Jour. Ind. Soc. Ag. Statistics
Vol. XXXIX, No. 2, (1987), pp. 205-215

BAYES DISCRIMINATORY ANALYSIS IN CULLING DAIRY COWS FOR BREEDING

V. K. BHATIA, MAHESH KUMAR,
P. K. MALHOTRA and PREM NARAIN
I.A.S.R.I., New Delhi-110012

(Received : February, 1987)

SUMMARY

Use of Bayes discriminatory analysis in culling dairy cows for breeding has been advocated not only for correct classification but also for screening of an optimal subset of possible informative variables. The approach is probabilistic i.e. posterior probabilities are assigned to a cow on the basis of the values observed on various production and reproduction variables.

The statistical model used is largely based on the assumption of independency between the variables, but one model parameter 'global association factor' is added in order to take dependency into account. The stepwise selection strategy has been used. The quadratic selection criterion is used in order to decide in each selection step which variable should be added. The use of this procedure has been illustrated with good amount of success in three different culling processes of dairy cattle of Military Dairy Farm, Ambala. It has been observed that their lactation yields are one of the important characteristics in arriving at the correct decisions.

Keywords : Bayes discriminant analysis; Prior and Posterior probabilities; Error rate; Allocation matrices.

Introduction

In animal breeding, genetic improvement in the productivity is achieved to some extent by retaining superior cows or disposing/culling unproductive cows. In other words these two terms 'retention' and 'culling' are part and parcel of the broad term longevity of the cow in a given herd. There are mainly two significant aspects of longevity in cattle

breeding i.e. economic value and culling of unproductive cows for replacement by heifers. The milk producer makes continual decisions as to which cow he would cull to make way for heifer replacements. In such decisions, he needs careful examination of the various production and other characteristics of an animal at the end of the lactation for deciding whether the animal is to be retained or culled. In past, various statistical methodologies have been proposed for exploring the variables affecting the culling process of dairy cattle. Robertson [8] had shown that culling of a cow could be regarded on the basis of truncation selection using a culling variate of which one component is the milk yield of the cow. Narain and Bhatia [7] had studied the relationship between yield characteristics and survival of a cow in the herd. These studies however do not serve the very purpose of classifying the animal into two broad classes 'retention' and 'culling' on the basis of characteristics observed at the end of a particular order of lactation. In the present study use of Bayes discriminatory analysis following Habbema and Gelpke [2] is advocated for such a problem. The study not only helps in classifying the cow into two different classes but also helps in selecting an optimal subset from a set of possible informative variables affecting the culling process. The use of study has been illustrated with the help of the data already collected for the culled animals at different orders of lactation from the Military Dairy Farm, Ambala.

2. Bayes Discriminant Analysis

In culling process studies, a cow has to be allocated into either 'culled' or 'retained' class after the completion of the particular order of lactation. If we do not measure any variables at all, on this cow, allocation is to be done according to the available prior knowledge, expressed quantitatively as prior probabilities.

Symbolically $P(C_j)$ = prior probability of class C_j; for $j = 1, 2$

When a vector of observations X on one or more variables is available for a cow whose actual category has to be identified, X provides probabilistic information. This second type of probability is denoted by

$P(X/C_j)$ = The probability of observation vector X for class C_j

These probabilities will have to be estimated from the observation on the sample of reference animals from each of the two classes. The prior probabilities have to be combined with the probabilities of the observations in order to get the probability of a class C_j, given the observation X, called posterior probability and denoted by

$P(C_j/X)$ = The posterior probability of class C_j

The estimated posterior probability is calculated from the prior and estimated observation probabilities by Bayes' theorem :

$$\hat{P}(C_j/X) = \frac{\hat{P}(C_j)\ \hat{P}(X/C_j)}{\sum_{m=1}^{2} \hat{P}(C_m)\ \hat{P}(X/C_m)} \quad \text{for} \quad j = 1, 2$$

Once the prior probabilities are specified, the posterior probabilities only depend on the probability estimates $\hat{P}(X/C_j)$ which have to be inserted in Bayes theorem.

It is assumed that an observation vector X consists of measurement on p variables $(x_1, x_2, \ldots, x_i, \ldots, x_p)$. Assuming all the p variables are multinomially distributed, one of the most simple statistical model for describing such an observation X, is the independent multivariate multinomial model i.e.

$$\hat{P}(X/C_j) = \prod_{i=1}^{p} P(x_i/C_j)$$

with the independent assumption as the justification that the probability for the total observation vector X, is just the product of the probabilities of the p variables. The probability of the single variable x_i is estimated by the fraction of reference cows from class C_j with value x_i for variable i or in the notation

$$\hat{P}(x_i/C_j) = \frac{1}{N_j} \sum_{t=1}^{N_j} I(x_{jti} = x_i)$$

With N_j the size of the sample of reference cows from class C_j, t the index running over these N_j reference cows, x_{jti} the ith component of X_{jt} the observation vector for the tth reference cow of class C_j, and $I(\cdot)$ the indicator function with as values $I(\text{true}) = 1$ and $I(\text{false}) = 0$. The assumption of independence between the variables allows for a generalizing modification which incorporates the interdependence between the variables in a general way, by means of a global association factor (B) (Hilden and Bjerregaard [4])

$$\hat{P}(X/C_j) = \left[\prod_{i=1}^{p} P(x_i/C_j) \right]^{B} \quad \text{with } (0 < B \leqslant 1)$$

The quantity B is interpreted as the proportion of independent information contained in the variables.

A generalization of the observation probability is also possible by allowing for a 'flattening constant' also called a Bayesian correction fac-

tor (Fienberg and Holland [1]) as

$$\hat{P}(x_i/C_j) = \frac{A/l_j + \sum_{i=1}^{N_j} I(x_{jit} = x_i)}{N_j + A} \quad \text{for } A > 0$$

The flattening constant 'A' can be interpreted as a device for avoiding probability estimates of zero. Such zero estimates make classes not just improbable but impossible (posterior probability = 0). Zero estimates in the usual formula will occur regularly when the sample sizes N_j are not large compared to the l_i-values (number of categories for variable i). The larger the value of A, the more probability estimates for the l_i categories will be pulled towards a common value $1/l_i$.

3. The Selection Criteria

The construction of the selection criteria used in the present study for screening the variables as well as discriminatory between the two classes 'culled' and 'retained' is defined below using the concept of error rate. The necessary steps are as follows.

Step 1. A 'penalty-score' q_{jt} is calculated for each reference cow $t(t = 1, 2, \ldots, N_j)$ from each class $C_j(j = 1, 2)$. The penalty score reflects the discrepancy between the actual class C_j and the posterior probability assigned to this class. The penalty function q is such that a zero penalty is incurred when 100% probability is assigned to the actual class, and a maximal penalty when zero probability is assigned to the actual class.

Step 2. An average penalty score is calculated for each of the two classes, by averaging the penalty scores of the N_j reference cows.

Step 3. The overall criterion score Q is obtained by weighing the average penalty score for each class with its prior probability.

$$Q = \sum_{j=1}^{2} P(C_j) \left[\frac{1}{N_j} \sum_{j=1}^{N_j} q_{jt} \right]$$

In order to decrease the complexity of the formulae the notation P_{tjr} will be used for $P(C_r/X_{jt})$ the posterior probability that cow t from the class j falls in category r (Thus P_{tjj} is probability assigned to the actual class).

Using this, one can define Q_1 (error rate) as one of the selection criterion

$$Q_1 = \sum_{j=1}^{2} P(C_j) \left[\frac{1}{N_j} \sum_{t=1}^{N_j} I\,(P_{tjj} \neq \max_r P_{tjr}) \right]$$

with the associate penalty function

$$q_1 = I\,(P_{tjj} \neq \max_r P_{tjr})$$

or to say it with words a penalty of 1 when the cow does not have the highest probability assigned to her actual class.

Since error rate is a real performance measure but it is not very sensitive to changes in posterior probabilities, thus the second criterion i.e. logarithmic criterion is defined as

$$Q_2 = - \sum_{j=1}^{2} P(C_j) \left[\frac{1}{N_j} \sum_{j=1}^{N_j} \log_e P_{tjj} \right]$$

with penalty function as $q_2 = - \log_e P_{tjj}$.

Criterion Q_2, is based on a continuous penalty function of the probability assessed to the actual class. The logarithmic criterion possesses a fundamental statistical optimality property (Mosteller and Wallace [6]). It has, however, one very serious drawback from an applied point of view : a zero probability assessment to the actual class is penalized with an infinity penalty score. This is the reason Hilden *et al.* [5] have proposed a modification as the $\in$-modified logarithmic scoring rule

$$Q_3 = - \sum_{j=1}^{2} P(C_j) \left[\frac{1}{N_j} \sum_{t=1}^{N_j} \log_e w_{tjj} + \varepsilon \sum_{r \neq j} \log_e (w_{tjr}/\varepsilon) \right]$$

with

$$w_{tjr} = (1 - \varepsilon)\, P_{tjr} + \varepsilon$$

The corresponding penalty function equals

$$q_3 = - [\log_e w_{tjr} + \varepsilon \sum_{r \neq j} \log_e (w_{tjr}/\varepsilon)]$$

The logarithmic criterion Q_2 only takes the probability assigned to the actual class into account, not the distribution of the wrongly assessed probability mass over the remaining classes. This distribution is taken

into account by the quadratic criterion

$$Q_4 = \sum_{j=1} P(C_j) \left[\frac{1}{N_j} \sum_{t=1}^{N_j} \left((1 - P_{tjj})^2 + \sum_{r \neq j} P_{tjr}^2 \right) \right]$$

with the corresponding penalty function

$$q_4 = (1 - P_{tjj})^2 + \sum_{r \neq j} P_{tjr}^2$$

4. The Selection Process

On the basis of selection criterion Q the stepwise forward selection process may start. In the first step p allocation rules are considered, each based on one of the variables x_i, $i = 1, 2, \ldots, p$. The score $Q(x_i)$ is estimated for these p rules. Variable x_{i_1} is selected when

$$Q(x_{i_1}) = \min_{i=1, 2, \ldots, p} Q(x_i).$$

In the second step another variable x_i is selected such that this new variable together with x_{i_1} gives a minimal criterion score for all pairs of variables containing x_{i_1}.

$$Q(x_{i_1}, x_{i_2}) = \min_{\substack{i=1, 2, \ldots, p \\ i \neq 1}} Q(x_{i_1}, x_i)$$

This process can be continued for p steps; we then arrive at an ordering of all the variables $(x_{i_1}, x_{i_2}, \ldots, x_{i_p})$. It is also possible to impose a stopping criterion on this selection process i e. choosing some threshold value Δ and stop after step r if after this step for the first time

$$Q(x_{i_1}, \ldots, x_{i_r}) - Q(x_{i_1}, \ldots, x_{i_{r+1}}) \leqslant \Delta$$

Using the error rate, the criterion simplifies to

$$\hat{P}_r(\text{error}) - \hat{P}_{r+1}(\text{error}) \leqslant \Delta$$

That is, stop the selection after r steps if the decrease in estimated probability of misallocation is less than or equal to Δ. In most applications, $\Delta = 0$ is a good stopping criterion.

5. Allocation Matrices

The criterion score reflects discriminatory performance in a very concise way. In most instances one will wish to have more detailed descrip-

tive information about the quality of discrimination. The first thing is to inspect the posterior probabilities because they contain the complete information about the quality of discrimination obtained by the selected set of variables An alternative, very informative way of summarizing these posterior probabilities is by means of allocation matrices (Habbema *et al.* [3]).

6. Application of this Procedure on Culling Process Studies

In order to illustrate the use of Bayes discriminatory analysis, the data from, Military Dairy Farm located at Ambala, for cows culled at different orders of lactation have been taken into account.

Starting with 247 crossbred cows, 47 were culled after the 1st lactation (200 retained), 50 after the 2nd lactation (150 retained) and 23 after the 3rd lactation (127 retained). Information recorded are as follows :

Level of Exotic Inheriance (LOEI); Age at : 1st calving (AFC), 2nd calving (ASC), and 3rd calving (ATC); Lactation length : first (FLL), second (SLL) and third (TLL); Lactation yield : first (FLY), second (SLY) and third (TLY); Milk yield per day of lactation : 1st lactation (FLY/LL), 2nd (SLY/LL) and 3rd (TLY/LL); Calving interval : first (FCI), second (SCI) and third (TCI); Milk yield per day of calving interval : first (FLY/CI), second (SLY/CI) and third (TLY/CI); Age at completion of Lactation : 1st lactation (ACL1), 2nd (ACL2) and 3rd (ACL3); Sum of 1st and 2nd Lactation yield (LY12); Sum of 1st, 2nd and 3rd lactation yield (LY/123); Second lactation milk yield per day of age at 3rd calving (SLY/ATC); Milk yield per day of age at: completion of 1st lactation (FLY/ACL1), 2nd calving (FLY/ASC), completion of 2nd lactation (LY12/ACL2), and 3rd calving (SLY12/ATC).

The analysis with quadratic criterion was carried out at three different stages by taking into consideration all possible variables into account. To examine the validation of this procedure, allocation matrices were worked out after adding one variable at each step and results are presented in Table 1 to 3 in terms of number of cows classified and misclassified into their actual class for the value of global association factor as 0.8.

7. Discussion

On looking into the results on allocation matrices (Table 1) obtained on the basis of first culling process, it has been observed that first lactation milk yield (FLY) is the most important characteristics for deciding the fate of an animal for its retention in the herd. This trait alone is

TABLE 1—PRIORITY OF VARIABLES SCREENED FOR CLASSIFYING THE ANIMALS INTO CULLED/RETAINED GROUPS BASED ON POSTERIOR PROBABILITIES AFTER COMPLETION OF FIRST LACTATION

Variable screened	*Error Rate*	*Classification based on posterior probabilities*			
		Out of 47 culled animals		*Out of 200 retained animals*	
		Culled	*Retained*	*Culled*	*Retained*
FLY	0.398	37 (78.7)	10 (21.3)	77 (38.5)	123 (61.5)
AFC	0.361	33 (70.2)	14 (29.8)	50 (25.0)	150 (75.0)
FCI	0.346	33 (70.2)	14 (29.8)	47 (23.5)	153 (76.5)
LOEI	0.332	34 (72.3)	13 (27.7)	44 (22.0)	156 (78.0)
FLY/CI	0.327	35 (74.5)	12 (25.5)	45 (22.5)	155 (77.5)
FLL	0.326	34 (72.3)	13 (27.7)	39 (19.5)	161 (80.5)

Figures in parenthesis denote the % classification.

TABLE 2—PRIORITY OF VARIABLES SCREENED FOR CLASSIFYING THE ANIMALS INTO CULLED/RETAINED GROUPS BASED ON POSTERIOR PROBABILITIES AFTER COMPLETION OF SECOND LACTATION

Variable screened	*Error Rate*	*Classification based on posterior probabilities*			
		Out of 50 culled animals		*Out of 150 retained animals*	
		Culled	*Retained*	*Culled*	*Retained*
SLY	0.415	38 (76.0)	12 (24.0)	61 (40.7)	89 (59.3)
FLY	0.362	39 (78.0)	11 (22.0)	40 (26.7)	110 (73.3)
SCI	0.334	39 (78.0)	11 (22.0)	32 (21.3)	118 (78.7)
LOEI	0.311	37 (74.0)	13 (26.0)	28 (18.7)	122 (81.3)
AFC	0.293	37 (74.0)	13 (26.0)	30 (20.0)	120 (80.0)
SLY/CI	0.282	38 (76.0)	12 (24.0)	26 (17.3)	124 (82.7)
ASC	0.276	42 (84.0)	8 (16.0)	22 (14.7)	128 (85.3)
FLY/ASC	0.273	40 (80.0)	10 (20.0)	27 (18.0)	123 (82.0)

Figures in parenthesis denote the % classification.

TABLE 3—PRIORITY OF VARIABLES SCREENED FOR CLASSIFYING THE ANIMALS INTO CULLED/RETAINED GROUPS BASED ON POSTERIOR PROBABILITIES AFTER COMPLETION OF THIRD LACTATION

Variable screened	*Error Rate*	*Classification based on posterior probabilities*			
		Out of 23 culled animals		*Out of 127 retained animals*	
		Culled	*Retained*	*Culled*	*Retained*
TLL	0.402	13 (56.5)	10 (43.5)	22 (17.3)	105 (82.7)
TLY	0.345	13 (56.5)	10 (43.5)	17 (13.4)	110 (86.6)
LOEI	0.315	15 (65.2)	8 (34.8)	26 (20.5)	101 (79.5)
TCI	0 285	15 (65.2)	8 (34.8)	18 (14.2)	109 (85.8)
AFC	0.262	16 (69.6)	7 (30.4)	17 (13.4)	110 (86.6)
SLY	0.243	17 (73.9)	6 (26.1)	15 (11.8)	112 (88.2)
FLY	0.229	18 (78.3)	5 (21.7)	17 (13.4)	110 (86.6)
SCI	0.214	19 (82.6)	4 (17.4)	12 (9.4)	115 (90.6)
FLL	0.204	19 (82.6)	4 (17.4)	11 (8.7)	116 (91.3)
ACL3	0.195	20 (87.0)	3 (13.0)	9 (7.1)	118 (92.9)
FCI	0.184	19 (82.6)	4 (17.4)	11 (8.7)	116 (91.3)
ASC	0.181	19 (82.6)	4 (17.4)	12 (9.4)	115 (90.6)
ACL2	0.179	21 (91.3)	2 (8.7)	13 (10.2)	114 (89.8)
ACLI	0.179	21 (91.3)	2 (8.7)	14 (11.0)	113 (89.0)

Figures in parenthesis denote % classification.

sufficient as 78.7% of animals culled could be classified correctly. On further adding of variables no doubt, there is slight decline in the percent correct classification of culled animals but there is substantial improvement of the order of 19% of correct classification of the retained group of animals. Table 2 indicates that SLY and FLY together are responsible for the correct classification of culled and retained group of animals to the extent of 78.0% and 73.3% respectively. In all SLY, FLY, SCI, LOEI, AFC, SLY/CI and ASC can improve this percentage of correct classification to 84.0 and 85.3 respectively. Table 3 reveals that TLL and TLY are sufficient for deciding the fate of retained group of animals to the extent of 86.6% correctly but for correct decisions for culling purposes we need large number of traits such as TLL, TLY, LOEI, TCI, AFC, SLY, FLY, SCI, FLL and ACL3 which in all can account 87.0% of the total decisions.

8. Conclusions

From the results of the above analyses at different culling processes, one can very well advocate the use of this objective way of evaluating the merit of the cow rather than just looking into one or two characteristics. This approach not only helps in correct classification but can also screen an optimal subset of variables required for correct classification. The results obtained from the reference animals can very well be applied on new set of test animals once the estimates of the observation and prior probabilities have been estimated or are known with the good amount of accuracy and precision. This may add to the precision of posterior probabilities as well as to the whole system of culling process at different stages of life of dairy cattle.

REFERENCES

[1] Fienberg, S. E. and Holland, P. W. (1972) : On the choice of flattening constants for estimating multinomial probabilities, *J. Multivar. Anal.*, **2** : 127-134.

[2] Habbema, J. D. F. and Gelpake, G. J. (1981) : A Computer programme for selection of variables in diagnostic and Prognostic Problems, *Comput. Prog. Biomed.* 13 : 251-270.

[3] Habbema, J. D. F., Hilden, J. and Bjerregaard, B. (1978) : The measurement of performance in probabilistic diagnosis I. The problem, descriptive tools and measures based on classification matrices, *Math. Inform. Med.*, **17** : 217-226.

[4] Hilden, J. and Bjerregaard, B. (1976) : Computer-aided diagnosis and the atypical case. *In* : F. T. de Dombal and F. Gremy (eds.), *Decision Making and Medical Care : Can Information Science Help?* North-Holland, Amsterdam, pp. 165-174.

[5] Hilden, J., Habbema, J. D. F. and Bjerregaard, B. (1978) : The measurement of performance in probabilities diagnosis. II. Trustworthiness of the exact values of the diagnostic probabilities, *Math. Inform. Med.*, **17** : 227-237.

[6] Mosteller, F. and Wallace, D. L. (1964) : *Inference and Disputed Authorship. The Federalist*, Addison-Wesley, Reading MA.

[7] Narain, P. and Bhatia, V. K. (1984) : *Some Statistical Aspects of Culling Patterns in Indian Herds of Dairy Cattle*, IASRI Silver Jubillee Souvenir (1959-1984), pp. 161-173.

[8] Robertson, A. (1966) : A mathematical model of culling in dairy cattle, *Animal Production*, 8 (2) : 241-252.