Listy Biometryczne
Vol. XXII 1985 No. 2
Biometrical Letters

LINGUISTIC VARIABLES AND THEIR APPLICATION TO AUTOMATICAL DIAGNOSIS OF BRONCHIAL ASTHMA AND CHRONIC BRONCHITIS (LINGUISTIC VARIABLES IN DIAGNOSIS)

EWA KRUSIŃSKA, JERZY LIEBHART

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

Praca wpłynęła 15 maja 1984; w wersji ostatecznej 17 grudnia 1984

Krusińska E., Liebnart J., 1985. Linguistic variables and tneir application to automatical diagnosis of bronchial asthma and chronic bronchitis (linguistic variables in diagnosis). Listy Biometryczne XXII, z.2. Wydawnictwo Naukowe Uniwersytetu im. Adama Mickiewicza w Poznaniu (Adam Mickiewicz University Press) pp.3-18, 3 tab., 2 fig., PL ISSN 0458-0036.

The aim of this work is to present linguistic variables as applied to the automatic recognition of bronchial asthma and chronic bronchitis. Linguistic variables as a conversion of the groups of original discrete features into the membership functions of fuzzy sets were introduced by Saitta and Torasso (Fuzzy Sets and Systems 5 (1981), 245-258). The quantitative and linguistic variables can be used together as predictor variables to calculate linear, quadratic, canonical and logistic discriminant functions. The results of discrimination obtained for the complete set of variables instead of only quantitative features are considerably better. The percentage of correctly classified individuals in the problem of recognition of bronchial asthma and chronic bronchitis increased from 82 to 95 per cent in the case of linear discriminant function. The variables with the greatest discriminative power (according to Wilks' Λ statistic) were also chosen. These were in turn the linguistic variables: character of dyspnoea, cough and X-ray examination of the chest. It appeared that the first of them had greater discriminative power than the set of 31 quantitative variables. The presented method permits the simultaneous use of qualitative and quantitative variables to discrimination and allows for missing values in discrete features. For this reason it seems to be very useful in such medical applications as ours.

1. AUTOMATICAL DIAGNOSIS OF RESPIRATORY DISEASES

The problem of recognition of respiratory diseases is very important in medical practice. The number of persons suffering from so_ called chronic obturative lung disease (it viz.: bronchial asthma, chronic bronchitis and lung emphysema) still increases and accounts over 20 percent of human adults (Sawicki 1977). The American Thoracic Society has established in 1983 a Task Force to evaluate the State of the Art of Screening for Adult Respiratory Disease. This official ATS statement confirms the significance of respiratory disease as a "social problem" (Tockman, Beckake et al 1983). Therefore the automatic assistance of medical diagnosis seems to be, for the future, necessary for early recognition and screening in this disease.

In a wide range of applications the object under consideration is described by both discrete and continuous variables. This was the case in our problem. Most laboratory findings were continuous in character whilst many other symptoms had to be coded as discrete. For this reason we had to apply a discriminant procedure which allowed for discrete variables.

One of them is the classical Bayes method. In this method the frequences of appearence for each state of every discrete symptom must be known. In practice only the frequences of occurence of some diseases have been estimated till now. It is impossible to perform expensive mass examinations which would be needed to estimate the frequences of occurence of each symptom's state for each disease.

Other procedures, worked out specially for discrete variables, seem to be also useless in the problem under consideration, because of missing values which occur in discrete features in our material which is characteristic for medical data. The additional problem is the cost of programming new specific procedures, which are often expensive for the computer. E.g. it is impossible to apply the location model (Krzanowski 1975) which is one of the best known procedures worked out for discrimination with both continuous and discrete variables. The method requires a partition into cells. In each cell every discrete variable is in a different state. In our case, where over 80 discrete features are considered, the number of cells would be enormous. Besides, a lot of cells would be empty for our data.

The other approaches to the discrimination with mixed variables are kernel method and logistic function. In both cases the problem of missing values must be specially solved.

A survey of methods for discrimination with mixed variables is given by Knoke (1982). There is also a rich bibliography of the problem. Some other informations about mentioned methods can be also found in the recent paper of Krzanowski (1983).

In the practical problem under consideration, we had to find a procedure which gives the possibility of the simultaneous use of continuous and discrete variables, allows for missing values and is inexpensive for the computer.

As is well known, the classical discriminant functions belong to the simplest ways to perform automatic diagnosis. But there is one difficulty: they are defined only for quantitative variables, under the assumption of normality of distributions.

There is, of course, a possibility to use the linear discriminant function as a kind of approximation, on discrete data. But in such a case a problem of missing values occurs just as in the methods mentioned before.

During several years' investigations, which were performed for preparing the computer system for automatized antiasthmatic consulting units, it appeared that the quantitative variables were not sufficient to make an exact diagnosis (A.Bartkowiak, J.Liebhart et al. 1981). Now the qualitative features ought to be also included in the analysis. The method which we used to solve the problem under consideration was based on a simple, preliminary transformation of the groups of original discrete features characteristic functions of linguistic variables. The method allows for missing values. Besides (after a transformation of the discrete features into linguistic variables) we obtained the reduction of the dimensionality of the problem. Then the standard methods of discrimination may be used. The quantitative and linguistic variables are treated together as the predictor variables for obtaining the discriminant functions.

2. THE MAIN IDEAS OF THE FUZZY SET THEORY

The theory of fuzzy sets was intoduced by Zadeh (1965). A fuzzy set is determined by a membership function which assigns to each object of the set a grade of membership with the value between 0 and 1.

Let us consider a simple example.

$$A = \{ x : x \text{ is } \text{ tall} \}$$

For the classical set we have only two possibilities: $x \in A$ or $x \notin A$. When we understand A as a fuzy set, for all the membership function is defined. It can be, for instance, defined as:

$$\phi(x) = \begin{cases} 0 & \text{when } x < 150 \text{ cm} \\ \frac{1}{30}\, x - 5 & 150 \leqslant x \leqslant 180 \text{ cm} \\ 1 & x > 180 \text{ cm} \end{cases}$$

The membership function $\phi(x)$ is presented at the Figure 1. Many symptoms are coded as discrete. The discretization is the simplest, but not always the best way, to characterize these symptoms. The fuzzy set theory makes it possible characterize them as the membership functions.

Let us consider a feature:

How often have you suffered from influenza?

0) not at all

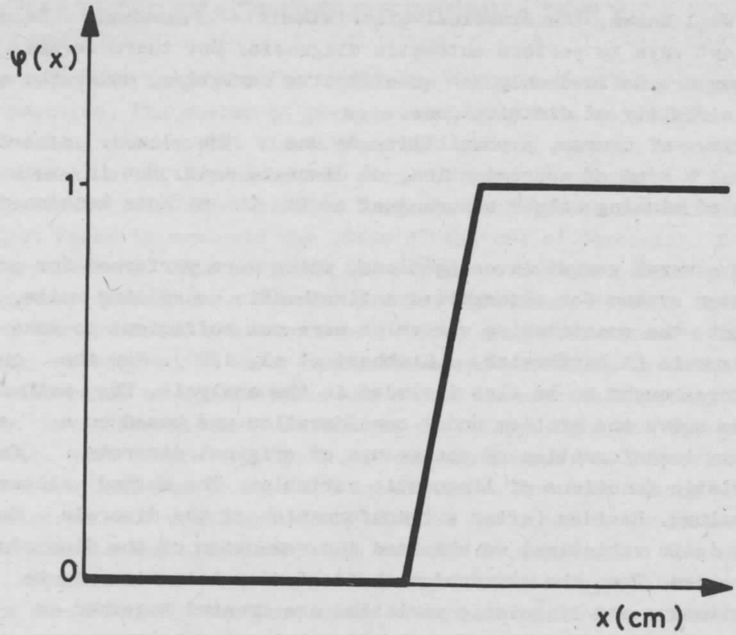

Fig. 1. The membership function $\phi(x)$ for the fuzzy set "tall"

1) occasionaly
2) often.

in reality, there is no sharp border between the answer "occasionally" and "often". The three states mentioned above are distinguished by discretization.

The lingüistic variables were also introduced by Zadeh (1975). The method of transformation into linguistic variables used in this work, following Saitta and Torasso (1981), is based on comparison of the values of discrete variables observed in all groups of ill ("abnormal") persons together and in the control group.

3. MATERIAL

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

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

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

III. control group which contains persons without signs of the, the diseases mentioned above (n = 128).

For each patient a questionnaire of 146 items was filled. It contained personal data, anamnesis, physical examination, laboratory findings including spirometry and gasometry and medical diagnosis. First, some laboratory findings were transformed into 31 quantitative 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) $\frac{FEV_1 \text{ actual } (cm^3)}{FEV_1 \text{ predicted } (cm^3)} * 100$,
10) VC actual (cm^3),
11) VC %/VC predicted (cm^3),
12) FEV_1 actual (cm^3),
13) FEV_1 %,
14) FEV_1 after salbutamol or histamine (cm^3),
15) $\frac{\Delta FEV_1 \text{ after salbutamol and histamine } (cm^3)}{FEV_1 \text{ predicted } (cm^3)}$,
16) pH,
17) p O_2,
18) p CO_2,
19) SaO_2,
20) HCO_3 standard,
21) HCO_3 actual,
22) BE,
23) $\frac{FEV_1 \text{ actual } (cm^3)}{VC \text{ predicted } (cm^3)}$,
24) Eosinophilla,
25) ΔFEV_1 after salbutamol or histamine (cm^3),
26) Age,
27) The number of cigarettes daily x years of smoking,
28) How many years ago did you stop smoking?
29) Heart's ventricular rate per minute,
30) Corticosteroid therapy counted in mg of prednison per day.

The earlier trial of automatical diagnosis was performed only on the first 26 variables of this set. The results obtained for the sample of 303 patients were presented by Bartkowiak, Liebhart et al. (1981). The fraction of correctly classified individuals was almost 70 per cent depending on the kind of discriminant function used for differentiation. So the results were not sufficiently good for practical application to automatical diagnosis. The continuous variables constitute only the smaller part of the questionnaire.Ignoring the qualitative variables we miss a lot of informations for each patient. For this reason at the next step of the statistical analysis a larger set of 112 variables was taken into consideration. For 81 discrete features the conversion of the original variables into linguistic ones was performed. We used the method of Saitta and Torasso (1981).

After transformation we obtained 14 linguistic variables:

31) Diseases of the childhood,
32) Symptoms of complications caused by the steroid therapy,
33) Nervous excitability,
34) Alergic diseases other than bronchial asthma,
35) Respiratory diseases other than chronic obturative lung disease,
36) Cough,
37) Character of dysponea,
38) Intensity of dysponea,
39) Symptoms of complications caused by antiasthmatic treatment,
40) Physical examination of the chest,
41) Qualitative laboratory findings other than ecg and X-ray examination of the chest,
42) X-ray examination of the chest,
43) Ecg,
44) Treatment.

To obtain these linguistic variables the transformation from groups of discrete features into characteristic functions was done. A lot of features coded as discrete in the questionnaire was obtained by discretization of some continuous variables which were difficult to describe. For this reason the fuzzy set theory approach seems to be natural in the preliminary analysis of these data.

To perform the comprehensive diagnosis our data were analysed in the following steps:

A. transformation of the groups of qualitative features into linguistic variables,

B. the choice of variables with the greatest discriminative power,

C.automatic diagnosis by use of linear, quadratic, canonical and logistic discriminant functions.

4. THE TRANSFORMATION OF THE DISCRETE FEATURES INTO LINGUISTIC VARIABLES

The transformation of groups of discrete features into linguistic variables was proposed by Saitta and Torasso (1981). Assume that each linguistic variable L_i $(i = 1,2,\dots,r)$ is described by the set of questions $Q^{(i)}$ taken from a given questionaire. Possible answers to these questions have a discrete character. A weight $\gamma_{jk}^{(i)} \in [-1,1]$ is associated with each answer $s_{jk}^{(i)}$ to question $q_j^{(i)} \in Q^{(i)}$ $(i = 1,2,\dots,r;\ j = 1,2,\dots,M_i$, where $M_i = |Q^{(i)}|$; $k = 1,2,\dots,A_{ij}$, where A_{ij} is the number of possible answers to the question $q_j^{(i)} \in Q^{(i)}$). If $\gamma_{jk}^{(i)}$ is near -1, it means that there is no agreement between the answer $s_{jk}^{(i)}$ and the variable L_i. When $\gamma_{jk}^{(i)}$ is near +1 the degree of agreement between answer $s_{jk}^{(i)}$ and the variable L_i is very high. As an example let us consider a linguistic variable "character of dyspnoea". It is described by 6 questions from the questionnaire. One of them is:

How long have you complained of attacks of dyspnoea?

When a patient does not complain of dyspnoea at all, the agreement degree $\gamma_{jk}^{(i)}$ equals -1. When she (he) complains of attacks of dyspnoea for over 20 years the agreement degree $\gamma_{jk}^{(i)}$ equals +1.

When $\gamma_{jk}^{(i)}$ equals 0 there is no information (possitive or negative)about the degree of agreement between the answer $s_{jk}^{(i)}$ and the variable L_i. Therefore $\gamma_{jk}^{(i)}$ equals 0 for each missing answer. So the method of transformation of discrete features into linguistic variables allows for missing values.

The weights $\gamma_{jk}^{(i)}$ are determined a priori. Next, we define the vector $\alpha^{(i)} = (\alpha_1^{(i)},\dots,\alpha_{M_i}^{(i)})$. The weights $\alpha_j^{(i)}$ are proportional to the discriminative power of the questions from the set $Q^{(i)}$ in differentiating between "normal" and "abnormal" individuals (in medical applications "abnormal" individual means an ill person). These weights can be chosen in various ways E.g. Saitta and Torasso (1981) advice to find them from the Student t-test performed on each question $q_j^{(i)}$ for comparing the mean values in two groups mentioned above. In our calculations we set the weight $\alpha_j^{(i)}$ equal to $t_j^{(i)}$ (where $t_j^{(i)}$ is the computed value of the Student t-statistic for the question $q_j^{(i)} \in Q^{(i)}$), when $t_j^{(i)}$ is greater or $t_j^{(i)}$is of the order of the critical value $t_{0,1}$. In the remaining cases we set $\alpha_j^{(i)}$ equal to 0.

Thus, the total weight of the answer $s_{jk}^{(i)}$ is given as

$$w_{jk}^{(i)} = \alpha_j^{(i)} \cdot \gamma_{jk}^{(i)}. \qquad (1)$$

Then Saitta and Torasso (1981) define, for each individual, variable m_i associated with the linguistic variable L_i as follows:

$$m_i = \sum_{j=1}^{M_i} w_{jk^*}^{(i)}, \qquad (2)$$

where $w_{jk^*}^{(i)}$ is the total weight (formula (1)) associated with this answer to question $q_j^{(i)} \in Q^{(i)}$, which was given by the individual (patient).

Using variable m_i the membership function μ_i can be calculated in the following way:

$$\mu_i(m_i) = \frac{1}{2} + \frac{1}{\pi} \operatorname{arctg} \frac{m_i - a_i}{b_i}, \qquad i = 1,2,\dots,r. \qquad (3)$$

The transformation by the function arctg reduces the values of m_i to the interval [0,1].

In our problem we set a_i in (3) equal to the mean value $\bar{m}_i$ in the whole sample of ill individuals and in the control group together. The values b_i are chosen as proposed by Saitta and Torasso (1981). This means, that b_i are calculated in such a way that

$$\mu_i(m_i') \leqslant 10^{-3},$$
$$\mu_i(m_i'') \geqslant 1-10^{-3},$$

where

$$m_i' = \sum_{j=1}^{M_i} \Big(\min_{1\leqslant k\leqslant A_{ij}} w_{jk}^{(i)} \Big),$$
$$m_i'' = \sum_{j=1}^{M_i} \Big(\max_{j\leqslant k\leqslant A_{ij}} w_{jk}^{(i)} \Big).$$

The membership function $\mu_i(m_i)$ in (3) are in practice a more precise representation of the values of linguistic variables L_i than the variables m_i in (2). Let us return to our example. The 7th linguistic variable "character of dyspnoea" is described by 6 questions. The answers to them are coded as discrete. The first question has two possible answers, each of succeeding four questions has 5 possible answers and the last question has 7 possible answers.

The last question is:

How long have you complained of attacks of dyspnoea?

0) not at all,

1) in the youth, for less than 3 years,

2) in the youth and during the last 5 years,

3) for 5 years,

4) for 5-10 years,

5) for 11-20 years,

6) for over 20 years.

The weights $\gamma_{6k}^{(7)}$ (k = 0,1,...,6) tell us about the degree of agreement between the received answer and the variable "character of dyspnoea". If the patient complains of grave asthmatic dyspnoea for over 20 years the agreement degree equals +1. When she (he) does not complain of dyspnoea at all, the weight equals -1. All weights for 6 questions describing the variable "character of dyspnoea" are presented in the table 1.

Let us consider an individual who answered 0 2 2 4 4 4 to the six questions describing the variable "character of dyspnoea". The variable m_i (2) equals for him 98.979 and the membership function (3) equals 0.99874. For an individual who answered 0 1 0 0 0 0 m_i has the value -88.3150 and the membership function equals 0.00303. It is obvious, that it is much easier to interpret the values from the interval [0,1]. For the predictor variables of the classical discriminant functions the normality of dis-

T a b l e 1. The weights for the questions from the set $Q^{(7)}$ describing the variable "character of dyspnoea"

Number of question		1	2	3	4	5	6
weights of questions		22.88	25.83	20.18	19.69	13.09	25.39
Weights of answers	0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0
	1	1.0	0.5	1.0	0.75	0.5	0.2
	2	-	0.1	1.0	1.0	0.75	0.2
	3	-	0.75	1.0	0.5	0.25	0.4
	4	-	1.0	1.0	0.5	1.0	0.6
	5	-	-	-	-	-	0.8
	6	-	-	-	-	-	1.0

tributions is assumed. For this reason we used the variables m_i instead of the membership functions $\mu_i(m_i)$ for calculating the discriminant functions. The transformation of our data was performed with the program LINGVAR (Krusińska 1984) and written in ALGOL 1900 for the ODRA 1305 computer.

5. THE METHODS OF DISCRIMINANT ANALYSIS

First, for the whole set of quantitative and linguistic variables, we can select variables with the greatest discriminative power to reduce the number of considered features. Wilk's Λ statistic (Rao 1965) can be used to solve this problem. It is defined as a ratio of two determinants

$$\Lambda = \frac{|W|}{|T|}, \tag{4}$$

where W is the within-groups adjusted squares and products matrix, T is the total adjusted squares and products matrix.

The Wilk's Λ statistic takes the values from the interval [0,1]. It has clear interpretation. When $\Lambda = 0$, it is complete discrimination between considered groups using the given set of variables. When $\Lambda = 1$, the given set of variables has no discriminative power at all. The Wilks' Λ statistic was used for the preliminary reduction of the number of variables. Then the various discriminant functions were evaluated on the basis of the obtained subset (as in the paper of Bartkowiak, Liebhart et al 1981).

In the problem under consideration for the whole or reduced set of variables the classical Fisher 's linear and quadratic discriminant functions (Lachenbruch 1975) were calculated. The a priori probabilities were equalled to the fractions of individuals observed in the groups. Next, the canonical discriminant functions as described by Ahrens and Läuter

(1977) were evaluated. In the problem under consideration only statistically significant canonical variates were used to classification. The dimensionality of the discriminant space was tested by χ^2 statistic . We calculated

$$\chi^2_r = (n-K-s+r+1)(\lambda_{r+1}+\lambda_{r+2}+\ldots+\lambda_t), \qquad r = 0,\ 1,\ldots,\ t-1, \tag{5}$$

where: n - the number of individuals in the sample, K - the number of considered groups, s - the number of predictor variables, t - the number of non-zero eigenvalues of matrix B (the form of B is given in (6)), $\lambda_{r+1} \geqslant \lambda_{r+2} \geqslant \ldots \geqslant \lambda_t$ - the eigenvalues of B.

Matrix B is given by

$$B = \frac{1}{n-K} N^{1/2}A'S^{-1}AN^{1/2}, \tag{6}$$

where: N - diag(n_i), n_i $(i = 1,2,\ldots, K)$ - the number of individuals in the ith group, S - the sample within-groups covariance matrix, A - $(\underline{x}_{1.} - \underline{x}_{..}, -\underline{x}_{2.} - \underline{x}_{..}, \underline{x}_{K.} - \underline{x}_{..})$, $\underline{x}_{i.}$ - the sample mean vector of predictor variables in the ith group, $\underline{x}_{..}$ - the general sample mean vector of predictor variables.

Under assumption of normality of multivariate distribution of the vector $\underline{x} = (x_1,x_2,\ldots,x_s)$ of s predictor variables (in each considered population) the statistics specified in (5) have the distribution $\chi^2_{(s-r)(K-r-1)}$ $(r = 0,\ 1,\ldots,\ t-1)$. Now let us assume that we obtained:

for r = l-1

$$\chi^2_{l-1} > \chi^2_{\alpha,(s-l+1)(K-l)}$$

and for r = l

$$\chi^2_l \leqslant \chi^2_{\alpha,(s-l)(K-l-1)}$$

(α - the significance level).

In such a case the dimension of discriminant space equalls l. Only l statistically significant canonical variates are used later for classification.

The logistic discriminant function defined for expotential family of distributions (Cox 1970) was also evaluated.

$$P(\pi_i/\underline{x}) = \frac{\exp\left(b_0^{(i)}+\underline{b}^{(i)'}\underline{x}\right)}{1+\sum\limits_{j=1}^{K-1}\exp\left(b_0^{(i)}+\underline{b}^{(i)'}\underline{x}\right)} \qquad (i = 1,2,\ldots, K-1), \tag{7}$$

where: K is the number of considered populations, $P(\pi_i/\underline{x})$ means a posteriori probability that an individual, $\underline{x}$ - $(x_1,x_2,\dots,x_s)$ belongs to the population π_i, $b^{(i)}$, $\underline{b}^{(i)} = \left(b_1^{(i)},\dots,b_s^{(i)}\right)$ $(i = 1,2,\dots, K-1)$ are the parameters of the logistic discriminant function.

The probability $P(\pi_K/\underline{x})$ is obviously given as

$$P(\pi_K/\underline{x}) = 1 - \sum_{j=1}^{K-1} P(\pi_{j}/\underline{x}).$$

To obtain the values of the logistic discriminant function (5) we must find an estimator of the parameters $b_0^{(i)}$, $\underline{b}^{(i)}$ $(i = 1,2,\dots, K-1)$. We find the maximum likelihood estimators using the iterative procedure of Jennrich and Moore (1975).

The computations can be performed using a standard program for the choice of variables with the greatest discriminative power and for evaluating linear, quadratic and canonical discriminant functions. In our case, the programs from the package SABA (Bartkowiak (1981)) and the program GJENN for logistic discrimination (Krusińska (1982)) were applied.The possibility of the simultaneous use of continuous and discrete features is a considerable advantage of the presented method. Both qualitative and quantitative variables may be used to the statistical analysis without additional costs of programing new, specific discriminant procedures.After a simple and inexpensive transformation into linguistic variables various standard programs may be applied to the data with a great number of discrete features.

6. COMPARISON OF THE RESULTS OF DISCRIMINATION OBTAINED FOR DIFFERENT SETS OF VARIABLES

First we selected variables with the greatest discriminative power from the whole set of 45 variables (31 continuous and 14 linguistic). The selection was performed stepwise. The most discriminative are in turn:

character of dyspnoea,
cough,
X-ray examination of the chest,
the number of cigarettes daily x years of smoking,
hemoglobin,
physical examination of the chest,
pulse rate,

$$\frac{\Delta \text{FEV}_1 \text{ after salbutamol or histamine } (\text{cm}^3)}{\text{FEV}_1 \text{ predicted } (\text{cm}^3)},$$

treatment,
intensity of dyspnoea.

There were six linguistic variables among the ten variables with the greatest discriminative power. The first variable "character of dyspnoea" has greater discriminative power than the set of 31 continuous variables.

The results obtained mathematically are consistent with current medical knowledge. As pointed out in the Report of the American Thoracic Society the symptoms "character of dyspnoea" and "cough " are the most important in the recognition of the chronic obturative lung disease (Tookman, Beckake et al. 1983).

The values of Wilk's Λ statistic as a function of the number of variables for the first 10 variables are presented at the Figure 2.

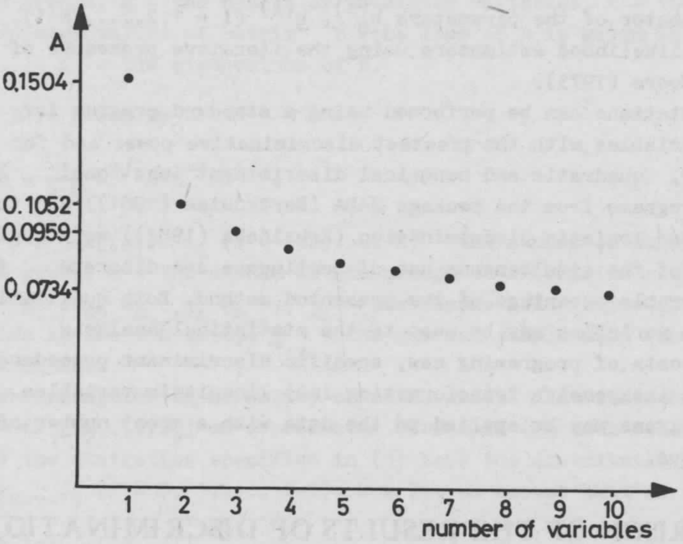

Fig. 2. The values of Wilk's Λ statistic as a function of the number of variables

The comparison of the values of Wilk's Λ statistic for various sets of variables is also presented in the table 2.

Table 2. The comparison of the values of Wilk's Λ statistic for various sets of variables

Set of variables	(a) 31 continuous variables	(b) the complete set of 45 variables	(c) 10 variables with the greatest disc. pow.	(d) 3 variables with the greatest disc. pow.
Wilk's Λ statistic	0.3048	0.0576	0.0743	0.0959

As we can see, the linguistic variables have a great discriminative power for our data. Wilk's Λ statistic decreases from 0.3048 (continuous variables) to 0.0576 (the whole set of variables). It is necessary to add these variables to the statistical analysis. We evaluated the discriminant functions for four sets of variables viz.

a) the continuous variables,

b) the whole set of variables,

c) 10 variables with the greatest discriminative power,

d) 3 variables with the greatest discriminative power.

Of course, it is possible to choose other sets e.g. 8 variables - 4 continuous and 4 linguistic. In the practical medical applications we will not use the set of 3 linguistic variables (d). This set only includes the variables which are transformations of some subjective informations.In practice, such a set must be enriched by objective continuous variables.

The presentation of results of discrimination for these 3 variables is done only for showing the importance of linguistic variables in the problem at issue.

The classical linear discriminant function needs the assumption of normality. The investigation on the normality of univariate probability distributions of the continuous variables and the variables m_i (formula 2) associated with L_i resulted, in almost all cases, in rejection of this hypothesis. The rejection of the hypothesis was obtained also for testing the homoscedasticity of covariance matrices. But the results obtained with the linear discriminant function are not much worse than the other results. We can see that this method is very proof against nonnormality and heteroscedasticity.

As a measure of goodness of the discrimination the percentage of correctly classified individuals is taken. We observe a considerable improvement of results using all quantitative and linguistic variables instead of only quantitative features. The percentage of correctly classified individuals increases from 82% to 95% in the case of linear discriminant function. A similar increase is observed for the other functions.

The results obtained for the reduced sets of 10 and 3 variables with the greatest discriminative power are also better than the results for quantitative variables. This is obvious because the set of quantitative variables has a lower discriminative power than these 2 reduced sets (see table 2).

The best results of discrimination were obtained with the logistic discriminant function (100 per cent or correctly classified individuals for the whole set of variables (a)). There is only one disadvantage of this function. The estimator of its parameters is found by use of the iterative method which needs much more computer time than the classical linear discriminant function. The quadratic discriminant function gives somewhat worse results than the logistic function. But when the number of the considred variables is great, there are difficulties with obtaining the determinants of the sample covariance matrices. In our calculations they are found by a special procedure (for sets (a) and (b)) with somewhat less accuracy.

T a b l e 3. The comparison of the results of discrimination

Kind of discr. function	Asthma (I), n = 171			Chronic Bronchitis (II), n = 59			Control (III), n = 128			Percentage of correctly classified ind.
	I	II	III	I	II	III	I	II	III	
Linear (a)	142	10	19	18	33	8	5	3	120	82 %
Quadratic (a)	159	3	9	19	39	1	8	3	117	88 %
Canonical (a)	121	29	21	11	40	8	4	5	119	78 %
Logistic (a)	151	10	10	16	36	7	5	2	121	86 %
Linear (b)	170	1	0	10	44	5	1	1	126	95 %
Quadratic (b)	171	0	0	25	34	0	1	0	127	93 %
Canonical (b)	170	1	0	9	47	3	1	2	125	96 %
Logistic (b)	171	0	0	0	59	0	0	0	128	100 %
Linear (c)	170	1	0	14	34	11	1	1	126	92 %
Quadratic (c)	164	7	0	7	50	2	1	5	122	94 %
Canonical (c)	170	1	0	11	40	8	1	1	126	94 %
Logistic (c)	167	4	0	7	49	3	1	2	125	95 %
Linear (d)	170	1	0	14	25	20	1	1	126	90 %
Quadratic (d)	168	3	0	13	40	6	1	5	122	92 %
Canonical (d)	169	2	0	14	31	14	1	2	125	91 %
Logistic (d)	168	3	0	10	42	7	1	2	125	94 %

(a) - continuous variables, (b) - all variables, (c) - 10 variables with the greatest discriminative power, (d) - 3 variables with the greatest discriminative power.

The results obtained for linear and canonical discriminant functions are at the same level. These two kinds of functions can be recommended for the practical use where quick and unexpensive computations are needful. The canonical discriminant functions have also clear graphical interpretation. The results of differentiation near 95 per cent of correctly classified individuals for complete and reduced set of 10 variables are sufficiently good for practical applications as the assistance in medical diagnosis.

7. CONCLUSIONS

The method presented in this work permits the simultaneous use of both quantitative and qualitative variables for discrimination. Therefore all important symptoms of the disease could be taken into consideration. The simple transformation of the groups of discrete features into linguistic variables can be performed quickly and inexpensively using a computer. Having performed such a transformation, the standard programs for linear, quadratic, canonical and logistic discrimination may be used. The percentage of correctly classified individuals obtained for continuous features is too low in order for the classical discriminant functions to be used practical in problems. The results obtained with linguistic and quantitative variables together are much better. The transformation of the original discrete features into linguistic variables leads to results of discrimination such that the classical discriminant functions may be applied as an assistance of medical diagnosis in daily practice. The authors consider that the method presented here can be applied, first of all, in automatized respiratory disease and occupational diseases consulting units (for factories, where air pollution occurs). In the future, it could be included to larger diagnosis systems.

Acknowledgements. The authors would like to thank Dr. Anna Bartkowiak of the Institute of Computer Science, Wrocław University, for helpful advice and comments. The authors are also grateful to the unknown referees for their valuable remarks.

REFERENCES

Ahrens H., Läuter J. (1977). Mehrdimensionale Varianzanalyse. Akademie Verlag, Berlin.

Bartkowiak A. (1981). SABA. Opis techniczny programów statystycznych w języku ALGOL 1900. Wydawnictwo Uniwersytetu Wrocławskiego, Wrocław.

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

Cox D.R. (1970). The analysis of binary data. London, Menthuen.

Jennrich R.I., Moore R.H. (1975). Maximum likelihood estimation by means of nonlinear least squares, in: Proceedings of the statistical computing section (Am. Statist. Assoc.), 57-65.

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

Krusińska E. (1982). Logistic discrimination for several groups of data. Raport Nr N-110, Instytut Informatyki Uniwersytetu Wrocławskiego, Wrocław.

Krusińska E. (1984). Linguistic variables and their application to discrimination. Raport Nr N-133, Instytut Informatyki Uniwersytetu Wrocławskiego, Wrocław.

Krzanowski W.J. (1975). Discrimination and classification using both binary and continuous variables. JASA 70, 782-789.

Krzanowski W.J. (1983). Stepwise location model choice in mixed-variable discrimination. Applied Statistics 32, 260-266.

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

Rao C.R. (1965), Linear statistical inference and its applications. Wiley, New York.

Saitta L., Torasso P. (1981). Fuzzy characterization of coronary disease. Fuzzy Sets and Systems 5, 245-258.

Sawicki F. (1977). Przewlekłe nieswoiste choroby układu oddechowego w Krakowie. PZH Warszawa.

Tockman M.S., Beckake M.R., Clausen J.L., Fontana P.S., O'Brien R.J., Permutt S., Petty T.L., Reed Ch., Reichman L.B., Stead W.W. (1983). Screening for adult respiratory disease. American Revue of Respiratory Diseases 128, 768-774.

Zadeh L.A. (1965). Fuzzy sets. Information and Control 8, 338-353.

Zadeh L.A. (1975). The concept of a linguistic variable and its application to approximate reasoning. Information Sci. 8, 199-250, 301-358; 9, 43-80.