IMPROVED BAYESIAN FEATURE SELECTION AND CLASSIFICATION METHODS USING BOOTSTRAP PRIOR TECHNIQUES

O. R. Olaniran [1], S. F. Olaniran [2], W. B. Yahya [1], A. W. Banjoko [1], M. K. Garba [1], L. B. Amusa [1], N. F. Gatta [1]

[1] Department of Statistics, University of Ilorin, PMB 1515, Ilorin, Nigeria
[2] Department of Statistics and Mathematical Sciences, Kwara State University, Malete, P.M.B 1530, Ilorin, Nigeria

Corresponding author: O. R. Olaniran, rid4stat@yahoo.com; olaniran.or@unilorin.edu.ng

ABSTRACT: In this paper, the behavior of feature selection algorithms using the traditional t-test, Bayesian t-test using MCMC and Bayesian two-sample test using proposed bootstrap prior technique were determined. In addition, we considered some frequentist classification methods like k- Nearest Neighbor (k-NN), Logistic Discriminant (LD), Linear discriminant analysis (LDA), Quadratic discriminant analysis (QDA) and Naïve Bayes when conditional independence assumption is violated. Two new Bayesian classifiers (B-LDA and B-QDA) were developed within the frame work of LDA and QDA using the bootstrap prior technique. The model parameters were estimated using Bayesian approach via the posterior distribution that involves normalizing the prior for the attributes and the likelihood from the sample in a Monte-Carlo experiment. The bootstrap prior technique was incorporated into the Normal-Inverse-Wishart natural conjugate prior for the parameters of the multivariate normal distribution where the scale and location parameters were required. All the classifiers were implemented on the simulated data at 90:10 training-test data ratio. The efficiencies of these classifiers were assessed using the misclassification error rate, sensitivity, specificity, positive predictive value, negative predictive value and area under the ROC curve. Results from various analyses established the supremacy of the proposed Bayes classifiers (B-LDA and B-QDA) over the existing frequentists and Naïve Bayes classification methods considered. All these methods including the proposed one were implemented on a published binary response microarray data set to validate the results from the simulation study.

KEYWORDS: *k*-Nearest Neighbour, Bayesian Linear Discriminant Analysis, Bayesian Quadratic Discriminant Analysis, Naïve Bayes, Bootstrap prior.

1. INTRODUCTION

Simultaneous feature selection and classification are two main problems in data mining. The two problems are related in the sense that, when good features are identified the classification scheme or method often yields high accuracy. This implies the performance of a feature selection algorithm can be assessed using the accuracy of resulting classification method that utilizes the identified subset features. Several studies have considered various feature selection algorithm or classifiers (classification method) that simultaneously select and classify clinical or non-clinical classification problems. In the recent time, Yahya ([Yah12]) considered feature selection algorithm using AUC (Area under the ROC Curve).
In this paper, we proposed a novel feature selection algorithm and two classification methods under the frame work of Bayesian discriminant analysis.

2. FEATURE SELECTION

Feature selection process is the removal of features that are irrelevant with respect to the output variable from the data set. Feature selection can be extremely useful in reducing the dimensionality of the data to be processed by the classifier, reducing execution time and improving predictive accuracy (inclusion of irrelevant features can introduce noise into the data, thus obscuring relevant features). Many feature selection methods have been proposed from the Bayesian and frequentist perspectives. Within the frequentist; p-value method (Golub [G+99]), Wilcoxon-Mann-Whitney rank sum test (Thomas [T+01]), the student-t test or its equivalent; the welch test (Nguyen & Rocke [NR02], Yahya et al. [Y+11]). Similarly within Bayesian; Cyber-T a Bayesian version of t statistic by (Baldi and Long [BL01]), among other methods based on t or ANOVA are ([WS03]), ([N+01]), ([Smy04]).

3. CLASSIFICATION

Predicting a qualitative response for an observation can be referred to as classifying that observation, since it involves assigning the observation to a category, or class. On the other hand, often the methods used for classification first predict the probability of each of the categories of a qualitative variable, as the basis for making the classification (Hastie et al. [H+13]). Classification can also viewed