International Journal of Statistical Sciences *ISSN 1683–5603*
Vol. 9(Special Issue), 2009, pp 199-208

Nonparametric Classification Based on Ranks

Mayer Alvo and Danielle Leger
Department of Mathematics and Statistics
University of Ottawa
Ottawa, ON K1N 6N5

[Received December 8, 2008; Revised January 8, 2009; Accepted February 17, 2009]

Abstract

A new nonparametric classification method based on the ranks of the observations is proposed. The probabilities of error are computed through various simulation studies. The method is shown to have some good properties.

Keywords and Phrases: Nonparametric Classification, Ranks, Bayes Rule.

AMS Classification: 62G99.

1 Introduction

In the two sample classification problem, it is desired to classify a new object into one of two classes, labelled π_1, π_2. The objects are classified on the basis of measurements on p random variables $X' = (X_1, ..., X_p)$. We shall assume that the population of X values differs from one class to another and that the populations can be described by probability density functions $f_1(x), f_2(x)$ and corresponding cumulative distribution functions $F_1(x), F_2(x)$respectively. Let A_i be the set of x values for which an object is classified as belonging to $\pi_i, i = 1, 2$. The conditional probability $P(2|1)$, defined to be the probability of classifying an object into π_2 given it belongs to π_1, is

$$P(2|1) = \int_{A_2} f_1(x)\, dx \tag{1.1}$$

Similarly, the conditional probability $P(1|2)$, defined as the probability of classifying an object into π_1 given it belongs to π_2, is given by

$$P\left(1|2\right) = \int_{A_1} f_2\left(x\right) dx \tag{1.2}$$

Let $c\left(2|1\right), c\left(1|2\right)$ be the corresponding costs of misclassification and let p_1, p_2 be the prior probabilities of π_1, π_2 respectively. It follows that the expected cost of misclassification

$$c\left(2|1\right) P\left(2|1\right) p_1 + c\left(1|2\right) P\left(1|2\right) p_2 \tag{1.3}$$

is minimized by choosing regions A_1, A_2 defined as

$$A_1 = \left\{ x | \ \frac{f_1\left(x\right)}{f_2\left(x\right)} \geq k \right\} \tag{1.4}$$

$$A_2 = \left\{ x | \ \frac{f_1\left(x\right)}{f_2\left(x\right)} < k \right\} \tag{1.5}$$

where $k = \frac{c(1|2)p_2}{c(2|1)p_1}$. When the misclassification costs are equal and the prior probabilities are equal, $k = 1$. Alternative criteria described in Johnson and Wichern (2002, p.503) may be used to arrive at a classification rule but these generally lead to variants of (1.4). As an example, suppose we are provided with training samples from each of two normal populations of dimension p, having common but unknown variance-covariance function. Let $\bar{x}_1, \bar{x}_2$ be the respective sample means and let S be the pooled estimate of the variance covariance matrix. Then, the classification rule which minimizes the expected cost of misclassification allocates a new observation x_0 to π_1 if and only if

$$\left(\bar{x}_1 - \bar{x}_2\right)' S^{-1} x_0 - \frac{1}{2}\left(\bar{x}_1 - \bar{x}_2\right)' S^{-1}\left(\bar{x}_1 + \bar{x}_2\right) \geq \ln k \tag{1.6}$$

Let $y = \hat{l}'x = \left(\bar{x}_1 - \bar{x}_2\right)' S^{-1} x, \bar{y}_i = \hat{l}'\bar{x}_i \ , i = 1, 2$ and define

$$\begin{aligned} \hat{m} &= \frac{1}{2}\left(\bar{x}_1 - \bar{x}_2\right)' S^{-1}\left(\bar{x}_1 + \bar{x}_2\right) && (1.7) \\ &= \frac{1}{2}\left(\bar{y}_1 + \bar{y}_2\right) && (1.8) \end{aligned}$$

Suppose that $k = 1$. We may then view the classification rule in (1.6) as classifying a new observation x_0 to π_1 if and only if $y_0 \geq \frac{1}{2}\left(\bar{y}_1 + \bar{y}_2\right)$.

Fisher arrived at the classification rule above by showing that the linear combination y maximizes the ratio $\frac{\left(\bar{y}_1 - \bar{y}_2\right)^2}{s_y^2}$, where $s_y^2 = \hat{l}S\hat{l}$, is the pooled sample variance of the $y's$. Fisher's idea was to consider linear combinations of the $x's$ since they are simpler functions of the data. He did not assume any distributional form for the data.

The classification rule in (1.4) requires knowledge of the underlying densities which characterize the populations. On the other hand, Fisher's rule, even though it is derived in a nonparametric manner, is a function of the means which may be unduly influenced by outliers. In section 3 , we propose a new nonparametric classification procedure based on the ranks of the observations and hence is less influenced by outliers. In section 4 we extend those results to the multivariate case. In section 5, we report on the results of a simulation study. It is shown that the error probabilities are small even under heavy tailed distribution such as Cauchy distribution.

2 Classification for two univariate populations

We begin by considering the univariate case, $p = 1$ and assume that the populations are continuous. Further assume that these populations differ in location so that the observations from one population tend to be larger than those of the other population. Let $X_1' = (X_{11}, ..., X_{1n_1})$ and $X_2' = (X_{21}, ..., X_{2n_2})$ be independent random samples from π_1, π_2 respectively. Let X_0 be a new random variable independent of X_1, X_2. Let $(X_{11}, ..., X_{1n_1}|X_0|X_{21}, ..., X_{2n_2})$ represent the pooled sample. Set $n = n_1 + n_2 + 1$. We shall assume that π_1 lies to the left of π_2.

Motivated by Alvo (2008), we consider the following general approach for classification based on the ranks of the observations. It consists of defining a set of permutations induced by the observations and two sets of permutations "most in agreement" with the new observation belonging to π_1 or to π_2 respectively. The test statistic is then based on a measure of the distance between these two sets. Specifically, we propose the following steps.

Let $\mathcal{P} = \{\mu : [\mu(1), ..., \mu(n)]\}$ be the set of all permutations of the integers $1, 2, ..., n$, and let $d(\mu, \nu)$ be a distance function between permutations μ and ν.

Step 1: Rank all the observations together so that the smallest gets rank 1, the next smallest rank 2 etc.. Let the $n-$dimensional vector

$$\pi = (\pi_1(1), ..., \pi_1(n_1)|\pi_0|\pi_2(1), ..., \pi_2(n_2)) \tag{2.1}$$

represent the ranks of X_0 and of $\{X_{il}\}, l = 1, ..., n_i, i = 1, 2,$ grouped by population. In view of the continuity assumption on the distributions, ties among the observations occur with probability zero.

Step 2: Define $\{\pi\}$ to be the subclass of permutations "equivalent" to the observable permutation π in the sense that ranks occupied by identically distributed random variables are exchangeable. This subclass having cardinality given by the product $(n_1!n_2!)$, consists of all the permutations where the rankings within each population are permuted among themselves only. The set $\{\pi\}$ accounts for sampling variation within each population.

Step 3: Define E_1, E_2 to be subclasses of $\mathcal{P}$ consisting of all permutations which are "most in agreement" with the new observation belonging to π_1, π_2 respectively. The extremal sets E_1, E_2 do not necessarily correspond to the entire critical region but rather consist of those permutations which provide the strongest evidence in favour of the classification in either π_1

or π_2 respectively. In the present context, we shall assume that permutations in E_1 are such that ranks occupied by observations from π_1 are always less than those from π_2. In practice, the direction can be determined from a histogram of values for the training samples. The cardinality of E_1 is equal to $(n_1+1)!n_2!$ whereas that of E_2 is $n_1!\,(n_2+1)!$

Step 4: For a given distance function $d\left(\mu,\nu\right)$, between two permutations $\mu,\nu,$ define the distance between the two sets $\{\pi\}$ and E *by computing the sum of all pairwise distances between them:*

$$d\left(\{\pi\},E\right)=\sum_{\mu\epsilon\{\pi\}}\sum_{\nu\epsilon E}d\left(\mu,\nu\right) \tag{2.2}$$

Step 5: The classification rule then classifies X_0 into π_1 if and only if

$$d\left(\{\pi\},E_1\right)\leq d\left(\{\pi\},E_2\right) \tag{2.3}$$

In what follows, we shall consider the Spearman distance between permutations defined by

$$\begin{aligned} d\left(\mu,\nu\right) &= \frac{1}{2}\sum_{i=1}^{n}\left[\mu\left(i\right)-\nu\left(i\right)\right]^2 \\ &= c-\sum_{i=1}^{n}\left[\mu\left(i\right)-\frac{n+1}{2}\right]\left[\nu\left(i\right)-\frac{n+1}{2}\right] \end{aligned} \tag{2.4}$$

where $c=\sum_{i=1}^{n}\left(i-\frac{n+1}{2}\right)^2=\frac{n\left(n^2-1\right)}{12}$.

Since the observations from π_1 are assumed to be smaller than those from π_2, the set E_1 consists of permutations of the type

$$(1,2,...,n_1+1|n_1+2,...,n+1)$$

where the first n_1+1 integers are permuted among themselves and the next n_2 integers are permuted among themselves. Hence,

$$\begin{aligned} d\left(\{\pi\},E_1\right) &= \sum_{\mu\epsilon\{\pi\}}\sum_{\nu\epsilon E_1}d\left(\mu,\nu\right) \\ &= \sum_{\mu\epsilon\{\pi\}}\sum_{\nu\epsilon E}c-\sum_{i=1}^{n}\sum_{\mu\epsilon\{\pi\}}\sum_{\nu\epsilon E}\left[\mu\left(i\right)-\frac{n+1}{2}\right]\left[\nu\left(i\right)-\frac{n+1}{2}\right] \\ &= \left(n_1!n_2!\right)^2\left(n_1+1\right)c \\ &\quad -\sum_{i=1}^{n}\left(\sum_{\mu\epsilon\{\pi\}}\left[\mu\left(i\right)-\frac{n+1}{2}\right]\right)\left(\sum_{\nu\epsilon E}\left[\nu\left(i\right)-\frac{n+1}{2}\right]\right) \end{aligned}$$

Setting $\bar{\pi}_i = \frac{\sum_{j=1}^{n_i} \pi_i(j)}{n_i}, i = 1, 2$, it follows that

$$\frac{1}{n_1!n_2!} \sum_{\mu\epsilon\{\pi\}} \left[\mu\left(i\right) - \frac{n+1}{2}\right] = \begin{cases} \left[\bar{\pi}_1 - \frac{n+1}{2}\right] & if \;\; i \leq n_1 \\ \left[\pi_0 - \frac{n+1}{2}\right] & if \;\; i = n_1 + 1 \\ \left[\bar{\pi}_2 - \frac{n+1}{2}\right] & if \;\; i > n_1 + 1 \end{cases} \tag{2.5}$$

and

$$\frac{1}{n_1!n_2!} \sum_{\nu\epsilon E_1} \left[\nu\left(i\right) - \frac{n+1}{2}\right] = \begin{cases} \left(n_1 + 1\right)\left(-\frac{n_2}{2}\right) & if \;\; i \leq n_1 + 1 \\ \\ \left(n_1 + 1\right)\left(\frac{n_1+1}{2}\right) & if \;\; i > n_1 + 1 \end{cases} \tag{2.6}$$

Similarly

$$\frac{1}{n_1!n_2!} \sum_{\nu\epsilon E_2} \left[\nu\left(i\right) - \frac{n+1}{2}\right] = \begin{cases} \left(n_2 + 1\right)\left(-\frac{n_2+1}{2}\right) & if \;\; i \leq n_1 + 1 \\ \\ \left(n_2 + 1\right)\left(\frac{n_1}{2}\right) & if \;\; i > n_1 + 1 \end{cases} \tag{2.7}$$

Using the fact that

$$n_1\bar{\pi}_1 + \pi_0 + n_2\bar{\pi}_2 = \frac{\left(n\right)\left(n+1\right)}{2} \tag{2.8}$$

the classification rule (2.3) becomes: Classify X_0 into π_1 if and only if

$$\left(n_1 - n_2\right)\left(\frac{\left(n^2 - 1\right)}{6} + n_1\left[\bar{\pi}_1 - \frac{n+1}{2}\right]\right) + \left(n_1 + 1\right)\left(\pi_0 - \frac{n+1}{2}\right) \leq 0 \tag{2.9}$$

When $n_1 = n_2$, the new observation is classified into π_1 if its rank in the combined samples is less than the average $\left(\frac{n+1}{2}\right)$. We note that in developing the classification rule in (2.9) we assumed that π_1 was located to the left of π_2. In practice such information can be obtained from knowledge of the histogram of values for the training data.

3 Classification for two multivariate populations

With a few simple modifications, the univariate case can be extended to the situation involving two multivariate populations. The p x n data matrix can now be expressed as

$$D = \begin{pmatrix} X_{11}^{(1)} & \cdots & X_{1n_1}^{(1)} & X_0^{(1)} & X_{21}^{(1)} & \cdots & X_{2n_2}^{(1)} \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ X_{11}^{(p)} & \cdots & X_{1n_1}^{(p)} & X_0^{(p)} & X_{21}^{(p)} & \cdots & X_{2n_2}^{(p)} \end{pmatrix} \tag{3.1}$$

Define the univariate ranking vector

$$\pi^{(l)\prime} = \left(\pi_1^{(l)}(1), ..., \pi_1^{(l)}(n_1)\,|\pi_0^{(l)}|\pi_2^{(l)}(1), ..., \pi_2^{(l)}(n_2)\right), l = 1, ..., p \tag{3.2}$$

and let $\left\{\pi^{(l)}\right\}$ be the corresponding permutation set generated from $\pi^{(l)}$. Let $d\left(\left\{\pi^{l}\right\}, E_j^{(l)}\right)$ be the univariate distance between the data vector and corresponding extremal set for the l^{th} variable when $X_0^{(i)}$ is classified in π_j. The vector of distances when the vector X_0 is classified in π_j can now be defined as

$$d(\{\Pi\}, E_j) = \left(d\left(\left\{\pi^{(1)}\right\}, E_j^{(1)}\right), ..., d\left(\left\{\pi^{(p)}\right\}, E_j^{(p)}\right)\right)' \tag{3.3}$$

where

$$\Pi = \begin{pmatrix} \pi^{(1)\prime} \\ \vdots \\ \pi^{(p)\prime} \end{pmatrix} \tag{3.4}$$

and for $j = 1, 2$,

$$E_j = \begin{pmatrix} E_j^{(1)} \\ \vdots \\ E_j^{(p)} \end{pmatrix} \tag{3.5}$$

A simple measure of the distance between the two sets may be given by the sum

$$d_1(\{\Pi\}, E_j) = \sum_{l=1}^{p} d\left(\left\{\pi^{(l)}\right\}, E_j^{(l)}\right) \tag{3.6}$$

The classification rule then consists of classifying X_0 in π_1 if and only if

$$d_1(\{\Pi\}, E_1) \leq d_1(\{\Pi\}, E_2)\,. \tag{3.7}$$

The classification rule can be computed by using the univariate procedures for each of the p variables separately and then summing. We note that this rule does not take into direct account correlations which may exist among the variables. This will be done rather through the resulting distribution. Alternatively, we may define the following measure of distance

$$S_j = d\left(\{\Pi\}, E_j\right)' \Sigma_j^{-1} d\left(\{\Pi\}, E_j\right), j = 1, 2 \tag{3.8}$$

where the matrix Σ_j represents the variance-covariance matrix of $d\left(\{\Pi\}, E_j\right)$. In that case, the classification rule consists of classifying X_0 in π_1 if and only if

$$S_1 \leq S_2 \tag{3.9}$$

The extension to more than two populations is straightforward. The distances from the data vector to the extremal set is computed for each population. The new observation is then classified into the population which is closest. We shall use this approach when dealing with Fisher's iris data in the next section.

4 Monte Carlo Simulation Study

Monte Carlo simulation has been conducted to determine the probability of correct classification for the proposed procedure. First, large sample simulations were performed using both univariate and multivariate data. For the univariate case, samples of observations for two populations were obtained for the Normal, Log-Normal, Logistic, Cauchy and Exponential distributions. Simulations were then done under four different conditions:

$1.f_1, f_2$ come from the same family of distributions and have equal scale parameters but unequal location parameters;

2. f_1, f_2 come from the same family of distributions and have unequal location and scale parameters;

$3.f_1, f_2$ come from different families of distributions and have equal scale parameters but unequal location parameters;

$4.f_1, f_2$ come from different families of distributions and have unequal location and scale parameters.

In the case of multivariate data, the classification rule in (3.7) was used where $\pi_i, i = 1, 2$ are assumed to be multivariate normal. Simulations were done under two different scenarios:

$5.f_1, f_2$ have unequal means but equal covariance matrices;

$6.f_1, f_2$ have unequal means and unequal covariance matrices.

Finally, the well-known iris data set, available in the statistical software program R, was used to test the rank classification procedure's capability when used with more than two multivariate populations.

For each simulation, the rank classification procedure was compared with the Bayes rule, the optimal parametric classification method. The performance of both classification procedures was measured using a leave-one-out cross-validation technique, which provides an estimate of the true error rate.

π_1	π_2	Bayes	Ranking
$N\,(0,1)$	$N\,(2,1)$	15.7	6.38
$N\,(0,1)$	$N\,(5,1)$	0.59	0.00
$LN\,(0,1)$	$LN\,(2,1)$	16.13	6.61
$LN\,(0,1)$	$LN\,(5,1)$	0.58	0.005
$C\,(0,1)$	$C\,(2,1)$	25.3	13.74
$C\,(0,1)$	$C\,(5,1)$	12.2	6.89
$C\,(0,1)$	$C\,(10,1)$	6.51	3.54
$EXP\,(0,1)$	$EXP\,(2,1)$	7.05	4.83
$LOG\,(0,1)$	$LOG\,(2,1)$	26.51	11.93
$LOG\,(0,1)$	$LOG\,(5,1)$	7.13	2.28
$C\,(0,1)$	$LOG\,(2,1)$	24.18	12.83
$LOG\,(0,1)$	$N\,(2,1)$	21.27	9.97
$C\,(0,1)$	$N\,(1,1)$	28.42	15.05

Table 1: Error rate for univariate densities: unequal location , equal scale parameters

Simulations were conducted using $10,000$ observations from each population. The simulations indicate that the new rank classification procedure performed better than the Bayes rule most of the time. Some selected results are displayed in Tables 1-3.

The tables indicate that the ranking procedure has a lower cross validation error rate than the Bayes rule. In part, this may be the result of using the additional information on the location of the populations relative to each other.

The Fisher iris data contains 150 4-dimensional observations, 50 for each of three different populations: iris setosa, versicolor and virginica. Assuming a multivariate normal distribution, we obtained results summarized in Table 4.

Notation. $N\left(\mu,\sigma^2\right)$: Normal density with mean μ and variance σ^2

$LN\left(\mu,\sigma^2\right)$: Log-Normal density with mean μ and variance σ^2
$C\,(\mu,\sigma)$: Cauchy density with location μ and scale σ
$EXP\,(\mu,\sigma)$: Exponential density with location μ and scale σ
$LOG\,(\mu,\sigma)$: Logistic density with location μ and scale σ
$N\,(\mu,\Sigma)$:Multivariate normal with mean μ and covariance Σ

5 Concluding Remarks

There exist many different classification procedures and each one has its own advantages and limitations. The total probability of misclassification, calculated in the previous section for various cases, along with the simulations presented in section 5, clearly indicate that the rank classification procedure proposed in this article provides a competitive nonparametric alternative to the Bayes rule. The rank classification procedure performed better than the optimal

π_1	π_2	Bayes	Ranking
$N\,(0,1)$	$N\,(2,0.5)$	8.3	2.67
$N\,(0,1)$	$N\,(2,2)$	22.83	10.68
$N\,(0,1)$	$N\,(5,2)$	4.12	0.92
$C\,(0,1)$	$C\,(2,0.5)$	18.81	11.18
$C\,(0,1)$	$C\,(5,1)$	28.65	15.91
$C\,(0,1)$	$C\,(10,1)$	8.65	5.09
$C\,(0,1)$	$C\,(5,2)$	15.96	9.77
$C\,(0,1)$	$C\,(10,0.5)$	4.19	2.51
$C\,(0,1)$	$C\,(10,2)$	8.62	5.33
$EXP\,(0,1)$	$EXP\,(2,0.5)$	6.49	5.35
$EXP\,(0,1)$	$EXP\,(2,2)$	6.72	3.48
$LOG\,(0,1)$	$LOG\,(2,0.5)$	18.9	8.99
$LOG\,(0,1)$	$LOG\,(2,2)$	28.99	14.17
$LOG\,(0,1)$	$LOG\,(5,0.5)$	3.14	0.78
$LOG\,(0,1)$	$LOG\,(5,2)$	14.49	6.6
$C\,(0,1)$	$LOG\,(3,2)$	24.27	13.22
$LOG\,(0,1)$	$N\,(3,1)$	15.28	12.64
$C\,(0,1)$	$N\,(3,2)$	21.87	9.83

Table 2: Error rate for univariate densities: unequal location and scale parameters

$N\,(\mu_1,\Sigma_1)$	$N\,(\mu_2,\Sigma_2)$	Bayes	Ranking
$\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1&0\\0&1\end{pmatrix}$	$\begin{pmatrix}2\\2\end{pmatrix},\begin{pmatrix}1&0\\0&1\end{pmatrix}$	8.25	1.63
$\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1&-\frac{1}{2}\\-\frac{1}{2}&1\end{pmatrix}$	$\begin{pmatrix}1\\-3\end{pmatrix},\begin{pmatrix}1&-\frac{1}{2}\\-\frac{1}{2}&1\end{pmatrix}$	6.6283	4.5
$\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}4&3\\3&4\end{pmatrix}$	$\begin{pmatrix}2\\2\end{pmatrix},\begin{pmatrix}4&3\\3&4\end{pmatrix}$	29.39	11.19
$\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1&0\\0&1\end{pmatrix}$	$\begin{pmatrix}2\\2\end{pmatrix},\begin{pmatrix}2&0\\0&3\end{pmatrix}$	11.63	2.73
$\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}4&2\\2&3\end{pmatrix}$	$\begin{pmatrix}1\\-3\end{pmatrix},\begin{pmatrix}5&-1\\-1&5\end{pmatrix}$	17.57	3.18

Table 3: Error rate for multivariate normal density:

Bayes procedure	Ranking procedure
2.67	0.67

Table 4: Iris data

parametric method most of the time. In fact, in many of these cases, the rank classification method had an estimated error rate that was nearly 50% smaller. For the occurrences where the rank procedure had a higher error percentage than the Bayes rule, the estimated error rates of both methods were comparable, with only a minor increase in the error rate of the rank procedure. Furthermore, unlike the Bayes rule, the rank classification procedure is entirely nonparametric which is useful given that in most applications little is known about the underlying distributions. The rank classification procedure also demonstrated that it has good classification properties for various underlying densities.

Knowing in advance the probabilities of error in classification can be helpful if the researcher wishes to set the cut-off points in order to achieve a specified probability of misclassification. Anderson (1973) addressed this issue in the parametric case involving normal distributions. In a future paper, we shall consider other approaches to classification including a method based on ranks and proposed by Mantel and Valand (1970).

References

[1] Alvo, M. (2008). Nonparametric tests of hypotheses for umbrella alternatives. *The Canadian Journal of Statistics*, 36, 143-156.

[2] Anderson, T. W. (1973). An asymptotic expansion of the distribution of the Studentized classification statistic W. *The Annals of Statistics*, 1, 964-972.

[3] Johnson, R. A. and Wichern, D. W. (2002). *Applied Multivariate Statistical Analysis*. Fifth Edition, *Prentice Hall, Inc*.

[4] Mantel, N. and Valand, R. S. (1970). A technique of nonparametric multivariate analysis. *Biometrics*, 26, 547-558.