Appl. Math. J. Chinese Univ.
2023, 38(2): 180-191

Misclassification analysis of discriminant model

HUANG Li-wen[1,2]

Abstract. This paper extends the criterion of the misclassification ratio of discriminant model and presents a new selection method of discriminant model. For selecting the discriminant model, this method establishes the rule of misclassification degree ratio through misclassification ratio of the discriminant model and misclassification degree of the samples. To test the effect of this method, this work uses seven UCI data sets. Numerical experiments on these examples indicate that this method has certain rationality and has a better effect to select a discriminant model.

§1 Introduction

Discriminant Analysis is a statistical method that is used to determine the sample type, and it was introduced by Fisher[1] for two-class problems. Over the past decades, many well-developed approaches have been proposed in order to improve the performance of discriminant models. Shinmura[2] summarized the problem of some discriminant methods and presented the new theory of discriminant analysis after Fisher. Among these methods, Song et al.[3], Li and Lei[4], and Hidaka et al.[5] have proposed new methods to improve the application of the method ; Chen and Li[6], Ji et al.[7], Huang and Su[8], Xu et al.[9], and Huang[10] have modified the corresponding discriminant analysis method in order to enhance the classification performance; Tang et al.[11], Yang et al.[12], Zhang and Wang[13], and Pacheco et al.[14] have mainly raised the effect of discriminant analysis from the perspective of variable selection or dimensionality reduction. However, most of the current approaches were designed to minimize the misclassification ratio (MR) or maximize the accurate rate. In fact, these methods implicitly assume that the misclassification cost of every sample is equal, but in many real-world domains, the misclassification cost is often different. For example, in medical diagnosis, when a healthy person is misclassified as catching a cold, its misclassification cost is often less than the person that is misclassified as a cancer patient.

Received: 2019-06-06. Revised: 2019-09-11.
MR Subject Classification: 62H30.
Keywords: discriminant model, misclassification ratio, misclassification degree.
Digital Object Identifier(DOI): https://doi.org/10.1007/s11766-023-3823-8.
Supported by the National Natural Science Foundation of China(52070119) and Key Laboratory of Financial Mathematics of Fujian Province University (Putian University) (JR201801).

For this reason, cost-sensitive learning methods of discriminant model have been an increasing interest in statistics, pattern recognition, data mining, machine learning and other fields, see Mcdonald[15], Pan et al.[16], Bahnsen et al.[17]. These methods usually use the minimum cost-sensitive risk to measure the performance of the discriminant model, and the misclassification costs of different samples play a crucial role in the construction of the cost-sensitive learning model. However, in many contexts of an imbalanced dataset, the misclassification cost cannot be determined (Cao et al., [18]). To avoid this kind of situation, a selection method of discriminant model has been proposed (Huang, [19]), and an evaluation of misclassification sample was established with the help of Analytic Hierarchy Process, but this method needs background knowledge of related questions to determine the impact of misclassified samples, and their results may be different due to the different evaluators. However, in multi-class classification, a sample belonging to one class may be misclassified as the other classes; in this situation, it is hoped that the impacts of the misclassified samples will be achieved a minimum degree for all possible misclassification. For this purpose, this paper presents a new approach and introduces the new concept of misclassification degree (*MD*), and its basic idea can be described as follows: if a sample belongs to one class, but it is misclassified as the other class; it is hoped that the difference between the original class and the predictive class should be kept as small as possible. This paper focuses on the selection methods of a discriminant model. The goal is to establish the evaluation rules of discriminant models based on the *MR* and the *MD*, namely the total misclassification degree (*TMD*) and the misclassification degree ratio (*MDR*). Then, according to the proposed rules, a method of selecting the discriminant model is discussed, the purpose of which is to make the misclassification degree of the selected model as small as possible.

In the following sections, the paper will discuss the misclassification degree ratio of the discriminant model in section 2, numerical experiments are presented to demonstrate the effectiveness of the proposed method in section 3, and conclusions are given in the last section.

§2 Misclassification degree ratio of discriminant model

Given $k > 0$, suppose that there are k classes $(G_1, G_2, \ldots, G_k)$, where $G_p : x^{(p)}_{(1)}, x^{(p)}_{(2)}, \ldots, x^{(p)}_{(n_p)}$. Let $\mu^{(p)}$ be the average of G_p, which is expressed as $\mu^{(p)} = \left(\mu^{(p)}_1, \mu^{(p)}_2, \ldots, \mu^{(p)}_m\right)'$. Denote n_p as the sample size of class $G_p, p = 1, 2, \ldots, k$. In this paper, assuming that x is an arbitrary given sample, μ is the average of the total training sample data with $\mu = (\mu_1, \mu_2, \ldots, \mu_m)'$, and $X = (x_{(1)}, x_{(2)}, \ldots, x_{(n)})'$, where $x_{(i)}$ is the i-th sample,$i = 1, 2, \ldots, n, n = \sum\limits_{p=1}^{k} n_p$.

2.1 Misclassification matrix

Suppose that the discriminant model has been established by the training samples. Let n_{ij} $(i \neq j)$ be the number of the samples belonging to G_i that are misclassified as G_j, set $n_{ii} = 0$ when $i = j$, then the results of misclassification are given in Table 1.

Table 1. Misclassification matrix.

Original class	Predictive class				Sample size
	G_1	G_2	$\ldots$	G_k	
G_1	0	n_{12}	$\ldots$	n_{1k}	n_1
G_2	n_{21}	0	$\ldots$	n_{2k}	n_2
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
G_k	n_{k1}	n_{k2}	$\ldots$	0	n_k

From Table 1, if $p\left(G_i|x_j\right)$ is the probability of the sample x_j $\left(x_j \in G_j\right)$ that is misclassified as G_i, then $p\left(G_i|x_j\right)$ can be expressed by the following form.

$$p\left(G_i|x_j\right) = \tfrac{n_{ji}}{n}.$$

Similarly, if the total number of misclassification samples is denoted by *TNM*, then

$$TNM = \sum_{i=1}^{k} \sum_{\substack{j=1 \\ i \neq j}}^{k} n_{ij}.$$

Thus, *MR* of the discriminant model can be computed by the following formula:

$$MR = \tfrac{TNM}{n} \times 100\%.$$

2.2 The measure of misclassification degree

The measure of *MD* is the key to the evaluation of discriminant model. If the expert evaluation method is adopted, the evaluation results of different evaluators may have some differences, which indicates that this method has certain subjectivity. So this paper tries to measure *MD* by using the difference between the sample and its corresponding predictive class, and then proposes a new method in order to select a better discriminant model.

In general, it is hoped that the *MD* can avoid the influence of dimension and the correlation between variables, and reflect the degree of misclassified samples as much as possible. Euclidean distance is a common method to measure the closeness of two research objects, but the method is affected by the dimension. To overcome this drawback, it should usually be dimensionless first. For a single variable, there are many methods for dimensionless processing. Common methods include standardization, equalization, Min-max normalization, Efficacy coefficient method, and so on. For multi-variables, although these methods can be used for dimensionless processing, they cannot eliminate the correlation between variables. The Mahalanobis distance overcomes these two problems, that is, it can eliminate not only the dimension, but also the correlation between variables. Therefore, with the advantage of the Mahalanobis distance, the concept of the *MD* between the sample and the class is introduced below.

Definition 2.1. Let $x \in G_p, \mu^{(p)} = \left(\mu_1^{(p)}, \mu_2^{(p)}, \ldots, \mu_m^{(p)}\right)'$ and V is the variance-covariance matrix of X, then the distance between the samples x and G_p is defined as follows:

$$d\left(x, G_p\right)=\sqrt{\left(x-\mu^{(p)}\right)' V^{-1}\left(x-\mu^{(p)}\right)}.$$

Given a sample $x \in G_p$, if x is misclassified as G_q, and the difference between G_p and G_q is smaller, then its *MD* is smaller. So the *MD* of the sample can be defined by the following form.

Definition 2.2. Given $x \in G_p$, if x is misclassified as G_q, then the *MD* of x is defined as follows:

$$C\left(G_q|x\right)=\tfrac{\left|d\left(x, G_q\right)-d\left(x, G_p\right)\right|}{d\left(x, G_p\right)}.$$

In Definition 2.2, for any given sample $x \in G_p$, if $q \neq p$, then $C(G_q|x) \geq 0$. This shows its *MD* is greater than or equal to zero when the sample of G_p is misclassified as G_q. In general, the *MD* of a sample is related to the difference between G_p and G_q according to the Definition 2.2, and when the difference between G_p and G_q becomes larger, the value of $C\left(G_q|x\right)$ also becomes larger.

Due to the above discussions, the misclassification degree has the following conclusions.

Theorem 2.1. *Let G_p and G_q be two different classes, and the relationship of two classes are not inclusion and disjoint each other. If $x \in G_p$, then $C\left(G_q|x\right)$ is dimensionless.*

Proof. Suppose G_p and G_q are two different classes, and the corresponding new classes are denoted by G_p^* and G_q^* after the units of the original variables are changed. Let $w^{(p)}, w, y$ be the sample average of G_p^*, the average of new training sample data, the new arbitrary given sample, respectively. Then there exists a diagonal matrix α such that $w^{(p)}=\alpha\mu^{(p)}, w=\alpha\mu, y=\alpha x$, where $\alpha=\operatorname{diag}\left(\alpha_1, \alpha_2, \ldots, \alpha_m\right)$ and $\alpha_i>0, i=1, \ldots, m$.

Let V be the variance-covariance matrix of X, then

$$\begin{aligned} V &= \frac{1}{n-1} \sum_{i=1}^{n}(x_{(i)}-\mu)(x_{(i)}-\mu)' \\ &= \frac{1}{n-1}(X-I\mu')'(X-I\mu') \end{aligned}.$$

where $I=(1,1,\ldots,1)'_{m1}$.

If Y is the data set of new variables, and V^* be the variance-covariance matrix of Y, then $Y=X\alpha$,

$$\begin{aligned} V^* &= \frac{1}{n-1}(Y-Iw')'(Y-Iw') \\ &= \frac{1}{n-1}(X\alpha-I(\alpha\mu)')'(X\alpha-I(\alpha\mu)') \\ &= \frac{1}{n-1}(X\alpha-I\mu'\alpha')'(X\alpha-I\mu'\alpha') \\ &= \frac{1}{n-1}\alpha'(X-I\mu')'(X-I\mu')\alpha \\ &= \alpha' V\alpha \end{aligned}.$$

If y is misclassified as G_q^*, then

$$\begin{aligned}
d\left(y, G_p^*\right) &= \sqrt{\left(y - w^{(p)}\right)' V^{*-1}\left(y - w^{(p)}\right)} \\
&= \sqrt{\left(x - \mu^{(p)}\right)' \alpha'\alpha^{-1} V^{-1} (\alpha')^{-1} \alpha \left(x - \mu^{(p)}\right)} \\
&= \sqrt{\left(x - \mu^{(p)}\right)' V^{-1}\left(x - \mu^{(p)}\right)} \\
&= d\left(x, G_p\right)
\end{aligned}$$

Similarly, $d\left(y, G_q^*\right) = d\left(x, G_q\right)$. Thus,

$$C\left(G_q^* | y\right) = \frac{\left|d\left(y, G_q'\right) - d\left(y, G_p'\right)\right|}{d\left(y, G_p'\right)} = \frac{\left|d\left(x, G_q\right) - d\left(x, G_p\right)\right|}{d\left(x, G_p\right)} = C\left(G_q | x\right)$$

That is, $C(G_q|x)$ is dimensionless. □

Theorem 2.2. *If X is eliminated the dimension by standardization and equalization respectively, $C\left(G_q|x\right)$ remains the same.*

Proof. After X is eliminated the dimension by standardization, let y and Y be the arbitrary sample and the new data set of new variables, and let the classes corresponding to G_p and G_q be G_p' and G_q' respectively. Similarly, after X is eliminated the dimension by equalization, let z and Z be the arbitrary sample and the new data set of new variables, and let the classes corresponding to G_p and G_q be G_p^* and G_q^* respectively. So y, z, Y and Z can be expressed as follows:

$$y = \alpha(x - \mu), z = \beta x, Y = (X - Iu')\alpha, Z = X\beta.$$

where $\alpha = diag(\frac{1}{\sigma_1}, \frac{1}{\sigma_2}, \ldots, \frac{1}{\sigma_m}), \beta = diag(\frac{1}{\mu_1}, \frac{1}{\mu_2}, \ldots, \frac{1}{\mu_m})$, and $I = (1, 1, \ldots, 1)'_{m1}$.

if V_1^* is the variance-covariance matrix of standardization, and V_2^* is the variance-covariance matrix of equalization, then

$$\begin{aligned}
V_1^* &= \frac{1}{n-1} Y'Y \\
&= \frac{1}{n-1} (X\alpha - I(\alpha\mu)')'(X\alpha - I(\alpha\mu)') \\
&= \frac{1}{n-1} (X\alpha - I\mu'\alpha')'(X\alpha - I\mu'\alpha') \quad . \\
&= \frac{1}{n-1} \alpha'(X - I\mu')'(X - I\mu')\alpha \\
&= \alpha' V \alpha
\end{aligned}$$

$$\begin{aligned}
d\left(y, G_p'\right) &= \sqrt{\left(y - \alpha(\mu^{(p)} - \mu)\right)' V_1^{*-1}\left(y - \alpha(\mu^{(p)} - \mu)\right)} \\
&= \sqrt{\left(x - \mu^{(p)}\right)' \alpha'\alpha^{-1} V^{-1} (\alpha')^{-1} \alpha \left(x - \mu^{(p)}\right)} \\
&= \sqrt{\left(x - \mu^{(p)}\right)' V^{-1}\left(x - \mu^{(p)}\right)} \\
&= d\left(x, G_p\right)
\end{aligned}$$

Using the similar method, $d\left(y, G_q'\right) = d\left(x, G_q\right), V_2^* = \beta' V \beta, d\left(z, G_p^*\right) = d\left(x, G_p\right)$ and $d\left(z, G_q^*\right) = d\left(x, G_q\right)$.

Hence, $C\left(G_q|x\right) = C\left(G_q'|y\right) = C\left(G_q^*|z\right)$.

That is, $C\left(G_q|x\right)$ remains the same. □

Theorem 2.3. *Let G_p and G_q be two different classes, and the relationship of two classes are not inclusion and disjoint each other. If $d\left(x_{(i)}^{(p)}, G_p\right) = d\left(x_{(j)}^{(p)}, G_p\right)$ and $d\left(x_{(i)}^{(p)}, G_q\right) < d\left(x_{(j)}^{(p)}, G_q\right)$, then $C\left(G_q|x_{(i)}^{(p)}\right) < C\left(G_q|x_{(j)}^{(p)}\right)$.*

Theorem 2.4. *Let G_p, G_q and G_t be three different classes, and these classes are not inclusion and disjoint each other. Let $x \in G_t$. If $d\left(x, G_p\right) < d\left(x, G_q\right)$, then $C\left(G_p|x\right) < C\left(G_q|x\right)$.*

From Theorem 2.1 to Theorem 2.4, the *MD* is dimensionless, after the original data is eliminated the dimension by standardization or equalization, its value remains the same, and its value is related to the distance between the sample and the predictive class. If the predictive class is the original class, then its *MD* is equal to zero. Furthermore, as the difference between the sample and the predictive class increases, the *MD* of the sample also increases.

2.3 Misclassification degree analysis

From section 2.2, the *MD* of each sample is usually not the same. For a discriminant model, on the one hand, it is hoped that the *MD* will be smaller; on the other hand, it is hoped that the misclassification probability of the sample will be kept as small as possible. Furthermore, if all the misclassified samples can achieve the minimum *MD* and the minimum misclassification probability, then the corresponding discriminant model works best. Thus, for selecting a better discriminant model, the concept of total misclassification degree (*TMD*) is introduced by the following form.

Let G be a class that includes all the misclassification samples, and let G_x be the class that the sample x is misclassified, where G_x is one of a class in $G_1, G_2, \ldots, G_k$, and G_x varies with the sample x, then the definition of *TMD* is outlined below.

Definition 2.3. *If $x \in G, C\left(G_x|x\right)$ is the MD of the sample x that is misclassified as G_x, and $p\left(G_x|x\right)$ is the probability of the sample x that is misclassified as G_x, then TMD of discriminant model is defined as follows:*

$$TMD = \sum_{x \in G} C\left(G_x|x\right) p\left(G_x|x\right)$$

In general, it is hoped that the value of *TMD* can be kept as small as possible. When its value is smaller, it can be considered that the effect of the discriminant model is better. So the criterion of selecting discriminant model can be described as follows:

Rule 2.1. *Suppose there are several discriminant models $v_1, \ldots, v_s, s > 0$. The TMD of discriminant model v_i is denoted by $TMD\left(v_i\right), i = 1, 2, \ldots, s$. If $TMD\left(v_t\right) = \min_{1 \le i \le s}\left\{TMD\left(v_i\right)\right\}$, then v_t is the best model in these discriminant models.*

As shown in Rule 2.1, this rule can help to select a relatively better model from multiple discriminant models. However, it can not give a valid evaluation for a single discriminant model. To overcome this drawback, it is necessary to discuss the misclassification degree ratio (MDR) of discriminant model.

Suppose $C\left(G_x|x\right)$ is the MD of the sample x that is misclassified as G_x, then the MDR of the sample x can be expressed by the following form:

$$\frac{C(G_x|x)}{\sum\limits_{i=1}^{k} C(G_i|x)} \times 100\%$$

For the sake of convenience, let $\overline{C}_x$ be the average misclassification degree (AMD) of the sample x, then $\overline{C}_x = \frac{1}{k-1}\sum\limits_{i=1}^{k} C\left(G_i|x\right)$. The value of $\frac{C(G_x|x)}{\sum\limits_{i=1}^{k} C(G_i|x)}$ may become small as the class increases, so the form of $\frac{C(G_x|x)}{\sum\limits_{k=1}^{k} C(G_i|x)}$ can be replaced by $\frac{C(G_x|x)}{\overline{C}_x}$ in order to avoid this from happening. Thus, the total average misclassification degree ($TAMD$) of discriminant model can be described as follows:

$$TAMD = \begin{cases} \frac{1}{n}\sum\limits_{x\in G} \frac{C(G_x|x)}{\overline{C}_x} \times 100\% & TNM \neq 0 \\ 0 & TNM = 0 \end{cases}$$

Generally, if the value of $TAMD$ is smaller, then the effect of the discriminant model is better. Combined with the misclassification ratio of discriminant model, it is hoped that $TAMD$ and MR can achieve the minimum value for a discriminant model, so the evaluation criterion can be described as follows:

Rule 2.2. *Let TAMD be the total average misclassification degree of discriminant model, MR is the misclassification ratio of discriminant model, then the misclassification degree ratio (MDR) of discriminant model can be expressed by the following formula:*

$$MDR = \min\{\sqrt{TAMD \times MR}, 100\}$$

According to the Rule 2.2 above, the MDR of discriminant model has the following propositions:

Proposition 2.1. $0 \leq \min\{TAMD, MR\} \leq MDR \leq \max\{TAMD, MR\} \leq 100.$

Proposition 2.2. *If* $C\left(G_x|x\right) = C$*, then* $MDR = MR$.

Proposition 2.3. *Let* G *be a group that includes all the misclassification samples, for any given sample* x*, if* $x \in G_j$ $(1 \leq j \leq k)$ *and* $C\left(G_x|x\right) < \overline{C}_x$*, where* $\overline{C}_x = \frac{1}{k-1}\sum\limits_{i=1}^{k} C\left(G_i|x\right)$*, then* $MDR \leq MR$.

Proof. if $TNM = 0$, then $MDR = 0 = MR$.

if $TNM \neq 0$ and $C\left(G_x|x\right) < \overline{C}_x$, then

$$\frac{C(G_x|x)}{\overline{C}_x} < 1$$

Hence,

$$TAMD = \frac{1}{n}\sum_{x\in G}\frac{C(G_x|x)}{\overline{C}_x} \times 100\% < \frac{1}{n}\sum_{x\in G} 1 \times 100\% = \frac{TNM}{n} \times 100\% = MR$$

Namely, $MDR = \sqrt{TAMD \times MR} < MR$.

To sum up, $MDR \leq MR$.

$\square$

Similar to the proof of the Proposition 2.3, the other proposition can be obtained as follows:

Proposition 2.4. *Let G be a group that includes all the misclassification samples, for any given x, if $x \in G_j$ $(1 \leq j \leq k)$ and $C(G_x|x) > \overline{C}_x$, here $\overline{C}_x = \frac{1}{k-1}\sum_{i=1}^{k} C(G_i|x)$, then MDR $\geq$ MR.*

From these *MDR* propositions mentioned above, it is easy to determine whether the *MD* of the sample is greater than its *AMD* through the comparative analysis between *MR* and *MDR*. If *MDR*<*MR*, then the *MD* of the sample is less than its *AMD*; if *MDR*>*MR*, then the *MD* of the sample is greater than its *AMD*.

Thus, for a discriminant model, an appropriate value of *MR* can be set as the threshold according to requirement of the actual problem, and if *MR*>*MDR*, then the corresponding model works well; if *MR*<*MDR*, then the corresponding model achieves poor effect, which shows that the *MD* of sample is relatively larger.

§3 Numerical experiments

To evaluate the effect of *MDR*, seven data sets are selected from UCI Machine Learning Repository (Dua and Karra Taniskidou, [20]). These data sets are Iris Data Set, Balance Scale Data Set, Banknote Authentication Data Set, Breast Tissue Data Set, Vertebral Column 2c Data Set, Vertebral Column 3c Data Set, and Ecoli Data Set, respectively. Table 2 lists the basic information of seven data sets. Subsequently, in order to test the effect of *MDR*, the discriminant models are established by the following discriminant analysis methods, Discriminant Method of SPSS 18 (SPSS), Bayes Stepwise Discriminant Method (BSDM), Fisher Stepwise Discriminant Method (FSDM), and Hierarchical Discriminant Method (HDM).

In general, a certain discriminant method can not achieve better results than other methods in any case, so it is important to select an appropriate method. For the above four methods, the discriminant method in SPSS is suitable for the discriminant problem of the small sample, and its algorithm efficiency is general. In particular, when the sample is larger, its memory consumption is also larger and the efficiency of the algorithm is lower too. BSDM is proposed based on the conditions that the discriminant data has normality and equal covariance matrix. When the discriminant data meets these two conditions, good results can be achieved. The algorithm runs fast and the computational complexity is $O(k * n * m)$. FSDM has no special

Table 2. Information of data set.

Data set	Sample number	Variable	Class number
Iris	150	4	3
Balance Scale	625	4	3
Banknote Authentication	1372	4	2
Breast Tissue	106	9	6
Vertebral Column 2c	310	6	2
Vertebral Column 3c	310	6	3
Ecoli	336	7	8

Table 3. Misclassification matrix.

Original Class	Predictive class			Sample size
	G_1	G_2	G_3	
G_1	0	0	0	50
G_2	0	0	2	50
G_3	0	1	0	50

requirement for the discriminant data. Its discriminant effect is related to the type of data. When the discriminant data is the large between-class difference and the small within-class difference, the effect is better. When the difference of each class is small or there is an inclusion relationship between the classes , the effect is poor. The efficiency of the algorithm is high, and the computational complexity is $O(k*n*m)$. HDM is an improved discriminant method based on FSDM. Theoretically, there is no special requirement for the discriminant data. This method has advantages of FSDM, and it can deal with the discriminant problem of one class surrounded by the other class. The algorithm does not run as fast as FSDM, and the computational complexity is $O(k*n*m) \sim O(k*nlogn*m)$.

Taking the Iris data set as an example, and the specific processes of the misclassification analysis using the SPSS method is as follows.

(1) From Section 2.1, the misclassification cases of the method (SPSS) are given in Table 3.

(2) As the results given in Table 3, there are three misclassification samples. From Section 2.2, the degree of each misclassification sample is given in Table 4.

(3) From Section 2.3, the *MDR* of the SPSS method can be computed by the Rule 2; Similarly, the *MDR* of other methods can be obtained by following the same steps, and all results are given in Table 5.

From the *MR*, the effect of four methods is as follows: FSDM > SPSS = HDM > BSDM. But from the *MDR*, the effect of four methods is as follows: FSDM > HDM> SPSS > BSDM.

Table 4. Misclassification degree.

NO.	Original Class	Predictive class	Misclassification degree		
			G_1	G_2	G_3
71	G_2	G_3	0.19	0.00	0.28
84	G_2	G_3	1.10	0.00	0.37
134	G_3	G_2	0.44	0.31	0.00

Table 5. Misclassification analysis.

Rule	Discriminant method			
	SPSS	BSDM	FSDM	HDM
MR	2.00%	4.00%	1.33%	2.00%
MDR	1.84%	3.61%	1.04%	1.53%

Table 6. Misclassification analysis of discriminant model.

Data set	Discriminant method			
	SPSS	BSDM	FSDM	HDM
Balance Scale	11.84%[a]/10.10%[b]	30.72%[a]/25.08%[b]	31.36%[a]/25.52%[b]	20.16%[a]/18.13%[b]
Banknote Authentication	2.33%[a]/ 2.33%[b]	2.33%[a]/ 2.33%[b]	2.33%[a]/ 2.33%[b]	0.80%[a]/ 0.80%[b]
Breast Tissue	25.47%[a]/21.06%[b]	30.13%[a]/22.89%[b]	47.17%[a]/28.02%[b]	33.02%[a]/24.98%[b]
Verbetra 2c	14.19%[a]/14.19%[b]	19.36%[a]/19.36%[b]	19.36%[a]/19.36%[b]	23.23%[a]/23.23%[b]
Verbetra 3c	18.17%[a]/17.45%[b]	19.36%[a]/18.78%[b]	31.29%[a]/27.34%[b]	26.45%[a]/23.89%[b]
Ecoli	11.31%[a]/ 8.17%[b]	15.18%[a]/ 9.84%[b]	41.67%[a]/28.82%[b]	17.26%[a]/ 8.57%[b]

[a] Misclassification ratio; [b] Misclassification degree ratio.

As can be seen from the above two results, the effect of selecting models with *MR* and *MDR* is basically the same. However, SPSS and HDM have the same *MR*, it is difficult to estimate the effect of two methods. But from the *MDR*, it is easy to know that HDM is superior to SPSS. In addition, results given in Table 4 indicate the *MD* of each misclassification sample is different, and it is hoped that they would be kept as small as possible. Results given in Table 5 indicate the values of *MDR* are less than the corresponding values of *MR*, which shows the misclassification degree of each sample is less than the corresponding average misclassification degree in each method, and if the threshold of *MR* is set to 15% (this value can be set according to actual problem), four methods have achieved good effect, and the corresponding discriminant model has a relatively small misclassification degree.

Therefore, compared with *MR*, *MDR* has several potential advantages: (1) Reflect the difference between the misclassification samples and each class. (2) Embody the relationship between the *MD* of the misclassification sample and its *AMD*, that is, when the ratio of the *MD* of the misclassification sample to its *AMD* is smaller, the value of *MDR* is smaller. (3) Since the *MR* treats the importance metrics of misclassification samples equally and ignores the differences between them, the *MDR* can better measure the effectiveness of the models.

Similar to misclassification analysis of the Iris data set, the results of other data sets are given in Table 6.

Results given in Table 6 indicate the discrimination results of each method are often different, and the single discriminant method is unlikely superior to other discriminant methods in any case. Therefore, for a given practical problem, it is important to choose a suitable method to establish a discriminant model. On the whole, the SPSS method achieves a better effect than the other three methods regardless of the classification performance measured by *MR*, or the classification performance measured by *MDR*. However, for a certain data set, if the *MDR* is greater than the corresponding *MR*, the *MD* of the misclassification samples is greater than their average *MD*.

As shown in the numerical experiments above, a good discriminant model should make its *MR* and its *MDR* as small as possible. In practical applications, for a single discriminant model, if the threshold of *MR* is set, then its performance can be measured by the comparative analysis of *MR* and *MDR*. For multiple discriminant models, the best model corresponds to the minimum *MDR*.

§4 Conclusion

This paper has extended the criterion of the *MR* of discriminant model and presented the *MDR* of discriminant model. In most practical applications, the misclassification cost of each sample is often not equal. Although the misclassification cost of the sample is difficult to determine, this paper overcomes this drawback through the *MD* of the sample. To select a better discriminant model, the criterion of *MDR* has been established by the *MR* and *MD* of the samples. Numerical experiments on illustrative examples indicate that the performance of discriminant model can be measured by the comparative analysis of *MR* and *MDR*, and the proposed method is helpful to select a better discriminant model.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

[1] R A Fisher. *The use of multiple measurements in taxonomic problems*, Annals of Human Genetics, 1936, 7(2): 179-188.

[2] S Shinmura. *New theory of discriminant analysis*, In: new theory of discriminant analysis after R Fisher, Springer, Singapore, 2016.

[3] F L Song, P Lai, B H Shen, G S Cheng. *Variance ratio screening for ultrahigh dimensional discriminant analysis*, Communications in Statistics Theory and Methods, 2018, 47(24): 6034-6051.

[4] Y F Li, J Lei. *Sparse subspace linear discriminant analysis*, Statistics, 2018, 52(4): 782-800.

[5] A Hidaka, K Watanabe, T Kurita. *Sparse discriminant analysis based on estimation of posterior probabilities*, Journal of Applied Statistics, 2019, 46(15): 2761-2785.

[6] S C Chen, D H Li. *Modified linear discriminant analysis*, Pattern Recognition, 2005, 38(3): 441-443.

[7] A B Ji, H J Qiu, M H Ha. *Fisher discriminant analysis based on choquet integral*, Applied Mathematics-A Journal of Chinese Universities, 2009, 24(3): 348-352.(in Chinese)

[8] L W Huang, L T Su. *Hierarchical discriminant analysis and Its application*, Communications in Statistics - Theory and Methods, 2013, 42(11): 1951-1957.

[9] X Z Xu, C W Huang, Y Jin, C Wu, L Zhao. *Speech emotion recognition using semi-supervised discriminant analysis*, Journal of Southeast University (English Edition), 2014, 30(1): 7-12.(in Chinese)

[10] L W Huang. *Modified Hybrid Discriminant Analysis Methods and Their Applications in Machine Learning*, Discrete Dynamics in Nature and Society, 2020, DOI: 10.1155/2020/1512391.

[11] E K Tang, P N Suganthan, X Yao, A K Qin. *Linear dimensionality reduction using relevance weighted LDA*, Pattern Recognition, 2005, 38(4): 485-493.

[12] W H Yang, D Q Dai, H Yan. *Feature extraction and uncorrelated discriminant analysis for high-dimensional data*, IEEE Transactions on Knowledge and Data Engineering, 2008, 20(5): 601-614.

[13] Q Zhang, H S Wang. *On BIC's selection consistency for discriminant analysis*, Statistica Sinica, 2011, 21(2): 731-740.

[14] J Pacheco, S Casado, S Porras. *Exact methods for variable selection in principal component analysis: Guide functions and pre-selection*, Computational Statistics and Data Analysis, 2013, 57(1): 95-111.

[15] R A Mcdonald. *The mean subjective utility score, a novel metric for cost-sensitive classifier evaluation*, Pattern Recognition Letters,2006, 27(13): 1472-1477.

[16] S R Pan, J Wu, X Q Zhu. *CogBoost: boosting for fast cost-sensitive graph classification*, IEEE Transactions on Knowledge and Data Engineering, 2015, 27(11): 2933-2946.

[17] A C Bahnsen, D Aouada, B Ottersten. *Ensemble of example-dependent cost-sensitive decision trees*, Expert Systems with Applications, 2015, 42(19): 6609-6619.

[18] P Cao, D Zhao, O Zaiane. *An optimized cost-sensitive SVM for imbalanced data learning*, In: Advances in Knowledge Discovery and Data Mining, Lecture Notes in Computer Science, 2013, 7819: 280-292.

[19] L W Huang. *A selection method of discriminant model*, Journal of Jiangxi University of Science and Technology, 2013, 34(1): 96-99. (in Chinese)

[20] D Dua, E Karra Taniskidou. *UCI machine learning repository, Irvine, CA: university of california*, school of information and computer science, http://archive.ics.uci.edu/ml.

[1]College of Mathematics and Computer science, Quanzhou Normal University, Quanzhou 362000, China.

[2]Key Laboratory of Financial Mathematics, Putian University, Putian 351100, China.

Email: livern@126.com