigma_0^2| \quad (9)$$

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

Step III: Find the posterior distribution function.

$$\pi(\theta|\underline{x}) = \frac{f(\underline{x}|\theta)\pi(\theta)}{\int_{-\infty}^{\infty} f(\underline{x}|\theta)\pi(\theta)d\theta} \quad (10)$$

So the posterior distribution function is

$$\theta|\underline{X} \sim N\left[\frac{n\tau^2\bar{X} + \mu\sigma_0^2}{n\tau^2 + \sigma_0^2}, \frac{\tau^2\sigma_0^2}{n\tau^2 + \sigma_0^2}\right] \quad (11)$$

Step IV: Replaced the estimators of hyper-parameter from Eq. (8) and Eq. (9)

in Eq. (11).

$$\theta | \underline{X} \sim N \left[\frac{n\hat{\tau}^2 \bar{X} + \hat{\mu}\sigma_0^2}{n\hat{\tau}^2 + \sigma_0^2}, \frac{\hat{\tau}^2 \sigma_0^2}{n\hat{\tau}^2 + \sigma_0^2} \right] \quad (12)$$

Step V: Compute the posterior predictive probability.

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

$$p(y | \underline{x}, \theta) = \int_{\theta} f(y | \theta) \pi(\theta | \underline{x}) d\theta \quad (13)$$

where

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

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

Sometimes the posterior predictive probability is not a tractable, so Markov Chain Monte Carlo (MCMC) technique (Guo & Chakraborty 2009) can be applied estimate $p(y | \underline{x}, \theta)$ as

$$\hat{p}(y | \underline{x}, \theta) \approx \frac{1}{M} \sum_{t=1}^M p(y | \underline{x}, \theta^{(t)}) \quad (14)$$

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

4. Simulation Results

Three characteristics of data are constructed by varying the values of skewness and kurtosis based on Shapiro (Shapiro et al. 1968) and Ramberg (Ramberg et al. 1979) criteria as (skewness, kurtosis); (0.25, 2.80), (0.75, 2.80) and (0.75, 6.00) for near normal, asymmetric short-tailed and asymmetric long-tailed, respectively. The simulation results indicated the percentages of correctly classified data in the case of unknown mean and known variance using EB and K-NN method when data are distributed as normal (Table 1), near normal, asymmetric short-tailed and asymmetric long-tailed (Table 2).

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. When data are distributed as normal, EB method yielded the highest correctly classified data for all sample sizes. While K-NN method is proposed by varying k ($k=3, 5, 7$) which $k=7$ produced the highest correctly classified data for all three levels of k . All the results are shown in Table 1.

TABLE 1: Percentages of correctly classified data when data are distributed as normal

Methods	Sample sizes (n)		
	100	200	500
EB	98.6667	98.0833	97.4667
3-NN	97.5000	97.5833	96.8333
5-NN	97.6667	97.6667	97.0667
7-NN	98.0000	97.9167	97.2000

TABLE 2: Percentages of correctly classified data when data are constructed as near normal, asymmetric short-tailed and asymmetric long-tailed.

Skewed distribution	Method	Sample sizes (n)		
		100	200	500
Near Normal	EB	96.8333	96.7500	97.0900
	3-NN	97.0000	96.8333	96.8800
	5-NN	97.1667	96.9167	97.1200
	7-NN	97.5000	96.9167	97.2600
Asymmetric long-tailed	EB	97.1667	97.9167	97.6000
	3-NN	97.1667	98.2500	97.6333
	5-NN	97.3333	98.1667	97.6333
	7-NN	97.8333	98.1667	97.6333
Asymmetric short-tailed	EB	98.6667	98.8333	99.3333
	3-NN	99.6667	99.4167	99.6333
	5-NN	99.8333	99.5000	99.7333
	7-NN	99.8333	99.5000	99.7333

Table 2 shows the percentages of correctly classified data when data are constructed as near normal, asymmetric short-tailed and asymmetric long-tailed. The results revealed that the 7-NN method yielded the highest correctly classified percentages for all sample sizes in near normal data. For asymmetric long-tailed data, the 7-NN and 3-NN yielded the highest correctly classified percentages as $n=100$ and $n=200$ respectively, while all 3 levels of k -NN yielded similar correctly classified percentages as $n=500$. For asymmetric short-tailed data the 5-NN and 7-NN method resulted in the highest correctly classified percentages in all sample sizes.

5. Conclusions

Simulation results suggest that, EB method outperforms K-NN method for all sample sizes in normally distributed data, whereas K-NN method performs better than EB method in the cases of near normal, asymmetric short-tailed and asymmetric long-tailed. In addition, the percentages of correctly classified data tend to be higher as the number of neighborhoods (k) increases.

Acknowledgement

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

- L. Cucala, J.-M. Marin, C.P. Robert, & D.M. Titterton. A Bayesian reassessment of nearest-neighbour classification. University Paris-Sud, Project Select, unpublished.
- Damoulas, T. & Girolami, M.A. 2009. Combining feature spaces for classification. *Pattern Recognition, Letters* 42, pp. 2671-2683.
- R. Johnson, & D. Wichern. 2002. Applied multivariate statistical analysis. Prentice – Hall.
- Piro, P., Nock, R., Nielsen, F. & Barlaud, M. 2012. Leveraging k-NN for generic classification boosting. *Neurocomputing*, 88 (2012), pp. 3-9.

- Aci, M., Inan, C. & Avci, M. 2010. A hybrid classification method of k nearest neighbor, Bayesian methods and genetic algorithm. *Expert Systems with Applications*, 37, pp. 5061-5067.
- Manocha, S. & Girolami, A.M. 2007. An empirical analysis of the probabilistic K-nearest neighbour classifier. *Pattern Recognition, Letters* 28, pp. 1818-1824.
- Samsudin, N.A. & Bradley, A.P. 2010. Nearest Neighbour group-based classification. *Pattern Recognition, Letters* 43, pp. 3458-3467.
- Guo, R. & Chakraborty, S. 2009. Bayesian Adaptive Nearest Neighbor. *Statistical Analysis and Data Mining*, DOI:10.1002/sam, pp. 92-105.
- Hastie, T. & Tibshirani, R. 1996. Discriminant Adaptive Nearest Neighbor Classification, *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, VOL. 18, No. 6.
- Williams, C.K.I. & Barber, D. 2008. Bayesian Classification With Gaussian Processes. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, VOL. 20, NO. 12.
- R.O. Duda, P.E. Hart, & D.G. Stork. 2001. Pattern Classification. John Wiley & Sons.
- Duarte-Mermoud, M.A. & Beltran, N.H. 2008. Classification of Chilean wines. *Bayesian Network, A Practical Guide to Applications*, pp. 281-299.
- Porwal, A. & Carranza, E.J.M. 2008. Classifiers for modeling of mineral potential. *Bayesian Network: A Practical Guide to Applications*, pp. 149-171.
- B.P. Carlin, & T.A. Louis. 2009. Bayesian Methods for Data Analysis. Chamman & Hall.
- Li, T.F. 2000. Bayes empirical Bayes approach to unsupervised learning of parameters in pattern recognition. *Pattern Recognition, Letters* 33, pp. 333-340.
- Chang, S. & Li, T.F. 2006. Empirical Bayes decision rule for classification on defective items in Weibull distribution. *Applied Mathematics and Computation*, 182, pp. 425- 433.
- Wei, L. & Chen, J. 2003. Empirical Bayes estimation and its superiority for two-way. *Statistics & Probability, Letters* 63, pp. 165-175.
- Ji, Y., Tsui, K.-W. & Kim, K. 2005. 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.
- Kelly, J.D. & Davis, L. 1991. A Hybrid Genetic Algorithm for Classification. *Learning and Knowledge Acquisition*.
- T. Koski. Bayesian Predictive Classification. School of Swedish Statistical Association, Alternative Perspectives On Statistical Inference, unpublished.
- Shapiro, S.S., Wilk, M.B. & Chen, H.J. 1968. A Comparative Study of Various Tests for Normality. *Journal of the American Statistical Association*, Vol. 63, No. 324, pp. 1343-1372.
- Ramberg, J.S., Tadikamalla, P.R., Dudewicz, E.J. & Mykytka, E.F. 1979. A Probability Distribution and Its Uses in Fitting Data. *Journal of the American Statistical Association*, Vol. 21, No. 2, pp. 201-214.