Academic Journal of Applied Mathematical Sciences
ISSN(e): 2415-2188, ISSN(p): 2415-5225
Vol. 6, Issue. 8, pp: 181-186, 2020
URL: https://arpgweb.com/journal/journal/17
DOI: https://doi.org/10.32861/ajams.68.181.186

Academic Research Publishing Group

Original Research Open Access

A Note on Different Types of Probabilities of Misclassification

Awogbemi Clement Adeyeye
Department of Statistics, National Mathematical Centre, Abuja, Nigeria
Email: awogbemiadeyeye@yahoo.com

Article History
Received: August 5, 2020
Revised: August 30, 2020
Accepted: September 7, 2020
Published: September 10, 2020

Abstract

Whenever a discriminant function is constructed, the attention of a researcher is often focused on classification. The underlined interest is how well does a discriminant function perform in classifying future observations correctly. In order to assess the performance of any classification rule, probabilities of misclassification of a discriminant function serves as a basis for the procedure. Different forms of probabilities of misclassification and their associated properties were considered in this study. The misclassification probabilities were defined in terms of probability density functions (pdf) and classification regions. Apparent probability of misclassification is expressed as the proportion of observations in the initial sample which are misclassified by the sample discriminant function. Different methods of estimating probabilities of misclassification were related to each other using their individual shortcomings. The status of degrees of uncertainties associated with probabilities of misclassification and their implications were also specified.

Keywords: Probabilities of misclassification; Classification regions; Estimated probability of misclassification; Discriminant function; Mahalanobis distance.

1. Introduction

Probability of misclassification expressed by P_{jk} is the probability of classifying an observation to population j when it is actually from population k. It occurs when there is a selection of criteria that is not suitable for classification [1, 2]. An observation X may be classified as belonging to population π_1 when it actually comes from population π_2, or vice versa. These errors are of serious concern in the choice of the procedure and as such, one is required as much as possible to reduce the errors or, more appropriately, their probabilities of occurrence [3, 4].

Let $f_1(x)$ and $f_2(x)$ be the probability density functions associated with X for populations π_1 and π_2 respectively. If R_1 is the set of values of X for which observations in π_1 are classified and R_2 is the set of values of X for which observations in π_2 are classified, then the probabilities of correctly or incorrectly classifying observations are:

- Pr (correctly classifying an object from π_1 to π_1) = $\Pr(1|1) = \Pr(X \in R_1 | \pi_1) = \int_{R_1} f_1(x)dx$
- Pr (misclassifying an object from π_2 to π_1) = $\Pr(1|2) = \Pr(X \in R_1 | \pi_2) = \int_{R_1} f_2(x)dx$
- Pr (correctly classifying an object from π_2 to π_2) = $\Pr(2|2) = \Pr(X \in R_2 | \pi_2) = \int_{R_2} f_2(x)dx$
- Pr (misclassifying an object from π_1 into π_2) = $\Pr(2|1) = \Pr(X \in R_2 | \pi_1) = \int_{R_2} f_1(x)dx$

The different probabilities of misclassification considered in this study are significant in the sense that the construction of a discriminant function would prompt a researcher to determine how this function performs on the validity of future samples [5]

2. Description of Probabilities of Misclassification

2.1. Optimum Probability of Misclassification

Optimum probability of misclassification assumes that the parameters of a distribution in the two populations are known and cannot be improved upon. According to John [6], the total optimum probability of misclassification is defined as:

$$\alpha(R, f) = P_1\int_{R_2} f_1(x)\,dx + P_2\int_{R_1} f_2(x)\,dx \,, \tag{1}$$

where R is the entire region of classification, f is the distribution of the observations that will be classified and $P_i,\ (i=1,2)$ refers to the a priori probability that an observation comes from population $\pi_i,\ (\mathrm{i}=1,2)$.

Let $X \in \pi_1 \sim N\left(\mu_1, \sigma^2\right)$ and $X \in \pi_2 \sim N\left(\mu_2, \sigma^2\right)$ with classification regions as follows:

$$R_1 : \left\{ X : X \le \frac{1}{2}(\mu_1+\mu_2) \right\}$$

$$R_2 : \left\{ X : X > \frac{1}{2}(\mu_1+\mu_2) \right\}. \tag{2}$$

The optimum probability of misclassification when observation from π_1 is misclassified is given by

$$\alpha_1(R, f) = \int_{R_2} f_1(x)\,dx = \Phi\left(\frac{\Delta}{2}\right), \tag{3}$$

where $f_1(x)$ is the probability density function associated with the random vector X for the population π_1, R_2 is the set of values of X for which observations into π_2 are classified, Φ is the cumulative standard distribution function and Δ is the mahalanobis distance between populations π_1 and π_2 defined by

$$\Delta = \left[(\mu_1-\mu_2)'\,\Sigma^{-1}(\mu_1-\mu_2)\right]^{\frac{1}{2}}.$$

Similarly, the optimum probability of misclassification when an observation from π_2 is misclassified is given as

$$\alpha_2(R, f) = \int_{R_1} f_2(x)\,dx = 1-\Phi\left(-\frac{\Delta}{2}\right), \tag{4}$$

where $f_2(x)$ is the probability density function associated with the random vector X for the population π_2, R_1 is the set of values of X for which observations into π_1 are classified, Φ is the cumulative standard distribution function and Δ is the mahalanobis distance between populations π_1 and π_2 defined by

$$\Delta = \left[(\mu_1-\mu_2)'\,\Sigma^{-1}(\mu_1-\mu_2)\right]^{\frac{1}{2}}.$$

Sedransk and Okamato [7], gave similar result on probability of misclassification when the variance in two populations, π_1, π_2, is given by σ^2.

Suppose X in populations, π_1 and π_2 has the density function

$$f_i(x) = (2\pi)^{-\frac{p}{2}}\left(|\Sigma|\right)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x-\mu_i)'\,\Sigma^{-1}(x-\mu_i)\right\}, \quad i=1,2 \tag{5}$$

The parameters, μ_i and $\sum$, satisfy the conditions, $-\infty < \mu_i < \infty$ and $\sum$ is a positive definite symmetric matrix of order p. The optimum probabilities based on the classification regions:

$$R_1 : \{X : D(X) : \mu_1, \mu_2, \Sigma) = \left[X - \frac{1}{2}(\mu_1+\mu_2)\right]' \Sigma^{-1}(\mu_1-\mu_2) > 0$$

$$R_2 : \{X : D(X) : \mu_1, \mu_2, \Sigma) = \left[X - \frac{1}{2}(\mu_1+\mu_2)\right]' \Sigma^{-1}(\mu_1-\mu_2) \le 0 \tag{6}$$

are given by

$$\alpha_1(R, f) = \Phi\left(\frac{\Delta}{2}\right)$$

$$\alpha_2(R, f) = 1 - \Phi\left(-\frac{\Delta}{2}\right) \tag{7}$$

2.2. Conditional Probability of Misclassification

The conditional probability of misclassification is usually calculated when a sample discriminant function is involved in the classification rule. Given a discriminant function, the probability can be described as the conditional probability that a randomly chosen member of $\pi_i, (i=1,2)$ is misallocated. It is not only conditional on the individual coming from one of populations π_1 or π_2, but also on the estimates of the means of the distribution in the two populations.

John [6], obtained the conditional probability of misclassification when an observation from population π_1 is misclassified as:

$$\alpha_1(R, f) = \begin{cases} 1 - \Phi\left[\sigma_1^{-1}\left(\frac{1}{2}\{\overline{X_1} + \overline{X_2}\} - \mu_1\right)\right]; if\ \overline{X_1} < \overline{X_2} \\ \Phi\left[\sigma_1^{-1}\left(\frac{1}{2}\{\overline{X_1} + \overline{X_2}\} - \mu_1\right)\right]; if\ \overline{X_1} \geq \overline{X_2}, \end{cases} \tag{8}$$

where Φ is the cumulative standard distribution function, σ_1^{-1} is the inverse of standard deviation fr $from\ population\ \pi_1, \mu_1\ is\ mean\ from\ population\ \pi_1,\ \overline{X_1}\ \ and\ \ \overline{X_2}$ are the sample means from populations $\pi_1\ \ and\ \ \pi_2$ respectively.

The conditional probability of misclassification when an observation from population π_2 is misclassified is given as

$$\alpha_2(\mathrm{R}, f) = \begin{cases} \Phi\left[\sigma_2^{-1}\left(\frac{1}{2}\{\overline{X_1} + \overline{X_2}\} - \mu_2\right)\right]; if\ \overline{X_1} < \overline{X_2} \\ 1 - \Phi\left[\sigma_2^{-1}\left(\frac{1}{2}\{\overline{X_1} + \overline{X_2}\} - \mu_2\right)\right]; if\ \overline{X_1} \geq \overline{X_2}, \end{cases} \tag{9}$$

where Φ is the cumulative standard distribution function, σ_2^{-1} is the inverse of standard deviation $from\ population\ \pi_1, \mu_1\ is\ mean\ from\ population\ \pi_1,\ \overline{X_1}\ \ and\ \ \overline{X_2}$ are the sample means from populations $\pi_1\ \ and\ \ \pi_2$ respectively.

2.3. Estimated Probability of Misclassification

Estimated probability of misclassification often referred to as the "plug-in estimate" was suggested by Fisher [8]. This was premised on the fact that the maximum likelihood estimates of the parameters are plugged in the discriminant function prior to classification. The total estimated probability of misclassification is given by

$$\alpha\left(\hat{R}, \hat{f}\right) = P_1 \int_{\hat{R}_2} \hat{f}_1(x)dx + P_2 \int_{\hat{R}_1} \hat{f}_2(x)dx, \tag{10}$$

where $R_1\ and\ R_2$ are respective sub-regions of classification corresponding to populations $\pi_1\ and\ \pi_2$, $f_1(x)\ and\ f_2(x)$ are the respective density functions of X in populations, $\pi_1\ and\ \pi_2$ and $P_1\ and\ P_2$ are the a priori probabilities that an observation comes from $\pi_1\ and\ \pi_2$, respectively.

2.4. Apparent Probability of Misclassification

Apparent probability of misclassification was suggested by, Smith [9] and defined as the proportion of observations in the initial sample which are misclassified by the sample discriminant function. If n_1 is the proportion of observation misclassified by the discriminant function in population π_1, and n is the total sample size in population π_1, then the apparent probability of misclassification is $\frac{n_1}{n}$.

2.5. Expected Probability of Misclassification

The expected probability of misclassification has been discussed in the literature as the expected value of the conditional probability of misclassification. It is otherwise known as unconditional probability of misclassification [6]. The total expected probability of misclassification is defined as:

$$E\left[\alpha\left(\hat{R}, f\right)\right] = E\left[P_1\int_{R_2} f_1(x)dx + P_2\int_{R_1} f_2(x)dx\right], \tag{11}$$

where R_1 and R_2 are respective sub-regions of classification corresponding to populations π_1 and π_2, $f_1(x)$ and $f_2(x)$ are the respective density functions of X in populations π_1 and π_2 and P_1 and P_2 are the a priori probabilities that an observation comes from π_1 and π_2 respectively.

The expressions for the expected probability of misclassification and its approximations were given by John [6] using the Anderson's classification statistic (W) as:

$$E\left[\alpha_1\left(\hat{R}, f\right)\right] = Q\left(a_{11}, a_{21}; \rho\right) + Q\left(a_{12}, a_{22}; \rho\right)$$

$$a_{11} = -\left[n_1^{-1}\sigma_1^2 + n_2^{-1}\sigma_2^2\right](\mu_2 - \mu_1), \quad a_{12} = -a_{11}$$

$$a_{21} = \frac{1}{2}\left[\sigma_1^2 + \frac{1}{4}\left(n_1^{-1}\sigma_1^2 + n_2^{-1}\sigma_2^2\right)\right]^{-\frac{1}{2}}(\mu_2 - \mu_1), a_{22} = -a_{21}$$

$$\rho = \frac{1}{2}\left[n_1^{-1}\sigma_1^2 + n_2^{-1}\sigma_2^2\right]^{-\frac{1}{2}}\left[\sigma_1^2 + \frac{1}{4}\left(n_1^{-1}\sigma_1^2 + n_2^{-1}\sigma_2^2\right)\right]^{-\frac{1}{2}}\left(n_1^{-1}\sigma_1^2 - n_2^{-1}\sigma_2^2\right)$$

and

$$Q\left(h, k; \rho\right) = \int_{-\infty}^{h}\int_{-\infty}^{k} q(u, v; \rho)\, du\, dv, \tag{12}$$

where μ_1 and μ_2 are the respective means from populations, π_1 and π_2, $q(\mu, v, \rho)$ is the standard bivariate normal density function with correlation coefficient ρ, μ_1 and μ_2 are the means from populations π_1 and π_2, n_1 and n_2 are the sample sizes from π_1 and π_2, and σ_1^2, σ_2^2 are the variances from π_1 and π_2.

3. Methods of Estimating Probabilities of Misclassification

3.1. Parameter Substitution Method

With this method, the probability of misclassification is estimated directly by substituting sample estimates of population parameters in the theoretical expression for the probability of misclassification. The method is a natural estimate and maximum likelihood estimator of the error rate. It is also said to be highly biased for small sample sizes [10].

3.2. Re-substitution Method

This procedure results to apparent error rate since the proportion of the sample incorrectly classified is used as the estimate of probability of misclassification [11]. Let π_1 *and* π_2 *be* the probabilities of misclassification of erroneously assigning an observation to group $i(P_1)$ when the observation comes from group $j(P_2)$, then $\hat{\pi}_1$ *and* $\hat{\pi}_2$ are the sample proportions of misclassified observations. The estimates are consistent, but can be severely biased for small sample sizes. This method underestimates the probability of misclassification since the data used for fitting and validating the model are the same [12].

3.3. Holdout Method

This method splits the total sample into two equal parts so as to overcome the shortcoming of re-substitution approach. One subsample is employed to construct the classification rule and second part for validation. However, it requires large samples; otherwise its estimate of misclassification suffers [13].

3.4. Cross-validation Method

The method uses all of the available data without serious bias in the estimated error rates. It holds out one observation at a time, estimates the distribution function based on $n_1 + n_2 - 1$ observations and classifies the held out observations. This process is repeated until all observations are classified. Let n_1 *and* n_2 be the number of sampled observations misclassified in groups P_1 *and* P_2 *respectively*, then the estimated classification error rates are $\hat{\pi}_1 = \frac{m_1}{n_1}$ *and* $\hat{\pi}_2 = \frac{m_2}{n_2}$

The method produces unbiased estimates of the probability of misclassification for a rule based on $n_1 - 1$ *and* $n_2 - 1$ observations, respectively [4]

3.5. Jacknife Method

In order to overcome the defects of methods (3.2) and (3.3), application of Jacknife was proposed by Lachenbruch [14]. According to this procedure, the linear discriminant function is fitted to all but one observation. The linear discriminant function is then applied to the (n-1) observations in the sample, and repeated n times [15]. This method was later examined in the context of the discrimination problem by Crask and Perreault [16]. Their work focused on the simultaneous use of its cross validation and Jacknife analysis. While cross validation method obtains good estimates of classification error rates, Jacknife analysis considers coefficient stability.

3.6. Boostrap Method

The bootstrap method is an extension of Jack-knife and might also be thought of as a finite sample Monte Carlo procedure. According to Samprit and Sangit [10], the method operates as follows:

- From the sample of the ith population ($i = 1,2, \dots g$), draw an independent sample of size n_i with each unit being drawn with a probability $\frac{1}{n_i}$ ($i = 1,2, \dots g$). The sample drawn from each of the G groups constitutes the bootstrap sample.
- On the basis of the bootstrap sample, the linear discriminant function is constructed and its performance is evaluated by classifying all the observations not included in the bootstrap sample. The proportion of observations correctly classified is observed.
- The aforementioned steps are repeated a large number of times and each trial generates an estimate of misclassification probability. The average of all the sample outcomes is taken as the bootstrap estimate, and the standard deviation of the estimates provides an estimate of the standard error.

4. Significance of Probabilities of Misclassification

A qualitative value of predictive performance of a classification model provided by uncertainty estimation is anchored on probabilities of misclassification. Low probability of misclassification is linked to low degree of misclassification which implies high reliability. High probability of misclassification is connected to high degree of improbability indicating propensity to generate erroneous classification.

5. Conclusion

Probability of misclassification is a decisive factor used to evaluate a classification procedure. Different approaches have been designed and related to one another in order to find the best possible way of estimating the true probabilities of misclassification. These methods have resulted to different types of probabilities of misclassification. The boostrap method has the advantage of not only furnishing the estimates of misclassification probabilities but also provides an estimate of the standard error of estimate.

References

[1] Awogbemi, C. A., 2020. "Graphical evaluation of probabilities of misclassification for normal and edgeworth series distributions." *Journal of Research and Innovations in Applied Science,* vol. 5, pp. 203-213.

[2] Sorum, M., 1972. "Three probabilities of misclassification." *Technometrics,* vol. 14, pp. 309-316.

[3] Awogbemi, C. A. and Onyeagu, S. I., 2019. "Errors of misclassification associated with edgeworth series distribution." *American Journal of Theoretical and Applied Statistics,* vol. 8, pp. 203-213.

[4] Onyeagu, S. I., 2003. *A first course in multivariate statistical analysis*. Awka: Mega Concept.

[5] Camilo, L. M., Kassio, M. G., and Francis, L. M., 2019. "Uncertainty estimation and misclassification probability for classification models based on discriminant analysis and support vector." *Analytica Chimica Acta,* vol. 1063, pp. 40-46.

[6] John, S., 1961. "Errors in discrimination." *Annals of Mathematical Statistics,* vol. 22, pp. 1125-1144.

[7] Sedransk, N. and Okamato, M. C., 1971. "Estimation of the probabilities of misclassification for the linear discriminant function in the univariate normal case." *Annals of Statistics,* vol. 23, pp. 419-427.

[8] Fisher, F. A., 1936. "The use of multiple measurements in taxonomic problems." *Annals of Eugenics,* vol. 7, pp. 179-188.

[9] Smith, C. A. B., 1947. "Some examples of discrimination." *Annals of Eugenics,* vol. 18, pp. 272-283.

[10] Samprit, C. and Sangit, C., 2007. "Estimating misclassification probabilities by boostrap methods." *Communications in Statistics and Computation,* vol. 12, pp. 645-656.

[11] Hills, M., 1966. "Allocation Rules and their error rates." *Journal of Royal Statistical Society,* vol. B28, pp. 1-26.

[12] Lachenbruch, P. A. and Mickey, M. R., 1968. "Estimation of error rates in discriminant analysis." *Technometrics,* vol. 10, pp. 1-11.

[13] William, R. D. and Matthew, G., 1984. *Multivariate analysis, methods and applications*. New York: John Wiley and Sons Inc.

[14] Lachenbruch, P. A., 1975. *Discriminant analysis*. New York: Hafner Press.

[15] Osuji, G. A., Onyeagu, S. I., and Ekezie, D. D., 2013. "Comparison of jackknife and resubstitution." *International Journal of Mathematics and Statistics Studies* vol. 1, pp. 29-37. Available: http://www.eajournals.org/wp-content/uploads/COMPARISON-OF-JACKKNIFE-AND-

RESUBSTITUTION-METHODS-IN-THE-ESTIMATION-OF-ERROR-RATES-IN-DISCRIMINANT-ANALYSIS.pdf
[16] Crask, M. R. and Perreault, W. D., 1977. "Validation of discriminant analysis in marketing research." *Journal of Marketing Research*, vol. 11, pp. 60-64.