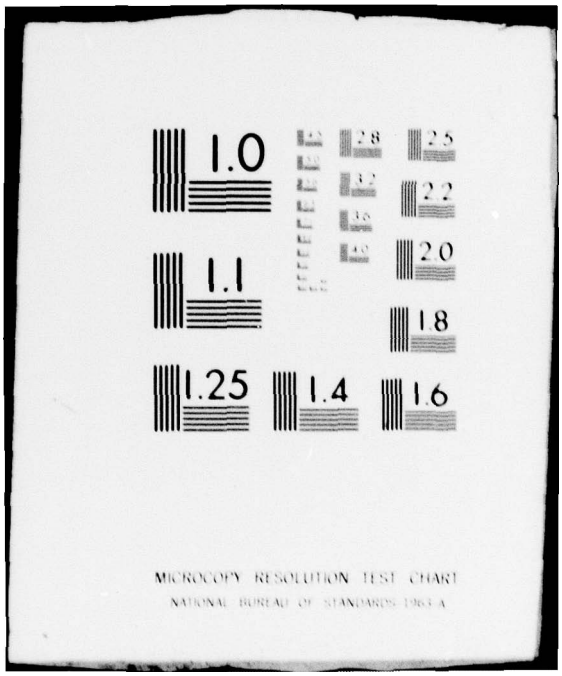
AD-A074 520 TEXAS UNIV AT AUSTIN CENTER FOR CYBERNETIC STUDIES F/G 9/4
INFORMATION THEORETIC STEPWISE SELECTION OF DISCRIMINATING DISC--ETC(U)
MAY 79 P BROCKETT, P HAALAND, A LEVINE N00014-75-C-0616
UNCLASSIFIED CCS-340 NL

END
DATE FILMED
11-79
DDC


1.0
1.1
1.25
1.4
1.6
2.8
2.5
3.2
2.2
3.6
2.0
4.0
1.8
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963-A

(12) LEVEL II

# CENTER FOR CYBERNETIC STUDIES

The University of Texas
Austin, Texas 78712

(12) LEVEL II

(9) (14) Research Report CCS-340

(6) INFORMATION THEORETIC STEPWISE SELECTION OF DISCRIMINATING DISCRETE VARIABLES.

by

(10) P. Brockett*
P. Haaland**
A. Levine***

(11) May 1979

(12) 15



*The University of Texas at Austin
**University of Miami
***Tulane University

(15) This research was partly supported by Project NR047-021, ONR Contracts N00014-75-C-0616 and N00014-75-C-0569 with the Center for Cybernetic Studies, The University of Texas at Austin. Reproduction in whole or in part is permitted for any purpose of the United States Government.

CENTER FOR CYBERNETIC STUDIES

A. Charnes, Director
Business-Economics Building, 203E
The University of Texas at Austin
Austin, TX 78712
(512) 471-1821

406197

**Abstract**

Often the scientist is faced with a large number of categorical variates which are of potential use in discriminating between two pregiven groups of objects. For example, an investor may wish to assign a particular firm to one of two possible risk groups based upon certain known characteristics of the firm (liquid to fixed asset ratio, etc.), or an engineer might wish to determine which of two models best describes a particular situation based upon the observed characteristics of situation. This is the general problem of variable selection in discriminant analysis. When obtaining and processing the numerous variables is expensive, one must select a "best subset" of variables which incorporates as much information for discriminating as possible. If time is also a factor, a stepwise procedure is mandated. We propose such a stepwise procedure here based upon information theoretic considerations.

**Key Words**

Variable selection
Information theory
Chi-squared distribution

## §1 Introduction

Variable selection in discriminant analysis is an important problem. For continuous multivariate normal population models, one can use procedures such as: a) selecting the variables with largest linear discriminant function coefficients, b) selecting the variables according to the relative sizes of the t-statistic for testing equality of means for the two groups on the variates, and c) a stepwise regression technique for selecting variables which maximally reduce the residual sum of squares. Logistic regression techniques can also be used (see Weiner and Dunn (1966) or Goldstein and Dillon (1978) for further discussions). For discrete binary data Goldstein and Dillon (1978) summarized several available methods, and show by example that their information approach (Goldstein and Dillon, 1977) performs quite well relative to the alternatives. The method we propose is related to, and extends the Goldstein-Dillon procedure.

Suppose $f_1(x)$ and $f_2(x)$ are the generalized densities with respect to some dominating measure $\lambda$ for populations 1 and 2 respectively. If $\Pi_i$ is the apriori chance of group $i$ membership, then $\ln \Pi_1/\Pi_2$ is the log-odds ratio in favor of group one membership. If an observation $x$ is made, then Bayes Theorem may be utilized to calculate the new log-odds ratio as $\ln \Pi_1 f_1(x)/\Pi_2 f_2(x)$. The difference between these two log-odds is a measure of the change in odds assessment due to having observed $x$. A simple calculation shows this is $\ln f_1(x)/f_2(x)$ which is known as the information gain in favor of group one membership due to the knowledge of $x$. (Kullback, 1959). The expected information gain for a group one member is

$$I(f_1|f_2) = \int f_1(x) \ln(f_1(x)/f_2(x)) \lambda(dx)$$

and is known as the directed information divergence between populations one and two

for discriminating in favor of group one membership. The measure $I(f_1|f_2)$ is asymmetric, however, a symmetric divergence can be formed by

$$J(f_1,f_2) = I(f_1|f_2) + I(f_2|f_1)$$

and has the interpretation of being the amount of information for discriminating between the two densities $f_1$ and $f_2$. For our purposes, we shall be concerned with categorical variates, and hence we take $\lambda(dx)$ as counting measure and obtain

$$I(p^{(1)}|p^{(2)}) = \sum_j p_j^{(1)} \ln p_j^{(1)}/p_j^{(2)}$$

and

$$J(p^{(1)}, p^{(2)}) = \sum_j (p_j^{(1)} - p_j^{(2)}) \ln (p_j^{(1)}/p_j^{(2)})$$

where $p_j^{(i)}$ is the probability a member of population $i$ responds in the $j^{th}$ category.

The general philosophy we shall use in choosing variables for inclusion is to include the variates in order of decreasing information furnished by the variate for discriminating between the groups. When the amount of additional information furnished by a variate falls below some given level $\gamma$, the selection scheme terminates.

When the precise probability distributions $p^{(1)}$ and $p^{(2)}$ are known this scheme can be implemented as stated. The more usual case, however, is where $p^{(1)}$ and $p^{(2)}$ are not precisely known and must be estimated from sample data from the groups. This situation is discussed in the next section.

§2 A sample based selection scheme.

Assume we are given a sample of $n_i$ units from population $i, i = 1,2,$ and we measure $m$ different variates on each unit. Our problem is to use the sample information to select the useful and throw out the worthless variates for discriminating between the two groups.

For a particular variable $X$, let $J(X)$ denote the divergence between the empirical probability distributions of group 1 and group 2 units over the question, i.e.

$$J(X) = \sum_{x=1}^{k} (\hat{p}_x - \hat{q}_x) \ln(\hat{p}_x/\hat{q}_x).$$

It tells us the amount of information furnished by variate $X$ for discriminating between the groups 1 and 2 with empirical response probabilities $\hat{p}$ and $\hat{q}$ respectively.

Our sequential procedure begins by choosing for first inclusion the most informative variable $X$ for discriminating. In this first step our procedure is similar in philosophy to that described by Levine (1974), Haaland, Brockett and Levine (1979) and by Goldstein and Dillon (1977), (see also Goldstein and Dillon 1978). Goldstein and Dillon assumed all their variates were dichotomous, however here we do not wish to assume that two categorical variates have the same number permissible categorical responses, e.g., the variables "Sex" and "income level" may have markedly different number of response categories. This prohibits us from using the Goldstein-Dillon procedure. We shall use the quantity $D(X) = n_1 n_2/(n_1+n_2)J(X)$ as a measure of information in variable $X$ for discrimination (cf. Kullback (1959), Gokhale and Kullback (1978)). Asymptotically $D(X)$ has a $\chi^2(k_x - 1)$ distribution,

and this is why direct comparison of the calculated $D(X)$ values (which was utilized by Goldstein and Dillon) is impossible in our situation. A variable with $k_X = 11$ categories is expected to have a $D(X)$ value of 10 while a variable with $k_X = 2$ would be expected to have a $D(X)$ value of 1 under the hypothesis that the variables do not discriminate. This does not directly imply, however, that the variable with 11 categories is more desirable than the variable with 2 categories.

Since $D(X)$ has a $\chi^2(k_X - 1)$ distribution under the null hypothesis that the two groups behave the same on that variable, and has a non-central $\chi^2(k_X - 1)$ distribution with a non-centrality parameter equal to the discriminatory power of the variable in the case where the alternative hypothesis holds and the variable actually discriminates, we shall use the p-value of the $D(X)$ statistic as a measure of discriminatory power of variable $X$. If $p_X = P[\chi^2(k_X - 1) \geq d]$ where $d$ is the observed value for $D(X)$, then the smaller $p_X$, the more informative is variable $X$. Although the values $D(X)$ for various different $X$'s may not be directly comparable in general (as they would be for example if $k_X$ was always the same), the p-values $p_X$ are always comparable and easily calculated from readily available $\chi^2$ tables. If no $\chi^2$ tables are available, one may easily calculate the required $p_X$ values via the probability integral transform. For $v = 1$ degree of freedom we have $P[\chi^2_{(1)} \geq d] = 2(1 - \Phi(\sqrt{d}))$, for $v = 2$ degrees of freedom we have $P[\chi^2_{(2)} \geq d] = e^{-d/2}$, while for $v \geq 3$ degrees of freedom we may use the fact that $(\chi^2_{(v)}/v)^{1/3}$ is approximately $N(1 - \frac{2v}{9}, \frac{2v}{9})$ to obtain $P[\chi^2_{(v)} \geq d] = P[\chi^2_{(v)}/v)^{1/3} \geq (d/v)^{1/3}] = 1 - \Phi(\frac{9(d/v)^{1/3} - 9 + 2v}{\sqrt{18v}})$, where $\Phi$ is the c.d.f. for the standard $N(0,1)$ variate. These formula make the calculation of the $p_X$ values quite simple.

Using the p-values, which are distributed uniformly over $[0,1]$ under the null hypothesis of no discriminatory power, we select as the first variable that $X$ with minimum $p_X$ value, provided this $p_X$ value is significant. We can assess significance for $\min_{1\le X\le m} p_X = U_{(1)}$ by using the distribution function $F(t) = 1 - (1-t)^m$ for $0 \le t \le 1$ as the c.d.f. for $\min_{1\le X\le m} p_X$. Thus the best variable has significant power at level of significance $\alpha$ if $\min_{1\le X\le m} p_X \le 1 - (1-\alpha)^{1/m}$. (The Goldstein-Dillon procedure does not employ the actual distribution for their selection statistic, and hence will not lead to a fixed type 1 error.) Having chosen the first variate for inclusion according to this procedure, we select the second variate for inclusion as that variable which yields the maximum <u>additional information</u> to the already selected first variable. For notational convenience, relabel the variables so that the first variable selected is called 1. We look at all variable pairs $(1,Y)$, $2 \le Y \le m$ and consider the joint probabilities $p_{xy}$ and $q_{xy}$ for the two groups over the possibile category pairs $(x,y)$ on variables $(1,Y)$.

The quantity $D(1,Y) = n_1n_2/(n_1+n_2) \sum_{x=1}^{k_1} \sum_{y=1}^{k_Y} (p_{xy} - q_{xy}) \ln p_{xy}/q_{xy}$ is a measure of the joint discriminatory power of the variable pair $(1,Y)$, and hence $D(1,Y) - D(1)$ is a measure of the increase in discriminatory information obtained by adding variable $Y$. Note that $D(1,Y) - D(1) = n_1n_2/(n_1+n_2) \sum_x \sum_y (p_{xy} - q_{xy}) \ln(p_{xy}/q_{xy}$

$- n_1n_2/(n_1+n_2) \sum_x (p_x - q_x) \ln(p_x/q_x) = n_1n_2/(n_1 + n_2) \sum_x \sum_y (p_{xy} - q_{xy}) \ln(p_{xy}q_x/p_xq_{xy})$

$= n_1n_2/(n_1 + n_2) \sum_x p_x I(p_{Y|x}|q_{Y|x}) + n_1n_2/(n_1 + n_2) \sum_x q_x I(q_{Y|x}|p_{Y|x})$ where $p_{Y|x}$ and $q_{Y|x}$ are the conditional probability distributions of groups 1 and 2 respectively over the categories of variable $Y$ given that variable 1 was $x$, i.e.

$P_{Y|x}(y) = p_{xy}/p_x$. This equality implies $D(1,Y) - D(1) \geq 0$ with equality only if Y contains no additional information given the category of variable 1 (i.e. the addition of Y can only improve things). This equation also shows that the distribution of $D(1,Y) - D(1)$ is a weighted sum of (non-independent) $\chi^2$ variables, the weights reflecting the probability of a particular category x to variable 1, and the $\chi^2$ variable measuring the information expected to be added by variable Y given that particular category x to variable 1. A stepwise regression analogue would consider $\{D(1,Y) - D(1)\}/D(1)$ as a measure of increased discriminatory power obtained by the addition of variable Y. The distributional properties of this ratio have not been explored in full, however when the variables are independent, one has $D(1,Y) = D(1) + D(Y)$ so the above ratio has (asymptotically) an F distribution with parameters $(k_Y - 1, k_1 - 1)$ (see Brockett, Haaland and Levine (1977)).

In this paper we shall also present a different approach based upon information theoretic analysis similar to that used in contingency table analysis (cf. Gokhale and Kullback (1978) and Kullback (1959)). One benefit of this technique over that of Goldstein and Dillon's is that it chooses which variables to include based upon the entire set of categories for the variable rather than individualized for the particular category of a unit under investigation. Thus, the order variable selection, and the particular choice of significant variables, is the same for all units under consideration. For applications such as categorical questionnaire analysis, this uniformity of presentation is quite desirable (even mandatory). For medical screening exams with dichotomous response variables, the more individualized Goldstein-Dillon procedure might be preferred.

Let $x_{ijy}$ denote the number of respondents in group i (i=1,2) who fall in category j on variable 1 and category y on variable Y, and p(i,j,y) represent the corresponding proportion of units in group i within these categories.

A test of the hypothesis that the inclusion of variable Y yields no additional discriminatory power can be obtained by testing the hypothesis that the conditional distribution of Y is independent of the group classification given the category of variable 1, i.e.

$$H_{1,Y}^{(0)}: p(iy|j) = p(i|j)p(y|j) .$$

The worth of adding variable Y is assessed by the p-value for rejecting $H_{1,Y}^{(0)}$. Using the directed divergence distance measure to test $H_{1,Y}^{(0)}$ yields the statistic

$$I(Y|1) = 2I(p(ijy)|p_{(i|j)}p_{(y|j)}p_{(j)}) = 2\sum_{i=1}^{2}\sum_{j=1}^{k_1}\sum_{y=1}^{k_Y} x_{ijy}\ell n\left(\frac{x_{ijy}x_{\cdot j\cdot}}{x_{ij\cdot}x_{\cdot jy}}\right)$$

where the dot replacing a subscript is the usual convention for "sum over that variable", i.e. $x_{ij\cdot} = \sum_{y=1}^{k_Y} x_{ijy}$ etc. The distribution of $I(Y|1)$ under $H_{1,Y}^{(0)}$ is asymtotically $\chi^2$ with $k_1(k_Y - 1)$ degrees of freedom (c.f. Kullback 1959, or Gokhale and Kullback 1978). Hence if $p_Y = P[\chi^2(k_1k_Y - k_1) \geq i(Y|1)]$ where $i(Y|1)$ is the observed value for $I(Y|1)$, we select the second variable to minimize $p_Y$, $2 \leq Y \leq m$. The exact level of significance can be found from the distribution function $F(t) = 1 - (1-t)^{m-1}$, $0 \leq t \leq 1$.

If the p-value for the second variable is significant, we proceed on to select the third variable by the same procedure. We test if the group classification and the category of variable Z, $3 \leq Z \leq m$ are conditionally independent given the categories on variables 1 and 2. The information statistic is

$$2\sum_{i=1}^{2}\sum_{j=1}^{k_1}\sum_{k=1}^{k_2}\sum_{\ell=1}^{k_\ell} x_{ijk\ell}\ln\left(\frac{x_{ijk\ell}x_{\cdot jk\cdot}}{x_{ijk\cdot}x_{\cdot jk\ell}}\right)$$

which is $\chi^2$ with $k_1k_2(k_\ell - 1)$ degrees of freedom. Again, the p-values for each variable $3 \leq \ell \leq m$ are compared to determine if the minimum p-value variable is significant. If it is, we include it as variable 3 and proceed until we obtain a non-significant result. When we finally find a non-significant result, we quit adding variables and consider the task complete. This procedure can reduce the number of variates included. The logic behind this stopping rule is that if there is not enough information in the variable for rejecting the hypothesis of equality of response probabilities for the two groups (given the previously selected variables), then the additional discriminatory power obtained by including this variable will be minimal. Using this procedure, the chance of falsely including a worthless variable will converge to $\alpha$ as the number of units increases. Goldstein and Dillon (1978) recommend a liberal inclusion level $\alpha$ for stepwise selection of variables, and for stepwise regression, Bendel and Afifi (1977) similarly recommend levels $.15 \leq \alpha \leq .25$. These considerations apply here also.

One problem with utilizing contingency table methods of analysis is the rapid proliferation of cells, resulting in possible empty cells. Empty cells here won't bother us, however, zero marginals will. The problem of zero marginals, and the resulting loss of degrees of freedom is discussed in Gokhale and Kullback (1978). Another approach is to break the variates into subsets of 3 or 4 variables each, each group of variables being treated separately. For example in a categorical questionnaire, a group of questions concerning economic status might be treated separately from a group concerning health, which in turn is treated separately from a group concerning education level. Each group is analyzed to obtain the questions in that particular group which should be included. The best subset of

each group is then combined to form the overall questionnaire. Still another approach to the sparse cells problem is the "nearest neighbor" approach of Hills (1966). One estimates $p_{ijk\ldots}$ by grouping together units whose categorical responses differ in only one place on one variable. This technique will increase (perhaps in a somewhat biased way) the number of observations in each cell. Essentially one is trading bias for sample size considerations.

## References


1. Bendel, R. B and A. A. Afifi (1977),"Comparison of stopping rules in forward stepwise regression." J. Am. Stat. Assoc. 72(357), 46-53.

2. Brockett, P. L., P. Haaland and A. Levine (1977), "Discriminant Analysis for Categorical Questionnaire Data", Tulane University mathematics preprint.

3. Gokhale, D. V. andS. Kullback (1978, The Information in Contingency Tables, New York, Marcel Dekker, Inc.

4. Goldstein, M. and W. R. Dillon (1977), "A stepwise discrete variable selection procedure", Comm. Stat., Theory and Methods 6, 1423-1436.

5. ————————(1978), Discrete Discriminant Analysis, New York, John Wiley and Sons.

6. Haaland, P., Brockett, P. L. and A. Levine (1977), "A characterization of Divergence with applications to Questionnaire Information", to appear Information and Control.

7. Hills, M. (1966), "Allocation rules and their error rates", J. Roy Stat. Soc. B 28, 1.

8. Kullback, S. (1959), Information Theory and Statistics, New York, John Wiley and Sons, Dover Press (1968), New York.

9. Levine, A. (1974), "A new approach to Discriminant Analysis in Screening Questionnaires", Int. Symp. on Epidemiological Studies in Psychiatry, Tehran.

10. Weiner, J, and O. J. Dunn (1966), "Elimination of Variates in linear discrimination problems", Biometrics 22, 268.

Unclassified

Security Classification

## DOCUMENT CONTROL DATA - R & D

*(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)*

| | |
|---|---|
| 1. ORIGINATING ACTIVITY (Corporate author)<br>Center for Cybernetic Studies<br>The University of Texas at Austin | 2a. REPORT SECURITY CLASSIFICATION<br>Unclassified<br>2b. GROUP |

3. REPORT TITLE

Information Theoretic Stepwise Selection of Discriminating Discrete Variables

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

5. AUTHOR(S) (First name, middle initial, last name)

P. Brockett, P. Haaland, A. Levine

| 6. REPORT DATE | 7a. TOTAL NO. OF PAGES | 7b. NO. OF REFS |
|---|---|---|
| May 1979 | 12 | 10 |

| 8a. CONTRACT OR GRANT NO. | 9a. ORIGINATOR'S REPORT NUMBER(S) |
|---|---|
| N00014-75-C-0616 and 0569 | CCS 340 |
| b. PROJECT NO.<br>NR047-021 | |
| c. | 9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report) |
| d. | |

10. DISTRIBUTION STATEMENT

This document has been approved for public release and sale; its distribution is unlimited.

| 11. SUPPLEMENTARY NOTES | 12. SPONSORING MILITARY ACTIVITY |
|---|---|
| | Office of Naval Research (Code 434)<br>Washington, DC |

13. ABSTRACT

Often the scientist is faced with a large number of categorical variates which are of potential use in discriminating between two pregiven groups of objects. For example, an investor may wish to assign a particular firm to one of two possible risk groups based upon certain known characteristics of the firm (liquid to fixed asset ratio, etc.), or an engineer might wish to determine which of two models best describes a particular situation based upon the observed characteristics of situation. This is the general problem of variable selection in discriminant analysis. When obtaining and processing the numerous variables is expensive, one must select a "best subset" of variables which incorporates as much information for discriminating as possible. If time is also a factor, a stepwise procedure is mandated. We propose such a stepwise procedure here based upon information theoretic considerations.

DD FORM 1 NOV 68 1473 (PAGE 1)

S/N 0101-807-6811

Unclassified

Security Classification

A-31408

| 14 KEY WORDS | LINK A | | LINK B | | LINK C | |
|---|---|---|---|---|---|---|
| | ROLE | WT | ROLE | WT | ROLE | WT |
| Variable selection<br>Information theory<br>Chi-squared distribution | | | | | | |