Listy Biometryczne
Vol. XXIII (1986) No. 2
Biometrical Letters

POLYCHOTOMOUS LOGISTIC DISCRIMINATION IN DIAGNOSIS OF BRONCHIAL ASTHMA AND CHRONIC BRONCHITIS

EWA KRUSIŃSKA, JERZY LIEBHART

Institute of Computer Science, Wrocław University
Department of Internal Diseases, Medical Academy of Wrocław

Praca wpłynęła 10 grudnia 1985; w wersji ostatecznej 13 listopada 1986

Krusińska E., Liebhart J., 1987. Polychotomous logistic discrimination in diagnosis of bronchial asthma and chronic bronchitis (Polichotomiczna dyskryminacja logistyczna w diagnozie astmy oskrzelowej i przewlekłego nieżytu oskrzeli). Listy Biometryczne XXIII, z. 2. Wydawnictwo Naukowe Uniwersytetu im. Adama Mickiewicza w Poznaniu (Adam Mickiewicz University Press), pp. 43-56, 1 table. ISBN 83-232-0091-2, ISSN 0458-0036.

W pracy pokazano zastosowanie polichotomicznej dyskryminacji logistycznej do automatycznej diagnozy astmy oskrzelowej i przewlekłego nieżytu oskrzeli. Estymacja parametrów funkcji logistycznej dla kilku populacji jest przeprowadzana przy użyciu własnej oryginalnej metody będącej rozszerzeniem metody Walkera i Duncana (Walker i Duncan (1967)) opracowanej dla przypadku dychotomicznego.

Algorytm polega na poprawianiu początkowych estymatorów parametrów funkcji na kolejnych osobnikach z próby.

Dyskryminacja logistyczna została przeprowadzona dla 10 zmiennych o największej sile dyskryminacji w różnicowaniu pomiędzy rozważanymi schorzeniami i kontrolą, wybranymi spośród 92 zmiennych przy użyciu statystyki Λ Wilksa. Wśród wybranych zmiennych 4 miały charakter ciągły, a 6 - dyskretny. Logistyczna funkcja dyskryminacyjna może być stosowana dla mieszanin zmiennych, gdyż jest zdefiniowana dla rodziny rozkładów wykładniczych.

Przeprowadzono porównanie wyników z klasyczną liniową funkcją dyskryminacją. Rezultaty dyskryminacji logistycznej były nieco lepsze niż klasycznej dyskryminacji liniowej. Frakcja poprawnie zaklasyfikowanych osobników wzrosła z 94 do 96%.

The aim of this paper is to present the polychotomous logistic discrimination as applied to the automatic diagnosis of bronchial asthma and chronic bronchitis. The

estimation of the parameters of logistic functions in the polychotomous case is made using an original method which was developed as an extension of the Walker and Duncan procedure (Walker and Duncan (1967) for the dichotomous case.

The algorithm consists in correcting the initial estimates of parameters with succeeding individuals.

The logistic discrimination was performed for 10 variables with the greatest discriminative power (according to Wilks Λ statistic) in differentiation between the considered diseases and a control chosen out of the set of 92 features. The selected variables were of both continuous and discrete character. The logistic discriminant function allows for the mixtures of variables and is defined for the expotential family of distributions.

A comparison with the classical Fisherian linear discrimination was performed. The results of the logistic discrimination were a little better than the classical linear discrimination results. The percentage of correctly classified individuals increased from 94 to 96 percent.

1. INTRODUCTION

Mixtures of continuous and discrete features are quite often met in various practical problems which may be solved by discriminant analysis methods. They are present in medical diagnostics and also in agricultural and economical research. In spite of this wide range of applications there are only a few methods solving this problem (Seber 1984)). These procedures have mostly been elaborated in the last 10 years. Lachenbruch (1975) in his monograph on discriminant analysis mentioned only the logistic approach (Anderson (1972)) as defined for the expotential family of distributions and the foundations of nonparametric classification rules.

Habbema et al (1974, 1978) developed a general kernel approach for mixed data based on nonparametric density estimation.

Krzanowski (1975) presented the location model technique for the mixtures of binary and continuous variables. The method requires the partition into cells. In each cell every discrete variable is in a different state. The classification rule consists in linear discrimination inside each cell separately. Lachenbruch and Golstein (1979) mentioned some possibilities of a generalization of this model. Krzanowski (1980) also gave an extension of his method to the mixtures of discrete and continuous variables.

The surveys of methods for discrimination with mixed variables are geven by Knoke (1982) and Vlachonikolis and Marriott (1982).There is also a rich bibliography of the problem.Seber (1984) in his monograph on multivariate analysis summarizes the actual state of knowledge concerning this problem.

The discrimination with both continuous and discrete variables was of interest to authors in connection with automatic diagnosis of the so--called chronic obturative lung disease (i.e., bronchial asthma, chronic

bronchitis and lung emphysema). The number of persons suffering from this disease still increases and accounts now over 20 percent of adult population (Sawicki (1977)). Therefore an automatic assistance of the diagnosis seems to be useful in early recognition of bronchial asthma and chronic bronchitis. The investigation for preparing the computer system for an automatized antiasthmatic consulting unit started in 1975. The first trials of computerized diagnosis were performed only for continuous variables. It appeared that they were not sufficient to make an exact diagnosis (Bartkowiak, Liebhart et al (1981)).

Then the discrete features were also taken into the analysis. The method used (Krusińska, Liebhart (1985, 1986)) was based on a simple, preliminary transformation of the groups of original discrete features into characteristic functions of linguistic variables. Then the classical discriminant functions were evaluated for the continuous and linguistic variables together.

Now the polychotomous logistic discrimination will be applied for differentiation between bronchial asthma, chronic bronchitis and a control group. The predictor variables are of both continuous and discrete character. The method used for estimating the parameters of the logistic functions is an original one. The main ideas of the method were presented by Krusińska (1985a) in Poznań during the 23rd Scientific Session of the Polish Biometric Society. This method is a generalization of the Walker and Duncan procedure (1967) for the dichotomous logistic discrimination.

2. WALKER AND DUNCAN PROCEDURE

Walker and Duncan (1967) described the method of estimating the parameters of the logistic discriminant function in the case of the dichotomous criterion variable.

Let us assume that we have a sample of N individuals. For each one the values of s predictor variables are measured. So each individual is characterized by the vector $\underline{x}_n' = (x_{n0}, x_{n1}, \ldots, x_{ns})(x_{n0} = 1)$ of s predictor variables and a criterion variable p_n $(n = 1,2,\ldots, N)$ which is binomial and defined as equal to 1 if $\underline{x}_n \in \Pi_1$ and 0 if $\underline{x}_n \in \Pi_2$.

For the whole sample we get an observation matrix

$$X = \begin{bmatrix} x_{10} & x_{11} & \cdots & x_{1s} \\ \vdots & & & \\ x_{N0} & x_{N1} & \cdots & x_{Ns} \end{bmatrix}.$$

The variable p_n $(n = 1, 2, \ldots, N)$ has its expected value $E(p_n) = P_n$. We assume that the probability P_n is of the form described by the logistic function

$$P_n = f(\underline{x}_n,\underline{\beta}) = [1+\exp(-\underline{x}_n'\underline{\beta})]^{-1} \quad (n = 1, 2,\ldots, N), \tag{1}$$

where $\underline{\beta}' = (\beta_0,\beta_1,\ldots,\beta_s)$.

The logistic function is defined for the expotential family of distributions which contains both continuous and discrete distributions (Lachenbruch (1975)).

The vector $\underline{\beta}$ comprising s+1 unknown parameters should be estimated from the data.

It follows that the considered model is of the form

$$p_n = f(\underline{x}_n,\underline{\beta})+\varepsilon_n, \quad E(\varepsilon_n) = 0, \quad \mathrm{var}(\varepsilon_n) = P_nQ_n \quad (n = 1, 2,\ldots, N), \tag{2}$$

where $Q_n = 1-P_n$.

Walker and Duncan (1967) rewrite the model (2) to a more suitable form by linearization. Let us expand (2) in a Taylor series about some initial guessed value of $\underline{\beta}$, say $\overline{\underline{\beta}}$, write $F(\underline{x}_n,\overline{\underline{\beta}})$ for $\partial f(\underline{x}_n,\underline{\beta})/\partial\underline{\beta}$ at $\underline{\beta} = \overline{\underline{\beta}}$ and obtain

$$p_n \approx f(\underline{x}_n,\overline{\underline{\beta}})+ F(\underline{x}_n,\overline{\underline{\beta}})(\underline{\beta}-\overline{\underline{\beta}})+\varepsilon_n \tag{3}$$

or

$$y_n^* \approx F(\underline{x}_n,\underline{\beta})\underline{\beta}+\varepsilon_n \quad (n = 1, 2,\ldots, N) \tag{4}$$

where „the working observation"

$$y_n^* = p_n-f(\underline{x}_n,\overline{\underline{\beta}})+F(\underline{x}_n,\overline{\underline{\beta}})\overline{\underline{\beta}}.$$

Noting that the vector of derivatives equals $F(\underline{x}_n,\overline{\underline{\beta}}) = \overline{P}_n\overline{Q}_n\underline{x}_n$ (where $\overline{P}_n = [1+\exp(-\underline{x}_n'\overline{\underline{\beta}})]^{-1}$, $\overline{Q}_n = 1-\overline{P}_n$) defining $\underline{x}_n^* = \overline{P}_n\overline{Q}_n\underline{x}_n$ and writing X^* for the matrix having $\underline{x}_n^{*\prime}$ as its nth row, the system (4) can be rewritten in the form

$$\underline{y}^* \approx X^*\underline{\beta}+\underline{\varepsilon} \tag{5}$$

where $\underline{y}^*$ and $\underline{\varepsilon}$ are the vectors composed of y_n^* and ε_n, respectively.

Equation (5) is of the form suitable to the application of a weighted iterative least squares procedure with a diagonal matrix W, where W is an inverse of the variance matrix of the vector $\underline{\varepsilon}$. Weights must be estimated from the data, because $W = \mathrm{diag}(1/P_nQ_n)$ depends on $\underline{\beta}$.

Now we can write normal equations as

$$X^{*\prime}WX^*\underline{\beta} = X^{*\prime}W\underline{y}^*. \tag{6}$$

Noting that $X^* = W^{-1}X$ and writing $\underline{y}$ for the vector of the rescaled working observations $W\underline{y}^*$ the system (6) can be rewritten as

$$X'W^{-1}X\underline{\beta} = X'W^{-1}\underline{y} \tag{7}$$

Proceeding from (7) we can write the estimate $\underline{b}$ of $\underline{\beta}$ in the form

$$\underline{b} = (X'W^{-1}X)^{-1}X'W^{-1}\underline{y}. \tag{8}$$

Standard iterative methods of solving (8) require inverting the matrix $V = (X'W^{-1}X)^{-1}$ on each iteration. When the rank of V is large such methods require a lot of computer time. To avoid this Walker and Duncan (1967) gave a recursive method of solving (8).

Let us assume that the starting estimate of $\underline{\beta}$, say $\underline{b}_0$ is given. It can be chosen as the classical linear discriminant function. Using the information given by the first individual we correct this estimate. Then we take the second individual and start correcting again etc. Let $\underline{b}_n$ be an estimate of $\underline{\beta}$ based on the first n individuals. The variance matrix of $\underline{b}_n$ is given by

$$V_n = \left(X_n'W_n^{-1}X_n\right)^{-1} \tag{9}$$

where X_n and W_n are based on the first n individuals.

It is obvious that (9) can be rewritten as

$$V_n = \left(X_{n-1}'W_{n-1}^{-1}X_{n-1}+\underline{x}_n w_n^{-1}\underline{x}_n'\right)^{-1} = \left(V_{n-1}^{-1}+\underline{x}_n w_n^{-1}\underline{x}_n'\right)^{-1} \tag{10}$$

where $w_n = 1/\hat{P}_n\hat{Q}_n$, $\hat{P}_n$ is an estimate of P_n based on the last known approximation of $\underline{b}$, i.e., $\underline{b}_{n-1}$.

Applying (8) and (10) Walker and Duncan (1967) give simple recursive formulas to obtain the estimates $\underline{b}_n$ and V_n. These are

$$V_n = V_{n-1}-V_{n-1}\underline{x}_n w_n^{-1}c_n w_n^{-1}\underline{x}_n'V_{n-1} \tag{11}$$

and

$$\underline{b}_n = \underline{b}_{n-1}+V_{n-1}\underline{x}_n w_n^{-1}c_n(p_n-\hat{P}_n) \tag{12}$$

where

$$c_n = \left(w_n^{-1}+w_n^{-1}\underline{x}_n'V_{n-1}\underline{x}_n w_n^{-1}\right)^{-1},$$

$$\hat{P}_n = \frac{1}{1+\exp(-\underline{b}_{n-1}'\underline{x}_n)}.$$

To avoid an effect of the starting estimates $\underline{b}_0, V_0$ on the final results of the recursive procedure (11), (12) Walker and Duncan (1967) proposed to correct $\underline{b}_0$ and V_0 on the n first individuals and then to calculate new estimates $\underline{b}_n$, V_n as

$$V_n = \left(V_n^{*-1} - V_0^{-1}\right)^{-1},$$

$$\underline{b}_n = V_n\left(V_n^{*-1}\underline{b}_n^{*} - V_0^{-1}\underline{b}_0\right),$$

where $\underline{b}_n^*$, V_n^* were obtained by correcting $\underline{b}_0$, V_0 on the n first individuals.

After obtaining the final estimate of the vector $\underline{\beta}$ the estimate $\hat{P}_n$ of the probability that an individual $\underline{x}_n$ belongs to the first population can be calculated. When $\hat{P}_n$ is greater or equal 0.5 the individual $\underline{x}_n$ is classified to the first population, otherwise it is classified to the second one.

A generalization of the Walker and Duncan method to the case of polychotomous criterion variable with unordered categories was proposed by Krusińska (1985a), (1985b).

Let us assume that we consider g populations. The a posteriori probabilities that $\underline{x}$ belongs to Π_i (i = 1, 2, ..., g) are given by the family of logistic discriminant functions (also defined as in the dichotomous case for the expotential family of distributions).

$$\Pr(\Pi_i|\underline{x}) = \frac{\exp(\underline{\beta}^{(i)\prime}\underline{x})}{1 + \sum_{j=1}^{g-1} \exp(\underline{\beta}^{(j)\prime}\underline{x})} \quad (i = 1, 2, \ldots, g-1) \tag{13}$$

where $\underline{\beta}^{(i)\prime} = (\beta_0^{(i)}, \beta_1^{(i)}, \ldots, \beta_s^{(i)})$ are the vectors of parameters of the logistic discriminant functions.

Let us assume as previously that we have a sample of N individuals. For each one the values of s predictor variables are measured. So each individual is characterized by the vector $\underline{x}_n' = (x_{n0}, x_{n1}, \ldots, x_{ns})$ of s predictor variables and the criterion variable $\underline{p}_n$ (n = 1, 2, ..., N). If the nth individual belongs to the group drawn out of the ith population $\underline{p}_n' = (0, \ldots, 0, 1, 0, \ldots, 0)$ where 1 appears in the ith position (i = 1,2,..., g-1).

The considered model can be now written in the form

$$\underline{p}_n = \underline{f}(\underline{x}_n, \underline{\beta}) + \underline{\varepsilon}_n, \quad E(\underline{\varepsilon}_n) = \underline{0},$$

$$\operatorname{var}(\underline{\varepsilon}_n) = \begin{bmatrix} P_{1n}Q_{1n} & -P_{1n}P_{2n} & \cdots & -P_{1n}P_{g-1,n} \\ -P_{2n}P_{1n} & P_{2n}Q_{2n} & \cdots & -P_{2n}P_{g-1,n} \\ \cdots & & & \\ -P_{g-1,n}P_{1n} & -P_{g-1,n}P_{2n} & & P_{g-1,n}Q_{g-1,n} \end{bmatrix} \tag{14}$$

$$(n = 1, \ldots, N),$$

where $\underline{f}(\underline{x}_n, \underline{\beta})$ is the family of logistic discriminant functions, $\underline{\beta}' = (\underline{\beta}^{(1)\prime}, \underline{\beta}^{(2)\prime}, \ldots, \underline{\beta}^{(g-1)\prime})$, P_{in} (i = 1, 2, ..., g-1) is the a posteriori probability in the logistic discrimination, $Q_{in} = 1 - P_{in}$.

Now, making analogous transformations of (14) as in the dichotomous case we can obtain the recursive formulas for $\underline{b}$ - the estimate of $\underline{\beta}$ and V - the variance matrix of $\underline{b}$.

Expanding in a Taylor series about $\bar{\underline{\beta}}$ the model (14) can be rewritten as

$$Y^* \approx X^*\beta + \varepsilon$$

where

$$\underline{v}^{*\prime} = (\underline{v}_1^{*\prime}, \underline{v}_2^{*\prime}, \ldots, \underline{v}_N^{*\prime}),$$

$$\underline{v}_n^* = \underline{p}_n - \underline{f}(\underline{x}_n, \underline{\beta}) - \underline{\xi}_n^{*\prime}\underline{b} \quad (n = 1, 2, \ldots, N),$$

$$\underline{\xi}_n^{*\prime} = t_n^{-1}\underline{\xi}_n',$$

$$\underline{\xi}_n' = \begin{bmatrix} \underline{x}_n' & & 0 \\ & \underline{x}_n' & \\ & & \ddots \\ 0 & & \underline{x}_n' \end{bmatrix} \text{ and has } g-1 \text{ rows},$$

$$t_n^{-1} = \operatorname{var}(\underline{p}_n),$$

$$X^* = \begin{bmatrix} \underline{\xi}_1^{*\prime} \\ \underline{\xi}_2^{*\prime} \\ \vdots \\ \underline{\xi}_N^{*\prime} \end{bmatrix},$$

$$\underline{\varepsilon}' = (\underline{\varepsilon}_1', \underline{\varepsilon}_2', \ldots, \underline{\varepsilon}_N').$$

Now we can write the normal equations as

$$X^{*\prime} W X^* \underline{\beta} = X^{*\prime} W Y^* \tag{15}$$

where

$$W = \begin{bmatrix} w_1 & & 0 \\ & w_2 & \\ & & \ddots \\ 0 & & w_N \end{bmatrix}.$$

$$w_n = [\operatorname{var}(\underline{\varepsilon}_n)]^{-1} \quad (n = 1, 2, \ldots, N).$$

Writing T for the matrix

$$\begin{bmatrix} t_1 & & 0 \\ & t_2 & \\ & \ddots & \ddots \\ 0 & & t_N \end{bmatrix}$$

and noting that $X^* = TX$ (where

$$X = \begin{bmatrix} \underline{\xi}'_1 \\ \underline{\xi}'_2 \\ \vdots \\ \underline{\xi}'_N \end{bmatrix},$$

it appears that the system (15) can be rewritten as

$$X'W^{-1}X\underline{\beta} = X'W^{-1}Y \tag{16}$$

where

$$W^{-1} = \begin{bmatrix} w_1^{-1} & & 0 \\ & w_2^{-1} & \\ & \ddots & \ddots \\ 0 & & w_N^{-1} \end{bmatrix},$$

$w_n^{-1} = \mathrm{var}(\underline{\varepsilon}_n)$ $(n = 1, 2, \ldots, N)$,

$Y = TY^*$.

Now we solve the system (16) analogously as in the dichotomous case. Let us assume that $\underline{b}_n$ is an estimate of $\underline{\beta}$ based on the first n individuals. The variance matrix of $\underline{b}_n$ is given as

$$V_n = \left(X'_n W_n^{-1} X_n \right)^{-1}$$

where X_n, W_n (the equivalents of X and W) are also based on the first n individuals.
It is obvious that the variance matrix V_n can be rewritten as

$$V_n = \left(X'_{n-1} W_{n-1}^{-1} X_{n-1} + \underline{\xi}_n w_n^{-1} \underline{\xi}'_n \right) = \left(V_{n-1}^{-1} + \underline{\xi}_n w_n^{-1} \underline{\xi}'_n \right)$$

where the matrix w_n is obtained on the basis of the last known approximation of $\underline{\beta}$, i.e., $\underline{b}_{n-1}$.

Now, after a little algebra (compare the properties given by Rao (1965)) we obtain recursive formulas for $\underline{b}_n$ and V_n.

The estimates $\underline{b}_n$ and V_n obtained by corrections of $\underline{b}_{n-1}$, V_{n-1} with the nth individual in the sample are given as

$$V_n = V_{n-1} - V_{n-1}\underline{\xi}_n w_n^{-1} C_n w_n^{-1} \underline{\xi}_n' V_{n-1} \tag{17}$$

and

$$\underline{b}_n = \underline{b}_{n-1} + V_{n-1}\underline{\xi}_n w_n^{-1} C_n (\underline{p}_n - \hat{\underline{P}}_n), \tag{18}$$

where

$$C_n = \left(w_n^{-1} + w_n^{-1}\underline{\xi}_n' V_{n-1} \underline{\xi}_n w_n^{-1} \right)^{-1},$$

$\hat{\underline{P}}_n = \left(\hat{P}_{1n}, \hat{P}_{2n}, \ldots, \hat{P}_{g-1,n} \right)$ is an estimate of $\hat{\underline{P}}_n$ based on $\underline{b}_{n-1}$.

The more detailed transformations performed to obtain formulas (17) - (18) are given by Krusińska (1985b).

It may be seen that (formulas (17), (18)) the method requires only inverting the $(g-1)x(g-1)$ matrix on each step of correcting the estimates of $\underline{b}$ and V. This gives the reduction of computing time in comparison with the standard iterative procedures for obtaining the maximum likelihood estimates (or in the other words: for solving the systems of nonlinear equations).

After the estimation process the individual $\underline{x}_n$ is classified to the ith population when the probability $\hat{P}_{in}$ is the greatest one. In practice, however, when the a posteriori probabilities are similar the final decision of classification should be left to doctors.

3. MATERIAL

Our data collected in the Department of Internal Diseases, Medical Academy of Wrocław, contain results of examinations of 358 persons. The whole sample of patients was divided into 3 groups.

I. uncomplicated bronchial asthma and bronchial asthma complicated by lung emphysema and partly by chronic cor pulmonale (n = 171),

II. uncomplicated and complicated chronic bronchitis (n = 59),

III. control group (persons without symptoms of these diseases) (n = 128).

For each examined patient a questionnaire of 146 items was filled out. It contained personal data, anamnesis, physical examination, laboratory findings including spirometry and gasometry and medical diagnosis.

At first some laboratory findings were transformed into 31 continuous variables. These were:

1) Pulse rate,
2) Systolic blood pressure,
3) Diastolic blood pressure,
4) Erythrocyte rate sedimentation after 1 hour,
5) Erythrocyte rate sedimentation after 2 hours,
6) Hemoglobin,
7) Erythrocytes,
8) Leucocytes,
9) $\frac{FEV_1 \text{ actual } (cm^3)}{FEV_1 \text{ predicted } (cm^3)} \times 100$,
10) VC (vital capacity) actual (cm^3),
11) $\frac{VC \text{ actual } (cm^3)}{VC \text{ predicted } (cm^3)} \times 100$,
12) FEV_1 (forced expiratory volumne in 1 second) actual (cm^3),
13) $\frac{FEV_1 \text{ actual } (cm^3)}{VC \text{ actual } (cm^3)} \times 100$,
14) FEV_1 after salbutamol or histamine (cm^3),
15) $\frac{\Delta FEV_1 \text{ after salbutamol or histamine } (cm^3)}{FEV_1 \text{ predicted } (cm^3)}$,
16) pH,
17) pO_2,
18) pCO_2,
19) SaO_2,
20) HCO_3 standard,
21) HCO_3 actual,
22) BE,
23) total CO_2,
24) $\frac{FEV_1 \text{ actual } (cm^3)}{VC \text{ predicted } (cm^3)} \times 100$,
25) Eosinophilia,
26) ΔFEV_1 after salbutamol or histamine (cm^3),
27) Age,
28) The number of cigarettes daily x years of smoking,
29) How many years ago did you stop smoking?
30) Heart ventricular rate per minute,
31) Corticosteroid therapy counted in mg of prednison per day.

The continuous features constitute only a smaller part of the questionnaire. It also contains the results gathered for 93 discrete variables. These contain the information about:

Diseases of childhood,
Symptoms of complications caused by the steroid therapy,

Nervous excitability,
Alergic diseases other than bronchial asthma,
Respiratory diseases other than the chronic obturative lung disease,
Cough,
Character of dyspnoea,
Intensity of dyspnoea,
Symptoms of complications caused by antiasthmatic treatment,
Physical examination of the chest,
Qualitative laboratory findings other than ecg and X-ray examination of the chest,
X-ray examination of the chest,
Ecg,
Treatment,
Skin tests,
Bronchial challenge tests.

Out of the complete set of 124 variables (31 continuous and 93 discrete) 92 variables without missing values were chosen because the used method did not allow for missing values.

Then the reduction of dimensionality was performed. The selection of variables was performed with the use of Wilks Λ statistic (Rao (1965)).

$$\Lambda = \frac{|W|}{|T|},$$

where W is the within-group adjusted squares and products matrix, T is the total adjusted squares and products matrix.
The Wilks Λ statistic takes the values from the interval $[0,1]$. When $\Lambda = 0$, it is a complete discrimination between considered groups for the given set of variables. When $\Lambda = 1$ the given set of variables has not the discriminatory power at all.

Using Wilks Λ statistic 10 variables with the greatest discriminative power were chosen out of the set of 92 variables without missing values. These were variables: 6 (hemoglobin), 9 $\left(\frac{FEV_1 \text{ actual}}{FEV_1 \text{ predicted}} \times 100\right)$, 11 $\left(\frac{VC \text{ actual}}{VC \text{ predicted}} \times 100\right)$, 28 („smoking index"), one discrete variable from the set describing „cough", two discrete variables from the set describing „character of dyspnoea", one discrete variable from the set describing „physical examination of the chest", one discrete variable from the set describing „ecg", one discrete variable from the set describing „X-ray examination of the chest". Only the last variable was binomial, the remaining discrete features were observed in more than two stages.

The chosen set of 10 variables with the greatest discriminative power was a basis for calculating the discriminant functions.

4. RESULTS

The set of 10 variables with the greatest discriminative power was used to calculate the coefficients of discriminant functions with the whole sample of 358 individuals. Then this sample was reclassified with the obtained discriminant functions. The results of this reclassification are presented in Table 1. The reclassification obtained with the use of the classical linear discriminant function is compared with the results of logistic discrimination. A small amelioration of the fraction of correctly classified individuals is observed in the case of logistic discrimination. The fraction increases from 94 to 96 percent of correctly classified individuals in reference to a prior medical diagnosis performed by a specialist.

5. CONCLUSIONS

The selected 10 variables with the greatest discriminative power contain sufficient information to perform the differentiation between the considered diseases with a high fraction of correctly classified individuals. From the medical point of view they have also a considerable diagnostic power. The goodness of reclassification obtained with the logistic discriminant function is greater than that for the linear discrimination. The amelioration of the percent of correctly classified individuals is not high and equals 2 percent (the fraction increases from 94 to 96 percent). This is due to the high level of correctly classified individuals in linear discrimination. In such situation 2 percent increase of the fraction of correctly classified individuals should be treated as a good result. The fractions of correctly classified individuals are similar in all groups in the case of logistic discrimination while for the linear one chronic bronchitis is reclassificated much worse than other groups. The considerable advantage of the logistic discriminant function consists in recursive approximation of its parameters. The obtained estimate can be corrected recursively with the information given by new individuals. This feature is very convenient for application in an automatized consulting unit because the discriminant functions can be permanently updated in a very simple way. In the case of linear discrimination the discriminant functions must be calculated anew on the basis of the larger sample. This is not necessary for the logistic discrimination because new individuals are added to the previously calculated estimates. Therefore the presented algorithm for the estimation of the parameters of the logistic discriminant functions seems to be very useful in medical applications for computerized diagnosis.

A c k n o w l e d g e m e n t. The theoretical part of this paper was elaborated when the first author stayed at the Computing Centre of Adam Mickiewicz University in Poznań as a research student. She would like to thank Professor Mirosław Krzyśko for his helpful comments and remarks.

T a b l e 1. The results of reclassification of the sample of 358 individuals

Kind of discriminant function	Group	Number of individuals	Reclassification I	II	III	Fraction of corr. class. ind.	Global fraction
Linear discrimination	bronchial asthma (I)	171	170	0	1	.99	
	chronic bronchitis (II)	59	9	39	11	.66	.94
	control (III)	128	1	1	126	.98	
Logistic discrimination	bronchial asthma (I)	171	168	3	0	.98	
	chronic bronchitis (II)	59	3	49	7	.83	.96
	control (III)	128	0	3	125	.98	

REFERENCES

Anderson J.A. (1972), Separate sample logistic discrimination, Biometrika 59, 19-35.

Bartkowiak A., Liebhart J., Liebhart E., Małolepszy J. (1981), Ocena trzech algorytmów dyskryminacyjnych dla cech ilościowych na przykładzie niektórych schorzeń układu oddechowego, Listy Biometryczne Nr 74, 1-20.

Habbema J.D.F., Hermans J., Reme J. (1978), Variable kernel density estimation in discriminant analysis, in: COMPSTAT 1978 (Corsten L.C.A., Hermans J., eds.), 178-185, Physica Verlag, Vienna.

Habbema J.D.F., Hermans J., Van Den Broek K. (1974), A stepwise discriminant analysis program using density estimation, in: COMPSTAT 1974 (Bruckman G., ed.), 101-110, Physica Verlag, Vienna.

Knoke J.D. (1982), Discriminant analysis with discrete and continuous variables, Biometrika 38, 101-110.

Krusińska E. (1985a), O pewnej metodzie estymacji parametrów logistycznej funkcji dyskryminacyjnej dla kilku populacji, presented during the 23rd Scientific Session of the Polish Biometric Society in Poznań.

Krusińska E. (1985b), Maximum likelihood estimates of the parameters of the logistic discriminant function, Report N-150, Institute of Computer Science, Wrocław University.

Krusińska E., Liebhart J. (1985), Linguistic variables and their application to automatical diagnosis of bronchial asthma and chronic bronchitis, Listy Biometryczne 22, No. XXII, 3-18.

Krusińska E., Liebhart J. (1986), A note on the usefulness of linguistic variables for differentiating between some respiratory diseases, Fuzzy Sets and Systems 18, 131-142.

Krzanowski W.J. (1975), Discrimination and classification using both binary and continuous variables, JASA 70, 782-790.

Krzanowski W.J. (1980), Mixtures of continuous and categorical variables in discriminant analysis, Biometrics 36, 493-499.

Lachenbruch P.A. (1975), Discriminant Analysis, Hafner Press, Macmillan, New York.

Lachenbruch P.A., Goldstein M. (1979), Discriminant analysis, Biometrics 35, 69-85.

Rao C.R. (1965), Linear Statistical Inference and Its Applications, Wiley, New York.

Sawicki F. (1975), Przewlekłe nieswoiste choroby układu oddechowego w Krakowie, PZH Warszawa.

Seber G.A.F. (1984), Multivariate Observations, Wiley, New York.

Vlachonikolis I.G., Marriott F.H.C. (1982), Discrimination with mixed binary and continuous data, Applied Statistics 31, 23-31.

Walker S.H., Duncan D.B. (1967), Estimation of the probability of an event as a function of several independent variables, Biometrika 54, 167-179.