Comparison of Multinomial Classification Rules

Ana M. Pires and João A. Branco

Department of Mathematics
Technical University of Lisbon (IST)
Av. Rovisco Pais, 1096 Lisboa Codex, Portugal

Abstract: In this paper two modifications of the full multinomial classification rule for the discrete feature discrimination problem are proposed. One is derived from a likelihood ratio test and the other from the Bayesian predictive approach. A simulation experiment was conducted to compare the performance of the new rules with two other existing rules for the same problem: the usual estimative or plug-in multinomial rule and the Dillon-Goldstein rule which is based on a distance principle. For equal group sample sizes all rules are equivalent, however for different sizes they behave differently. The simulation study has shown that the three alternative rules lead to similar results, in terms of expected actual error rates, which are better than those for the plug-in rule.

Keywords: Classification rules; Discrete data; Discriminant analysis; Multinomial model.

1. Introduction

Suppose that we want to discriminate between two distinct populations or groups, Π_1 and Π_2, on the basis of a vector of discrete random variables $\mathbf{X} = (X_1, \ldots, X_k)$, each assuming a finite number of distinct values, $d_1, \ldots, d_k$. Assume that, conditional on Π_i, $\mathbf{X}$ follows the multinomial distribution, denoted by $f_i(\mathbf{x})$, with

$$s = \prod_{j=1}^{k} d_j$$

possible states and characterized by the (s-dimensional) vector of probabilities $\mathbf{p}_i$ ($i = 1, 2$).

The optimal rule for the classification of an entity of unknown origin with feature vector $\mathbf{x}$ is, assuming equal costs of misclassification and *a priori* probabilities of the populations,

classify in Π_1 if $f_1(\mathbf{x}) > f_2(\mathbf{x})$,

classify in Π_2 if $f_1(\mathbf{x}) < f_2(\mathbf{x})$,

classify randomly if $f_1(\mathbf{x}) = f_2(\mathbf{x})$.

The usual (estimative) sample rule derived from the optimal rule is obtained replacing the unknown probabilities by their maximum likelihood estimates, that is:

classify in Π_1 if $\frac{n_1(\mathbf{x})}{n_1} > \frac{n_2(\mathbf{x})}{n_2}$,

classify in Π_2 if $\frac{n_1(\mathbf{x})}{n_1} < \frac{n_2(\mathbf{x})}{n_2}$,

classify randomly if $\frac{n_1(\mathbf{x})}{n_1} = \frac{n_2(\mathbf{x})}{n_2}$,

where n_i is the number of observations of group i in the training sample and $n_i(\mathbf{x})$ is the corresponding observed absolute frequency of the state defined by $\mathbf{x}$. We will hereafter call this rule the estimative rule or the M rule.

As pointed out by Dillon and Goldstein (1978) one of the undesirable properties of the M rule is the way it treats zero frequencies. If $n_1(\mathbf{x}) = 0$ and $n_2(\mathbf{x}) \neq 0$, a new observation with feature vector $\mathbf{x}$ will be allocated to Π_2 irrespective of the sample sizes n_1 and n_2. In other words the rule does not take into account the fact that the occurrence of a zero frequency in only one group is more probable when the total number of observations is not large and n_1 and n_2 are of different order.

Those authors proposed a rule based on Matusita's distributional distance which is more sensitive to this issue. The rule, subsequently called the Dillon-Goldstein rule or the D rule, is (using the notation $n_i(\mathbf{x}) = n_{ij}$ if $\mathbf{x}$ belongs to state j)

$$\text{classify in } \Pi_1 \text{ if } \frac{[n_{2j}(n_{1j}+1)]^{1/2} + \sum_{k\neq j}(n_{1k}n_{2k})^{1/2}}{[n_{1j}(n_{2j}+1)]^{1/2} + \sum_{k\neq j}(n_{1k}n_{2k})^{1/2}} < \left[\frac{n_2(n_1+1)}{n_1(n_2+1)}\right]^{1/2},$$

in Π_2 if the reverse inequality holds and randomly else. We shall note that if $n_1 = n_2$ the D rule reduces to the M rule and also that Krzanowski (1987) showed their asymptotic equivalence. However, if $n_1 < n_2$ and $n_{1j} = 0$ but $n_{2j} > 0$, the D rule will classify $\mathbf{x}$ in Π_1 only if

$$\sqrt{n_{2j}} < \left[\left(\frac{n_2(n_1+1)}{n_1(n_2+1)}\right)^{1/2} - 1\right] \sum_{k\neq j}(n_{1k}n_{2k})^{1/2}.$$

Goldstein and Dillon (1978, p. 46) present one example where the previous inequality is satisfied ($n_{1j} = 0$, $n_{2j} = 1$, $n_1 = 25$, $n_2 = 275$ and $\sum_{k\neq j}(n_{1k}n_{2k})^{1/2} \simeq 69.28$).

In sections 2 and 3 two different rules are proposed and it is shown that they can also handle the zero frequency problem. In section 4 the results of a simulation study designed to compare the four rules are presented and discussed.

2. The likelihood ratio criterion

This criterion is described in Anderson (1984) but, as far as we know, has been applied only to the multivariate normal model. The basic idea is to derive a rule from a generalized likelihood ratio test for the hypotheses

$$H_0\text{: } \mathbf{x},\, \mathbf{x}_{11},\ldots,\mathbf{x}_{1n_1} \text{ follow } f_1(\mathbf{x}) \text{ and } \mathbf{x}_{21},\ldots,\mathbf{x}_{2n_2} \text{ follow } f_2(\mathbf{x})$$
$$\text{against}$$
$$H_1\text{: } \mathbf{x}_{11},\ldots,\mathbf{x}_{1n_1} \text{ follow } f_1(\mathbf{x}) \text{ and } \mathbf{x},\, \mathbf{x}_{21},\ldots,\mathbf{x}_{2n_2} \text{ follow } f_2(\mathbf{x}).$$

For the multinomial model under consideration the test statistic, which is a function of $\mathbf{x}$, becomes

$$\mathcal{L}(\mathbf{x}) = \frac{\left(1+\frac{1}{n_1(\mathbf{x})}\right)^{n_1(\mathbf{x})}(n_1(\mathbf{x})+1)}{\left(1+\frac{1}{n_2(\mathbf{x})}\right)^{n_2(\mathbf{x})}(n_2(\mathbf{x})+1)} \times \frac{\left(1+\frac{1}{n_2}\right)^{n_2}(n_2+1)}{\left(1+\frac{1}{n_1}\right)^{n_1}(n_1+1)}, \tag{1}$$

where if $n_1(\mathbf{x}) = 0$ ($n_2(\mathbf{x}) = 0$) the numerator (denominator) of the first fraction must be replaced by 1. The classification rule derived is simply (again for equal costs of misclassification and *a priori* probabilities):

classify in Π_1 if $\mathcal{L}(\mathbf{x}) > 1$, in Π_2 if $\mathcal{L}(\mathbf{x}) < 1$ and randomly else.

We will, from now on, call this the likelihood rule or the L rule. Again, for $n_1 = n_2$ the L rule and the M rule are equivalent, because the numerator of the first fraction in (1) is a strictly increasing succession of $n_1(\mathbf{x})$.

The L rule also solves the zero frequency problem. A new observation $\mathbf{x}$ with $n_1(\mathbf{x}) = 0$ will be classified in Π_1 if and only if

$$\left(1+\frac{1}{n_2(\mathbf{x})}\right)^{n_2(\mathbf{x})}(n_2(\mathbf{x})+1) < c$$

where c is the value of the second fraction in (1). In the example of Goldstein and Dillon (1978), referred in the previous section, $n_1(\mathbf{x}) = 0$ and $n_2(\mathbf{x}) = 1$ would also lead to classification in Group 1. In fact, by the L rule $n_2(\mathbf{x}) = 3$ would still lead to classification in Group 1 whereas using the D rule would result on classification in Group 2. We can thus expect a somehow different behaviour of the L and D rules.

3. The predictive approach

If the non-informative conjugate prior distribution for the parameter $\mathbf{p}_i$ of the multinomial model is chosen, that is the Dirichlet distribution with parameter $\boldsymbol{\alpha}= \mathbf{1}$, then the posterior distribution will be a Dirichlet distribution with parameter $\mathbf{z}_i + 1$, where $\mathbf{z}_i = (n_{i1}, \ldots, n_{is})^T$ (note that $\sum_{j=1}^{s} n_{ij} = n_i$). The Dirichlet distribution with parameter $\boldsymbol{\alpha}$ has a density function given by

$$f(\mathbf{p}) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_s)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_s)} \prod_{i=1}^{s} p_i^{\alpha_i - 1}, \;\; 0 < p_i < 1, \;\; \sum_{i=1}^{s} p_i = 1, \alpha_i > 0.$$

Therefore the predictive density is simply

$$\begin{aligned} h_i(\mathbf{x}|\mathbf{z}_i) \;&=\; \int_{\mathbf{p}} p_{ij} \frac{\Gamma(s + n_i)}{\Gamma(1 + n_{i1}) \cdots \Gamma(1 + n_{is})} \prod_{k=1}^{s} p_{ik}^{n_{ik}} d\mathbf{p} = \quad (0 < p_{ij} < 1, \sum_{j=1}^{s} p_{ij} = 1) \\ &=\; \frac{n_{ij} + 1}{n_i + s} = \frac{n_i(\mathbf{x}) + 1}{n_i + s}, \end{aligned}$$

which leads to the predictive rule (or the P rule)

classify in Π_1 if $\frac{n_1(\mathbf{x}) + 1}{n_1 + s} > \frac{n_2(\mathbf{x}) + 1}{n_2 + s}$,

classify in Π_2 if $\frac{n_1(\mathbf{x}) + 1}{n_1 + s} < \frac{n_2(\mathbf{x}) + 1}{n_2 + s}$,

classify randomly if $\frac{n_1(\mathbf{x}) + 1}{n_1 + s} = \frac{n_2(\mathbf{x}) + 1}{n_2 + s}$.

Once again for $n_1 = n_2$ this rule is equivalent to the M rule. The P rule also avoids the zero frequency problem. For instance $n_1(\mathbf{x}) = 0$ and $n_2(\mathbf{x}) < (n_2 + s)/(n_1 + s) - 1$ leads to classification in Π_1.

As far as we know the present rule has not been derived explicitly before although it is implicit in the predictive rule for mixed variables (Vlachonikolis, 1990).

4. Simulation study and discussion

In order to assess the performance of the four rules a small simulation study was designed. Let $\mathbf{X}$ be a random vector with 64 states (the actual values of the variables leading to each state are irrelevant under the multinomial model) and probabilities given by

$$\mathbf{p}_1 = \tfrac{1}{157} \times [\ 0(j = 1, \ldots, 15),\ 1(j = 16, \ldots, 21),\ 2(j = 22, \ldots, 28),\ 3(j = 29, \ldots, 36),$$
$$4(j = 37, \ldots, 43),\ 5(j = 44, \ldots, 54),\ 4(j = 55, \ldots, 58),\ 3(j = 59, 60, 61),\ 2(j = 62, 63),$$
$$1(j = 64)]$$

and

$$p_{2j} = p_{1,64-j+1},$$

and assuming, as previously, equal costs of misclassification and *a priori* probabilities of group membership. The optimal error rate under these conditions is $e_{opt} = 32/157 \simeq 0.2038$, admitting random equiprobable classification when $p_{1j} = p_{2j}$.

Six combinations of sizes (n_1, n_2) were selected: (13,100); (25,100); (50,100); (25,200); (50,200) and (100,200). Remember that with $n_1 = n_2$ and apart from the random classification all rules must give the same results. 100 training samples were generated for each size case. Then for each training sample the four classification rules were computed and the apparent and actual error rates, per group and overall, were evaluated (all these computations, including the random number generation, were carried out by Minitab Macros running on a DEC 7620 computer under Open VMS/AXP). The results obtained, mean and standard deviation over the 100 replications, are reported in Table 1 (apparent error rates) and Table 2 (actual error rates).

(Place Table 1 and Table 2 near this point)

Comparing the results in Table 1 with those in Table 2 it is possible to see that:

- Even though it is well known that the apparent error rate is optimistically biased, the numbers obtained are worse than expected. Note, however, that the bias decreases when n_1 and n_2 increase.

- The apparent error rate for the D rule is larger than for the other rules. This is positive, meaning that for this rule the bias towards the actual error rate is smaller (Dillon and Goldstein, 1978, have also observed this fact in their simulation study).

The comparisons between the rules must be based on the analysis of the actual error rates in Table 2. The overall error rate is the most important but it is also interesting to look at the differences between the error rates per group.

- A common and expected aspect to all the rules is that their performance deteriorates as n_1, n_2 and n_1/n_2 decrease. This meaning that the overall error rate increases and also that the differences between the error rates of the groups are wider. An exception occurs with the D rule for which the last difference increases when n_1/n_2 is kept constant but n_1 and n_2 increase.

- The overall error rate of the M rule is the worst in all cases but one. However the difference to the best result is reduced when n_1, n_2 and n_1/n_2 increase. A paired comparison (necessary because of the simulation design) indicates that for $n_1 = 100$ and $n_2 = 200$ that difference is no longer significant, the actual error rate of the M rule was smaller or equal to that of the P rule 35 times out of 100. For $n_1 = 13$ and $n_2 = 100$ the corresponding figure is only 2.

- The P, L and D rules all lead to similar overall error rates — perhaps with a slight advantage of the P rule, though not significant — except in the first case where the L rule is clearly the worst and the D rule the best.

- In terms of the difference between the group error rates the worst rules are: the L rule for the smallest sample size; the D rule in the three cases where $n_2 = 200$, and the M rule in the remaining.

- It is also observed that for the L rule the highest group error rate is always for Π_2 while for the other rules it is always for Π_1 (except in one case for each rule).

It is worth noting that the different behaviour of the four classification rules in this empirical study is not due only to the zero frequency problem, but also to a diversified handling of other small frequencies, especially where ties occur. For instance, when $n_1 = 50$ and $n_2 = 100$ the classification scheme for selected values of $n_1(\mathbf{x})$ and $n_2(\mathbf{x})$ is the following

$n_1(\mathbf{x})$	$n_2(\mathbf{x})$	M	P	L
0	0	random	Π_1	Π_1
0	≥ 1	Π_2	Π_2	Π_2
1	2	random	Π_2	Π_1
2	4	random	Π_2	Π_1

These assignments apparently justify the different behaviour of the P and L rules regarding the group error rates. Similar situations occur with the other sample sizes (the D rule cannot be included in this comparison because the allocation rule does not depend solely on $n_1(\mathbf{x})$ and $n_2(\mathbf{x})$).

5. Conclusions

It is possible to conclude from this study that discrete discriminant analysis with sparse and unbalanced data is more successful under the multinomial model using one of the discussed alternatives to the usual estimative rule.

When sparseness is very high then the Dillon-Goldstein rule is the best choice, but in this situation perhaps the wisest thing to do is to abandon the full multinomial model and to adopt a more parsimonious one (a lot of such models are referred, for instance, in McLachlan, 1992, Ch. 7).

For mild sparseness we would recommend the predictive rule because of its simplicity and good behaviour.

References

ANDERSON, T.W. (1984), *An Introduction to Multivariate Statistical Analysis* (2nd ed.), New York: John Wiley & Sons.

DILLON, W.R., and GOLDSTEIN, M. (1978), "On the Performance of Some Multinomial Classification Rules," *Journal of the American Statistical Association, 73*, 305-313.

GOLDSTEIN, M., and DILLON, W.R. (1978), *Discrete Discriminant Analysis*, New York: John Wiley & Sons.

KRZANOWSKI, W.J. (1987), "A Comparison Between Two Distance Based Discriminant Principles," *Journal of Classification, 4*, 73-84.

MCLACHLAN, G.J. (1992), *Discriminant Analysis and Statistical Pattern Recognition*, New York: John Wiley & Sons.

VLACHONIKOLIS, I. G. (1990), "Predictive Discrimination and Classification with Mixed Binary and Continuous Variables," *Biometrika, 77*, 657-662