Tassanee Nacharoen 2014: Discriminant Analysis for Class - Imbalanced Data :
A Case Study of Diabetes Risk Group. Master of Science (Statistics),
Major Field: Statistics, Department of Statistics. Thesis Advisor:
Assistant Professor Lily Ingsrisawang, Ph.D. 99 pages.

This study aimed to compare the classification performance between parametric discriminant analysis and nonparametric discriminant analysis in a three-group classification application of class-imbalanced data. Data from a healthy project for 599 staffs in a government hospital in Bangkok were obtained for the classification problem. The staffs were diagnosed into one of three diabetes risk groups: non-risk (90%), risk (5%), and diabetic (5%). The original data including these variables of diabetes risk group, age, gender, cholesterol, and BMI were analyzed and bootstrapped up to 50 samples, 599 observations per sample, for additional estimation of misclassification error rate. The 50 bootstrap samples consisted of two patterns of three diabetes risk groups of non-risk: risk: diabetic as 90%: 5%: 5% and 80%: 10%: 10%. Each data set was explored for the departure of multivariate normality and the equality of covariance matrices of the three risk groups. Both the original data and the bootstrap samples show nonnormality and unequal covariance matrices. The parametric linear discriminant function, quadratic discriminant function, and the nonparametric k-nearest neighbors' discriminant function were performed over 50 bootstrap samples and applied to the original data. In finding the optimal classification rule, the choices of prior probabilities were set up for both equal proportions and unequal proportions. The results indicated that the k-nearest neighbors approach when k = 3 or k = 4 and the prior probabilities of non-risk: risk: diabetic as 0.90: 0.05:0.05 or 0.80: 0.10: 0.10 gave the smallest error rate of misclassification.

/