ELSEVIER

Computational Statistics & Data Analysis 38 (2001) 139–160

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

A comparison of discriminant procedures for binary variables

Ognian K. Asparoukhov[a], Wojtek J. Krzanowski[b,*]

[a] *AI Insight, Inc., 602 Courtland Street, Suite 400, Orlando, FL 32804-1342, USA*
[b] *School of Mathematical Sciences, University of Exeter, Laver Building, North Park Road, Exeter EX4 4QE, UK*

Received 1 July 2000; received in revised form 1 February 2001; accepted 1 February 2001

Abstract

Thirteen discriminant procedures are compared by applying them to five real sets of binary data and evaluating their leave-one-out error rates. Three versions of each data set have been used, containing respectively "large", "moderate" and "small" numbers of variables. To achieve the latter two categories, reduction of variables was first carried out using the all-subsets approach based on Kullback's information divergence measure. Sample size, number of non-empty multinomial cells and Empirical Integrated Rank are taken into account in assessment of classifier effectiveness. While the data sets are ones that arose during day-to-day statistical consulting, the empirical basis for drawing widespread conclusions is inevitably limited. Nevertheless, the study did highlight the following interesting features. The Kernel, Fourier and Hall's k-nearest neighbour classifiers had a tendency to overfit the data. The mixed integer programming classifier was clearly better than the other linear classifiers, and linear discriminant analysis had better results than logistic discrimination especially for small sample sizes. The second-order Bahadur procedure was generally very effective when the number of variables was large, but only if the sample size was large when the number of variables was small. The second-order log-linear models were very effective when the number of variables was small or when the sample sizes were large. Quadratic discrimination and Hills' k-nearest neighbour classification both performed poorly. The traditional statistical classifiers did not cope well with sparse binary data; the non-traditional classifiers such as neural networks or mixed integer programming classifiers were much better in such circumstances.

Keywords: Linear classifiers; Quadratic classifiers; Nearest neighbours; Kernel estimators; Fourier procedure; Mixed integer programming; Neural networks; Leave-one-out error estimator

* Corresponding author.
E-mail addresses: oasparoukhov@medai.com (O.K. Asparoukhov), w.j.krzanowski@ex.ac.uk (W.J. Krzanowski).

0167-9473/01/$ - see front matter
PII: S 0167-9473(01)00032-9

1. Introduction

We consider a classical problem of discriminant analysis: an individual is to be allocated to one of c distinct classes $\omega_1,\ldots,\omega_c$, whose members are described by an m-component vector of binary variables $\mathbf{x}=(x_1,\ldots,x_m)'$. These binary variables can be viewed equivalently as a single multinomial variable having $s=2^m$ states. A design set of n individuals is available, n_i of which are known to have come from ω_i ($n=n_1+n_2+\cdots+n_c$). Design set individual j from ω_i is described by the vector $\mathbf{x}_{ij}$ consisting of the m binary variable values, and the design set for each class can be summarised by a contingency table having s cells.

The optimal decision-theoretic solution to the problem, assuming a 0/1 cost function, is to allocate the new individual $\mathbf{x}$ to the class that has the greatest estimated posterior probability $\hat{p}(\omega_i|\mathbf{x})$. The posterior probabilities are evaluated from the estimated conditional distributions $\hat{p}(\mathbf{x}|\omega_j)$ and the prior probabilities $\hat{p}(\omega_j)$ via Bayes' theorem. The problem has received considerable attention in the literature, and many different methods exist for estimating the posterior probabilities and constructing resultant classifiers (McLachlan, 1992; Hand, 1997).

In this paper we give the results from a comparison of 13 such classifiers applied to real binary data sets. The use of real data allows us to avoid the criticism that artificial data favour particular methods (for example, the generation of data using mutually independent binary variables might be expected to favour those procedures that assume such a distribution). To achieve generalizable results, we have used five different data sets and the leave-one-out method of error rate estimation (Lachenbruch and Mickey, 1968). Attention has also been paid to the role of the number of variables in the investigation of classifier effectiveness. Three versions of each data set have been used, containing respectively "large", "moderate" and "small" numbers of variables. To obtain the latter two categories, we have performed variable reduction using the all-subsets approach based on Kullback's information divergence measure (Hills, 1967).

The 13 classifiers are described in Section 2, the five data sets are presented in Section 3, and the variable selection procedure is outlined in Section 4. The results of the experiments are tabulated in Section 5 and discussed in Section 6. Conclusions are presented in Section 7.

2. The classifiers

We have grouped the 13 classifiers into four broad categories: linear classifiers, quadratic classifiers, nearest neighbour classifiers, and other non-parametric classifiers.

2.1. Linear classifiers

The independent binary model (*IBM*). If we assume the binary variables to be independent then (Bailey, 1964; Boyle et al., 1966):

$$\hat{p}_i(\mathbf{x})=\prod_{j=1}^{m}\{\Pr(x_j=1|\omega_i)\}^{x_j}\{1-\Pr(x_j=1|\omega_i)\}^{1-x_j}.$$

This procedure is very simple and has a tendency to overoptimistic estimation of class probabilities, but the assumption of independence is rarely fulfilled (Hand, 1993).

The linear discriminant function (*LDF*). Under the assumptions of multivariate normal distributions with known parameters and equal covariance matrices in the classes, linear classifiers provide optimal classification (Anderson, 1984). Fisher's (1936) LDF, with unbiased estimates in place of unknown parameters, maximizes the ratio of the between-sample variance to the within-sample variance.

Logistic discrimination (*LD*). This semiparametric method avoids the problems of density estimation by assuming a logistic form for the conditional probability $\hat{p}(\omega_i|\mathbf{x})$ (Cox, 1966; Day and Kerridge, 1967; Anderson, 1972). An adjustment to this form is then made to compensate for any difference between the proportions of classes in the design set and the assessed incidence rate.

Mixed integer programming—based classification (*MIP*). Recently, various non-parametric mathematical programming (MP)-based techniques have attracted considerable research attention (Stam, 1997). They facilitate a geometric interpretation and have intuitive appeal because of their potentially robust properties. A number of studies (Duarte Silva, 1995; Joachimsthaler and Stam, 1988, 1990; Koehler and Erenguc, 1990; Rubin, 1990) have confirmed that MP methods can indeed yield effective classification rules under certain non-normal data conditions, for instance if the data set is outliers-contaminated or highly skewed. However, the MP-based approach in general lacks a probabilistic foundation, so involves an ad hoc assessment of its classification performance.

Asparoukhov and Stam (1997) proposed the following mixed integer programming formulation for linear two-class classifier construction in the presence of binary variables:

$$\begin{aligned}
&\min \sum_s (|n_{1s} - n_{2s}|z_s + \min(n_{1s}, n_{2s})) \\
&\text{subject to} \\
&\begin{array}{ll} \mathbf{x}_s^T\theta - Mz_s \leqslant c & \text{if } n_{1s} \geqslant n_{2s}, \\ \mathbf{x}_s^T\theta + Mz_s > c & \text{if } n_{1s} < n_{2s}, \end{array} \qquad n_{1s} + n_{2s} \neq 0, \\
&z_s \in \{0, 1\}
\end{aligned}$$

where the cut-off value $c \in \mathbf{R}$ and the components of the weight vector $\theta \in \mathbf{R}^m$ have to be calculated; $\mathbf{x}_s$ is the sth cell binary vector; n_{is} is the number of sth cell design set observations from the class ω_i, $i = 1, 2$; M is a sufficiently large positive number. This formulation not only has a geometric interpretation, but is also inspired from the Bayesian decision theoretic approach for the case of a 0/1 cost function and prior probabilities equal to n_i/n.

2.2. Quadratic classifiers (or classifiers with quadratic terms)

The quadratic discriminant function (*QDF*). Under the same assumptions as for the LDF, but with unequal covariance matrices in the classes, optimal classification is based on Smith's (1947) QDF.

The second-order log-linear model (*LLM*(2)). Log-linear models are well-known techniques for analysis of contingency tables, and allow $\log \hat{p}_i(\mathbf{x})$ to be estimated as a linear function of main effects and interactions between binary variables (Agresti, 1990). However, these models often require awkward decisions to be made as to which main effects and interactions may be set to zero, and involve subsequent fitting of each class model. Empty cells can cause troublesome problems in the estimation.

The second-order Bahadur model (*Bahadur*(2)). This uses first-order correlation terms:

$$\hat{p}_i(\mathbf{x}) = \prod_{j=1}^{m} \theta_{ij}^{x_j}(1-\theta_{ij})^{1-x_j} \left\{ 1 + \sum_{j<k} \rho_{ijk} \frac{(x_j-\theta_{ij})(x_k-\theta_{ik})}{\sqrt{\theta_{ij}(1-\theta_{ij})\theta_{ik}(1-\theta_{ik})}} \right\},$$

where

$$\rho_{ijk} = \underset{i}{E} \left\{ \frac{(x_j-\theta_{ij})(x_k-\theta_{ik})}{\sqrt{\theta_{ij}(1-\theta_{ij})\theta_{ik}(1-\theta_{ik})}} \right\}, \quad \theta_{ij} = \mathrm{P_r}(x_j = 1 | \omega_i) \text{ (Bahadur, 1961).}$$

2.3. Nearest neighbour classifiers

Hills' nearest neighbour estimator (*kNN—Hills*). In order to avoid problems with empty cells Hills (1967) proposed the following smoothed estimates of the probabilities:

$$\hat{p}_i(\mathbf{x}) = n_i^{-1} \left\{ \sum_{j=0}^{L} n_{ij}(\mathbf{x}) \right\} \Big/ \left\{ \sum_{j=0}^{L} \binom{m}{j} \right\}, \quad i = 1, \ldots, c,$$

where $n_{ij}(\mathbf{x})$ is the number of design set individuals $\mathbf{x}_{ij}$ from ω_i whose 'distance' from (number of disagreements with) $\mathbf{x}$, defined by $d(\mathbf{x}, \mathrm{x}_{ij}) = (\mathbf{x}-\mathbf{x}_{ij})'(\mathbf{x}-\mathbf{x}_{ij})$ is equal to j; L is called—the order of the procedure.

The adaptive weighted near neighbour estimator (*kNN—Hall*). Hall (1981b) proposed an estimator of the form $\hat{p}_i(\mathbf{x}) = n_i^{-1} \sum_{j=0}^{L} w_{ij} n_{ij}(\mathbf{x})$. The weights for $\mathbf{w}_i = (w_{i0}, w_{i1}, \ldots, w_{iL})'$ are chosen to minimize $\Delta(w_{i0}, w_{i1}, \ldots, w_{iL}) = \sum_{b=1}^{2^m} E\{\hat{p}_i(\mathbf{b}) - p_i(\mathbf{b})\}$, where $\mathbf{b}$ denotes a multinomial cell.

It is possible for Hall's estimator sometimes to give negative estimates, but this usually arises when the probabilities are small and the design set is not sufficiently large.

2.4. Some other non-parametric classifiers

The kernel estimator (*Kernel*). This non-parametric estimator has the form

$$\hat{p}_i(\mathbf{x}; \lambda) = n_i^{-1} \sum_{j=1}^{n_i} \lambda^{m-d(\mathbf{x}_{ij}, \mathbf{x})}(1-\lambda)^{d(\mathbf{x}_{ij}, \mathbf{x})}, \quad \left(\frac{1}{2} \leqslant \lambda \leqslant 1 \right),$$

where λ is a smoothing parameter, estimated by a cross-validatory or 'leave-one-out' method using an estimate of the likelihood function (Aitchison and Aitken, 1976). Unfortunately this leads to an adaptive estimator which can behave very erratically when empty or near-empty cells are present. To overcome these difficulties, Hall (1981a) proposed an alternative estimator which is designed to minimize a global function of the mean squared error.

The kernel and nearest neighbour estimators are based on only one assumption—high correlation between nearby cells.

Fourier procedure. Ott and Kronmal (1976) proposed a model based on an orthogonal expansion of the density in terms of discrete Fourier series: $\hat{p}_i(\mathbf{x}) = 2^{-m}\sum_r d_{ir}\varphi_r(\mathbf{x})$, where $d_{ir} = 2^{-m}n_i^{-1}\sum_{j=1}^{n_i}\varphi_r(\mathbf{x}_{ij})n_{i0}(\mathbf{x})$ and $\varphi_r(\mathbf{x})$ is the rth Walsh function defined by $\varphi_r(\mathbf{x}) = (-1)^{\mathbf{x}'\mathbf{r}}$; $\mathbf{r}$ is simultaneously a binary vector and an index of a multinomial cell.

Multilayer perceptron neural networks (*MLP*). Multilayer perceptrons have become very popular recently (Ripley, 1994). The MLP architecture consists of input, hidden and output layers of fully connected units (neurons). The input units are the original classification variables. The outputs of the neurons in every layer are inputs for the neurons in the next layer. The output of every neuron is a fixed function (ϕ_{h} for the hidden layer units and ϕ_{o} for the output layer units) of a weighted sum of its inputs and a bias term. Usually ϕ_{h} is a logistic function and ϕ_{o} is a logistic, linear or threshold function. The weights of these functions are calculated by minimization of some error criterion, for example least squares. For one hidden layer MLP it has the form

$$E = \frac{1}{2}\sum_{p=1}^{n}\sum_{i=1}^{c}\left(t_i^p - \phi_{\mathrm{o}i}\left(w_0 + \sum_{h=1}^{H} w_{i\mathrm{h}}\phi_{\mathrm{h}}\left(w_{\mathrm{h}0} + \sum_{j=1}^{m} w_{\mathrm{h}j}x_j\right)\right)\right)^2,$$

where the w's are weights, H is the number of hidden layer units and t_i^p is the pth observation's desired output vector (the ith component of this vector is 1 if this observation belongs to the ith class, otherwise it is 0). The most widely used technique for the minimization of MLP error criterion is the so-called back-propagation algorithm (Hertz et al., 1991).

Learning vector quantization neural networks (*LVQ*). LVQ neural networks (Kohonen, 1990) drastically reduce the number of computations at every classification decision. Each class is represented by a relatively small number of codebook vectors $\mathbf{x}_{\mathrm{c}}$, placed within each class zone such that the decision borders are approximated by the nearest neighbor rule. The classification rule is: allocate the given observation to the closest codebook class in terms of Euclidean distance. The codebook vectors are computed in a two-step iterative learning process:

(a) Initialization of the codebooks by unsupervised learning—a minimization of the average squared discretization error $\int ||\mathbf{x} - \mathbf{x}_{\mathrm{c}}||^2 p(\mathbf{x})\,\mathrm{d}\mathbf{x}$, where $p(\mathbf{x})$ is the overall probability density function.

(b) LVQ supervised learning algorithm for steepest-descent gradient-step minimization of the average squared discretization error, using the class membership information.

3. The data sets

Our purpose in this paper is simply to compare the performance of the above discrimination methods on real data sets. To obtain generalizable results we need to use a range of data sets that encompass a variety of types and conditions, and we hope that the five sets detailed below provide just such a variety. They are all recent examples, the binary variables encompass all types from straight yes/no responses to dichotomized quantitative scales, the groups range from clearly defined disease categories to overlapping binary splits of the data, and the objectives of each study go beyond the confines of simple classification. We stress, however, that we do not attempt to give full (or even apposite) analyses of the data as presented, or to answer any of the underlying research questions. Our sole intention is to use these sets as vehicles for comparison of methods. Note that in all these sets, categories of binary variables are scored as 0—no, 1—yes unless otherwise stated.

3.1. Psychology data: 16 variables

Researchers in the Department of Psychology and Pedagogy at the University of Sofia studied the ability of 184 children (aged from 5 to 6 years) to construct grammatical structures in their own Bulgarian language. As a result of the research the children were divided into three groups: 66 children who could not correctly and consciously construct grammatical structures; 31 children who could correctly, but not consciously, construct grammatical structures and 87 children who could do both of these. Variables were as follows:

1. Sex: 0—female; 1—male.
2. Do you know any language other than your maternal language?
3. Sociometric status: 0—not preferable partners; 1—preferable partners.
4. Able to sequence gender of adjectives and nouns?
5. Able to sequence number of adjectives and nouns?
6. Able to construct the present, past and future tense?
7. Able to construct words with prefixes?
8. Able to construct words with suffixes?
9. Able to construct new words out of the same root?
10. Able to construct complex words?
11. Able to construct a short simple and simple extended sentence?
12. Able to construct complex sentences with the conjunction ‘and’, ‘but’, ‘if’?
13. Able to construct a complex compound sentence with an adverbial clause of place?
14. Able to construct a complex compound sentence with an adverbial clause of reason?
15. Able to construct a complex compound sentence with an adverbial clause of time?
16. Able to construct a complex compound sentence with an adverbial clause of way?

3.2. Pulmonary data: 15 variables

These data were collected during a retrospective study aimed at the development of a nonspecific screening index for predicting patients at high risk of appearance of postoperative Pulmonary Embolism (Asparoukhov, 1985). They consisted of 390 patients who underwent surgical intervention—144 with postoperative Pulmonary Embolism and 246 without it. Variables were as follows:

1. Concomitant diseases with high risk (neoplasm, syndrome postphlebiticum, decompensated cardiac insufficiency, chronic lung disease).
2. Concomitant disease—atherosclerosis general.
3. Concomitant diseases with moderate risk (varices, compensated cardiac insufficiency, hypertonia, local atherosclerosis or myocardiosclerosis, diabetes, emphysema, superficial non-artificial thrombophlebitis).
4. Other concomitant diseases (except those enumerated in items 1, 2 and 3).
5. Obesity.
6. Reoperation.
7. Emergency surgical intervention.
8. Main disease—cancer.
9. Coma after operation.
10. Sex: 0—female; 1—male.
11. Localization of the surgical intervention in the field of abdomen cavity—abdominal, urological and big vessels (abdominal aorta and vena cava inferior) operations.
12. Localization of the surgical intervention in the field of brain and chest cavities—neurosurgical, cardiosurgical, chest and big vessels (thoracic aorta and vena cava superior) operations.
    If both 11 and 12 are scored 0, then the localization of the surgical intervention is in the field of peripheral vessels, extremities and spinal cord.
13. Age: 0—age < 60; 1—age $\geqslant 60$.
14. Smoker.
15. Blood group: 0—A, B, AB; 1—O.

Variables 11, 12, 13 and 15 were transformed into binary form following consultations with the physicians who carried out the research.

3.3. Thrombosis data: 15 variables

These data were collected during a prospective study aimed at the development of a specific index for risk assessment of the presence/absence of Embolic Thrombosis for patients with thrombophlebitis. They consisted of 102 patients—34 with Embolic Thrombosis and 68 without it. Variables were as follows:

1. Sex: 0—female; 1—male.
2. Concomitant disease—neoplasm.
3. Concomitant disease—syndrome postphlebiticum.

4. Concomitant disease—chronic lung disease.
5. Concomitant disease—atherosclerosis.
6. Concomitant disease—hypertonia.
7. Concomitant disease—diabetes.
8. Concomitant disease—varices.
9. Concomitant disease—superficial non artificial thrombophlebitis.
10. Status: 0—well; 1—not well.
11. Blood group: 0—O; 1—A, B, AB.
12. Reoperation.
13. Platelets: 0—norm ($\leqslant 170\,000$); 1—over the norm ($>170\,000$).
14. F1+2: 0—norm ($\leqslant 10$); 1—over the norm (>10).
15. D-dimer level: 0—under the norm (<500); 1—norm and over the norm (>500).

Variables 11, 13 and 15 were transformed into binary form following consultation with the physicians who carried out the research.

3.4. Epilepsy data: 15 variables

The purpose of this study was to develop a moderate-term prognosis for Craniocerebral Trauma Epilepsy in children aged up to 14 years (Vladimirova et al., 1995). The data consisted of 129 patients—81 without the condition and 48 with it. Variables were as follows:

1. Sex: 0—female; 1—male.
2. Pregnancy: 0—normal; 1—pathological.
3. Delivery: 0—normal; 1—pathological.
4. Previous psychoneurogical diseases.
5. Consciousness disturbances.
6. Headache.
7. Vomiting.
8. Treatment of trauma disturbances.
9. Skull fractures and cutaneous injuries.
10. True closed cerebral trauma.
11. New complains during the first posttraumatic year.
12. Seizures during the first posttraumatic year.
13. Control neurological status: 0—normal; 1—abnormal.
14. Paroxysmal abnormalities in control EEG.
15. Non-paroxysmal abnormalities in control EEG.

3.5. Aneurysm data: 17 variables

This data set consisted of 242 patients from the National Center of Emergency Medicine in Bulgaria, 102 of whom were diagnosed with Dissecting Aneurysm and 140 were diagnosed with other similar diseases (40 with pulmonary embolism, 50

with angina pectoris and 50 with myocardial infarction). Variables were as follows:

1. Pain.
2. Haemopericardium.
3. Haemothorax.
4. Haememediastinum.
5. Hydrothorax.
6. Haemoptoe.
7. Haematemesis.
8. Rectorrhagia.
9. Consciousness disturbances.
10. Bronchospasm.
11. Circulation disturbances.
12. Aortic murmur.
13. Albuminuria.
14. Oligoanuria.
15. Pulsation haematoma.
16. Rhythm and conduction disturbances.
17. Heard frequency changes.

4. Varying the size of data set

All five data sets have between 15 and 17 variables, which might be considered a "large" number in typical discrimination problems. In order to investigate whether the number of variables in the data set affects the performance of a particular discrimination method, we chose several subsets of variables for each data set to reflect different "sizes" of data. Ten variables was deemed to be a "moderate" number, and six a "small" number. To provide the best subsets for comparison of discrimination methods, we used an all-subsets search (McCabe, 1975) together with the method proposed by Hills (1967) to measure discrimination between two (respectively, three) populations based on any subset $x_1, x_2, \ldots, x_p$ of $x_1, x_2, \ldots, x_m$ (Kullback's, 1952 information divergence between distributions):

$$D(x_1, \ldots, x_p) = \sum_{i=1}^{s} (p_{i1} - p_{i2}) \ln p_{i1}/p_{i2}, \quad p_{i1}, p_{i2} \neq 0,$$

$$D(x_1, \ldots, x_p) = \sum_{i=1}^{s} \left\{ \sum_{j=1}^{3} \frac{(p_{ij} - p_i/2)}{p_i} \right\}, \quad p_i = p_{i1} + p_{i2} + p_{i3},$$

where $s = 2^m$, p_{ij} $(i = 1, \ldots, s; j = 1, 2)$ are the multinomial parameters derived from $x_1, x_2, \ldots, x_p$.

Of course, in practice, variable selection might form part of one's overall analysis strategy, particularly for some of the discrimination methods (such as IBM). However, to include variable selection as an extra factor in the present set of experiments would introduce a further layer of complexity and would make comparison of methods more difficult. For reasons of simplicity, therefore, we restrict comparisons to the three fixed numbers of variables.

5. Experimental results

The number of variables (nv) in each of the five sets is categorized as 'large', because each of the sets has at least 15 variables. The results of the variable selection

Table 1
Some characteristics of the data sets and the experiments. Legend: No—number

	Data set				
	Psychology	Pulmonary	Thrombosis	Epilepsy	Aneurysm
No. variables	16	15	15	15	17
No. individuals	184	390	102	129	242
No. classes	3	2	2	2	2
No. non-empty cells	169	212	76	119	121
Hall's smoothing	0.943	0.942	0.972	0.948	0.962
parameter for kernel	0.948	0.983	0.946	0.938	0.988
estimator	0.951				
The 10 variables	1,4,5,7,9,10, 11,12,14,15	1,2,3,5,6,8, 9,10,11,15	1,3,4,5,7,9, 10,11,13,15	1,2,5,6,7, 8,10,11,12,14	1,2,3,4,9,11, 12,13,15,16
Kullback measure	5.922	1.501	1.922	2.000	1.892
No. non-empty cells	142	139	60	99	62
Hall's smoothing	0.905	0.935	0.968	0.938	0.971
parameter for kernel	0.914	0.982	0.940	0.925	0.990
estimator	0.939				
The six variables	7,8,11,12,14,15	1,2,3,5,6,9	1,6,8,9,12,13	4,6,7,8,11,12	1,2,3,4,15,17
Kullback measure	4.605	1.155	1.491	1.674	1.821
No. non-empty cells	50	34	43	48	20
Hall's smoothing	0.941	0.970	0.980	0.861	0.993
parameter for kernel	0.976	0.999	0.881	0.894	0.999
estimator	0.936				

procedure to find the best "moderate" and "small" subsets in each case are given in Table 1, which also contains additional information about the number of non-empty cells and the values of kernel smoothing parameters, calculated according to Hall (1981a).

The effectiveness of a classifier can be measured in various ways. In our case, the sample sizes are too small to allow splitting into separate training and test sets with any confidence either about stable discriminant rules or accurate error rate estimates, so we had to resort to sample re-use methods. While bootstrap methods are possible, we decided to use the *leave-one-out* (LOO) *error rate* $\sum_i p(\omega_i)e_i/n_i$, where $p(\omega_i)$ is the ith class prior probability and e_i is number of leave-one-out misclassifications in the ith class. However, prior probabilities must be estimated. The usual method is either to assume equal prior probabilities in the classes, or to use the design set sample sizes in the estimation. To cover both eventualities we carried out six experiments for every data set, taking $p(\omega_i) = 1/c$ and $p(\omega_i) = n_i/n$ in turn with each of the three settings of nv.

We used simple MLP architecture consisting of one hidden layer and fully connected units. The input unit number is equal to the variable number and the output unit number is equal to the class number. The experiments were carried out with 3, 4, 5 and 6 units in the hidden layer (but we quote only the best results). We used

Table 2
Leave-one-out error rate (multiplied by 10^3) for large (15–17) number of variables[a]

Discriminant procedure	Data set									
	Psychology		Pulmonary		Thrombosis		Epilepsy		Aneurysm	
	$=$	$\neq$	$=$	$\neq$	$=$	$\neq$	$=$	$\neq$	$=$	$\neq$
Linear classifiers										
IBM	242	255	144	146	250	265	201	209	22	21
LDF	237	245	141	159	279	294	168	163	48	41
LD	251	277	170	159	324	255	224	217	56	50
Quadratic classifiers										
Bahadur(2)	190	196	163	164	257	235	207	209	34	25
QDF	—	—	—	—	—	—	278	264	—	—
Nearest neighbour classifiers										
kNN-Hills(L)	235(4)	223(3)	295(3)	277(3)	265(3)	275(3)	203(5)	248(5)	299(3)	318(3)
kNN-Hall(L)	213(5)	212(3)	201(4)	208(3)	243(4)	245(4)	224(3)	202(3)	110(3)	103(3)
Other non-parametric classifiers										
Kernel	204	190	190	172	243	265	220	217	65	54
Fourier	367	348	253	239	265	245	346	326	68	62
MLP(4)		261		156		245		186		70
LVQ(c)		228(4)		213(2)		245(4)		209(4)		58(2)

[a] $=$—equal prior probabilities; $\neq$—prior probabilities are equal to the class individuals' number divided by the design set size; kNN-Hills(L) and kNN-Hall(L)—L is order of the procedure; MLP(h)—h is hidden layer neuron number; LVQ(c)—c is number of codebooks per class.

the BFGS quasi-Newton method (Gill et al., 1981) for the least squares minimization and replaced the 0/1 values of the variables with $-0.5/0.5$ (Wasserman, 1989).

The experiments with LVQ were carried out with 2, 4 and 6 codebook vectors (we quote only the best results in Tables 2–4).

The LLM(2) method was applied only for the 'moderate' and 'small' settings of *nv*.

The MIP classification was done using LINDO (Schrage, 1991). We carried it out only for the 'moderate' and 'small' settings of *nv* with two classes, since this classification requires a lot of computational time. Note, however, that for this method the solution might not be unique. Let us assume that for the leave-one-out classification of the *i*th individual we have s_i solutions that are different (from a classification point of view), for $e(s_i)$ of which it is misclassified. Then $e_i = e(s_i)/s_i$ in contrast to all other methods, where $e_i = 0$ or 1.

For the other procedures we used our own programs. All experiments were carried out with a Pentium 200 MHz PC. The results are presented in Tables 2–4. Results for nearest neighbour classifiers are given for order $L \geqslant 1$ (2, 3) in the case of small (moderate, large) *nv*, since some of the individuals have indeterminate leave-one-out classification—they have no neighbours of order less than the above mentioned values of L.

Table 3
Leave-one-out error rate (multiplied by 10^3) for moderate (10) number of variables[a]

Discriminant procedure	Data set									
	Psychology		Pulmonary		Thrombosis		Epilepsy		Aneurysm	
	$=$	$\neq$	$=$	$\neq$	$=$	$\neq$	$=$	$\neq$	$=$	$\neq$
Linear classifiers										
IBM	250	283	176	139	243	282	205	209	23	25
LDF	254	266	147	139	272	265	147	171	53	46
LD	273	321	154	187	257	284	203	202	33	29
MIP		—		165		220		124		36
Quadratic classifiers										
Bahadur(2)	260	250	165	162	338	245	193	209	38	33
LLM(2)	277	250	156	169	390	353	232	186	28	25
QDF	—	—	—	—	302	294	201	202	—	—
Nearest neighbour classifiers										
kNN-Hills(L)	237(2)	239(2)	233(2)	259(2)	250(3)	314(2)	168(3)	225(3)	260(2)	310(2)
kNN-Hall(L)	252(2)	228(3)	173(2)	185(2)	154(2)	157(2)	159(3)	209(3)	77(2)	66(2)
Other non-parametric classifiers										
Kernel	238	239	170	174	184	206	201	202	60	54
Fourier	327	342	251	213	191	147	298	279	65	50
MLP(4)		261		151		196		147		62
LVQ(c)		234(4)		159(4)		216(2)		171(2)		62(2)

[a] $=$—equal prior probabilities; $\neq$—prior probabilities are equal to the class individuals' number divided by the design set size; kNN-Hills(L) and kNN-Hall(L)—L is order of the procedure; MLP(h)—h is hidden layer neuron number; LVQ(c)—c is number of codebooks per class.

We carried out 10 experiments (two for every data set, with equal and non-equal prior probabilities, respectively) for each of the three *nv* cases (large, moderate and small). Every experiment resulted in leave-one-out (LOO) error rates of between nine and 13 classifiers. To generalize the behaviour of the ith classifier for a group of experiments we introduce the following Empirical Integrated Rank (EIR):

$$\mathrm{EIR}_i(\%) = 100 \left\{ \sum_{e=1}^{e_g} \sum_{\substack{j=1 \\ j \neq i}}^{c_e} r_e(i,j) \right\} \Bigg/ \left\{ \sum_{e=1}^{e_g} (c_e - 1) \right\},$$

$$r_e(i,j) = \begin{cases} 1 & \text{if } \mathrm{LOO}_e(i) < \mathrm{LOO}_e(j), \\ 0.5 & \text{if } \mathrm{LOO}_e(i) = \mathrm{LOO}_e(j), \\ 0 & \text{if } \mathrm{LOO}_e(i) > \mathrm{LOO}_e(j), \end{cases}$$

where e_g is the number of experiments in the group, c_e is the number of classifiers compared in the eth experiment, $r_e(i,j)$ is the rank of the paired comparison between

Table 4
Leave-one-out error rate (multiplied by 10^3) for small (6) number of variables[a]

Discriminant procedure	Data set									
	Psychology		Pulmonary		Thrombosis		Epilepsy		Aneurysm	
	=	≠	=	≠	=	≠	=	≠	=	≠
Linear classifiers										
IBM	239	283	147	139	228	284	168	186	23	21
LDF	244	223	147	139	272	275	147	147	53	46
LD	229	266	147	167	302	343	193	186	28	25
MIP		—		139		325		101		23
Quadratic classifiers										
Bahadur(2)	219	201	152	139	243	343	197	155	23	21
LLM(2)	206	207	151	144	250	304	153	116	23	21
QDF	215	207	—	—	265	314	153	178	—	—
Nearest neighbour classifiers										
kNN-Hills(L)	273(2)	266(1)	187(1)	185(1)	265(1)	294(1)	153(2)	217(1)	363(1)	306(1)
kNN-Hall(L)	230(2)	217(2)	166(1)	144(1)	257(1)	324(2)	102(2)	124(1)	23(1)	21(1)
Other non-parametric classifiers										
Kernel	247	234	166	144	243	363	143	124	23	21
Fourier	271	223	166	154	316	373	177	147	23	21
MLP(3)		207		139		284		132		25
LVQ(c)		190(6)		146(6)		226(6)		109(6)		37(4)

[a] =—equal prior probabilities; ≠—prior probabilities are equal to the class individuals' number divided by the design set size; kNN-Hills(L) and kNN-Hall(L)—L is order of the procedure; MLP(h)—h is hidden layer neuron number; LVQ(c)—c is number of codebooks per class.

the ith and jth classifiers, and $\mathrm{LOO}_e(i)$ is the LOO for the ith classifier in the eth experiment.

Obviously the internal sum in the numerator is equal to the rank of the ith classifier for the eth experiment, which takes the extreme values $(c_e - 1)$ or zero when this classifier is the best or worst one for the considered experiment. EIR is thus equal to the percent of paired comparisons in which the ith classifier's LOO is less than the LOO of any other classifier participating in the considered group of experiments. It is an empirical measure of the average classifier effectiveness. Of course it is only a rough measure of the classifier's effectiveness since the rank value $r_e(i,j)$ is not based on a statistical test of significance of the difference between $\mathrm{LOO}_e(i)$ and $\mathrm{LOO}_e(j)$, but it is useful for generalizing the results of a comparatively small number of experiments as in our study.

We calculated the EIR values of all classifiers for nine groups of experiments (Table 5) for all data sets, large sample size (PE and DA) data sets and small sample size (ET, CTE) data sets respectively for every (large, moderate and small) number of variables.

Table 5
The values of empirical integrated rank (EIR)[a]

NV	Classifier	EIR	NPC	Large sample size data			Small sample size data		
				Classifier	EIR	NPC	Classifier	EIR	NPC
Large	QDF	11.11	18	kNN-Hills	00.00	32	QDF	11.11	18
	Fourier	19.51	82	kNN-Hall	21.88	32	Fourier	26.47	34
	kNN-Hills	26.22	82	Fourier	21.88	32	LD	27.94	34
	LD	39.02	82	LVQ	33.33	18	kNN-Hills	33.82	34
	LVQ	51.09	46	Kernel	46.88	32	Kernel	48.53	34
	kNN-Hall	52.44	82	MLP	55.56	18	LDF	55.88	34
	Kernel	56.71	82	LD	64.06	32	IBM	60.29	34
	MLP	59.78	46	Bahadur	75.00	32	LVQ	65.79	19
	LDF	62.80	82	LDF	79.69	32	Bahadur	70.59	34
	IBM	68.90	82	IBM	96.88	32	kNN-Hall	70.59	34
	Bahadur	76.83	82				MLP	81.58	19
Moderate	Fourier	27.55	98	kNN-Hills	2.63	38	LLM(2)	21.43	42
	kNN-Hills	31.12	98	kNN-Hall	21.05	38	kNN-Hills	33.33	42
	QDF	34.52	42	Fourier	21.05	38	QDF	34.52	42
	LD	43.88	98	Kernel	39.47	38	IBM	35.71	42
	LLM(2)	44.90	98	LVQ	47.73	22	LD	38.10	42
	Bahadur	50.51	98	MLP	52.27	22	Bahadur	38.10	42
	kNN-Hall	55.10	98	MIP	59.09	22	Fourier	45.24	42
	IBM	55.10	98	LD	63.16	38	LDF	63.10	42
	Kernel	58.16	98	Bahadur	65.79	38	Kernel	65.48	42
	LDF	63.27	98	LDF	75.00	38	LVQ	72.92	24
	MLP	65.18	56	LLM(2)	75.00	38	kNN-Hall	73.81	42
	LVQ	66.07	56	IBM	81.58	38	MIP	79.17	24
	MIP	69.57	46				MLP	87.50	24
Small	kNN-Hills	18.00	100	kNN-Hills	00.00	38	LD	14.29	42
	LD	27.00	100	LVQ	22.73	22	Fourier	17.86	42
	Fourier	31.00	100	LD	35.53	38	Bahadur	33.33	42
	LDF	51.00	100	LDF	47.37	38	kNN-Hills	39.29	42
	Kernel	51.00	100	Fourier	47.37	38	QDF	41.67	42
	QDF	54.03	62	kNN-Hall	55.26	38	IBM	54.76	42
	IBM	57.00	100	Kernel	55.26	38	Kernel	59.52	42
	Bahadur	58.00	100	MLP	56.82	22	LDF	60.71	42
	kNN-Hall	59.50	100	LLM(2)	63.16	38	kNN-Hall	65.48	42
	MLP	64.91	57	MIP	63.64	22	MIP	66.67	24
	MIP	65.22	46	Bahadur	71.05	38	LLM(2)	66.67	42
	LVQ	68.42	57	IBM	78.95	38	MLP	68.75	24
	LLM(2)	69.00	100				LVQ	95.83	24

[a]NV—number of variables; NPC—number of paired comparisons.

6. Discussion

6.1. The data sets

The minimum LOO error rates for each data set are as follows: 19.02% for Psychology; 13.85% for Pulmonary; 14.71% for Thrombosis; 10.08% for Epilepsy and

2.07% for Aneurysm. Thus the groups in the last data set are very well separated and those in all other data sets are well separated. What is more—from a practical point of view we can say that all the medical data sets (i.e. all except Psychology) are very well separated since their minimum LOO error rates are all less than 15%.

The physicians felt that the Pulmonary and Aneurysm data sets contained either uncorrelated or weakly correlated variables, and this is supported by the greatest calculated values of EIR for IBM coming from these sets. Most of the variables of the Thrombosis data are similarly weakly correlated, the exceptions being the last three blood parameters (Platelets, F1 + 2 and D-dimer level). The small *nv* version of the Thrombosis data can thus be considered as having uncorrelated or weakly correlated variables since it includes only one (Platelets) of these three latter parameters. More variables of the Psychology and Epilepsy data sets are obviously correlated. The correlations between the variables could easily be tested by standard statistical criteria.

6.2. The classifiers

Independent binary model. This classifier gives the best result (minimum LOO) for all the experiments with the Aneurysm data, four experiments with the Pulmonary data and one of the small *nv* experiments with Thrombosis data. The stronger correlations between the binary variables for the other experiments leads to a decrease of the classifier effectiveness. It is a highly effective classifier for large sample sizes (EIR is about 80% for small and moderate *nv* and over 95% for large *nv*). It thus seems that IBM should be used only in the cases of generally uncorrelated or weakly correlated variables, or when *nv* is large. Similar findings were reported by Asparoukhov and Danchev (1997), Nordyke et al. (1971), Moore (1973), Ott and Kronmal (1976) and Young et al. (1981).

Linear discriminant function. The LDF is frequently applied to binary variables (Gilbert, 1968; Moore, 1973; Krzanowski, 1977; Trampisch, 1978; Dillon and Goldstein, 1978; Goldstein and Dillon, 1978; Hand, 1983; Ganeshanandam and Krzanowski, 1989, 1990) and almost all the results suggest that it can be recommended because of its expected stability as the number of variables increases (the exceptions being special cases such as the "reversal" of l.l.r. quoted by Moore). What is more, the losses incurred by the use of Fisher's LDF under non-optimal conditions compared with other procedures are small enough not to be of any practical importance.

In our study LDF gives the best result for three large *nv* experiments, three moderate *nv* experiments (in one of them together with IBM), and two small *nv* experiments (together with several other classifiers). It is very effective (EIR $>$ 75%) in the case of large and moderate *nv* if the sample size is large. Obviously small numbers of binary variables or small sample sizes can decrease the accuracy of the LDF.

It should be noted that both LDF and IBM give poor results in all the experiments with the Psychology data irrespective of *nv* setting. This fact is in agreement with the conclusion of Dillon and Goldstein (1978): "The presence of correlated variables can decrease the accuracy of classification drastically in the case of the first-order Bahadur and LDF procedures".

Logistic discrimination. LD performance has been studied by a number of authors in relation to its close affinity with LDF (McLachlan, 1992, Chapter 8). The general consensus (Krzanowski, 1988) is that LD is to be preferred to LDF when the distributions are clearly non-normal (as with binary variables) or when the dispersion matrices are clearly unequal. Surprisingly, in our study the comparison between LD and LDF definitively favours the latter. The advantage is greater for the small sample sizes than for those sets (Pulmonary, Aneurysm) with large size. However, the exceptions are those data sets with smallest number of non-empty multinomial cells. In these cases LD gives either better results (for the Aneurysm data with small or moderate nv) or comparable results (for the Thrombosis data with large or moderate nv and for the Pulmonary data with small nv) to the LDF. It should be noted that there are greater differences between the LOO error rates for the two kind of prior probabilities for LD than for LDF. It seems that LD is more sensitive than the LDF to the sample size and the values of the prior probabilities.

Mixed integer programming—based classification. The MIP classifier participates in eight experiments, in three of which its result is the best and in one it is the second-best. From the EIR values we can conclude that it is: (a) the most effective classifier for moderate nv; (b) among the three most effective classifiers for small nv; (c) obviously better than the other linear classifiers (in accordance with the conclusions of Asparoukhov and Stam, 1997; Asparoukhov and Danchev, 1997) and (d) more effective for the small sample size data than for sets with large sample size. It should be considered a serious contender for applications having moderate or small numbers of binary variables. Unfortunately, all known mixed integer programming-based formulations are very time consuming since they are NP-hard, and there is no hope for fast (polynomial time) algorithms to be obtained for their solution unless $P = NP$. The lack of mixed integer programming classification "branch and bound" algorithms and software that takes into account the specificity of the binary variables, strongly limits the application of this promising approach for the solution of classification problems. This is in contrast to the case of continuous variables—see Duarte Silva and Stam (1997) and Rubin (1997).

Quadratic discriminant function. In some data sets there are binary variables for which either all individuals from one of the classes have the same value or only one individual has a different value from the rest. In these cases the calculations are impossible due to singularity of covariance matrices. In the other cases QDF rarely performs as well as LDF; the same result was obtained by Moore (1973). Consideration of the EIR values suggests that this classifier is not effective for large nv, or moderate nv and small sample size data. We do not recommend its use with binary variables.

Second-order Log-linear model. This classifier gives the best result for three experiments with small nv and the second-best result for two experiments with small nv. Clearly this is a very effective procedure for: (a) small nv irrespective of the sample size (where EIR values are greatest) and (b) moderate nv for large sample size data (where EIR values are among the three greatest).

The use of the LLM (2) classifier may be limited because of the presence of empty multinomial cells and the awkward decision as to which main effects and

interactions to include. However its major problem is connected with the number of variables, as most well-known statistical packages only provide a log-linear analysis for a small number of variables.

Second-order Bahadur model. This procedure gives the best result for two experiments and the second-best result for three experiments in the case of large *nv*. For small *nv* it is best in three cases (jointly with other classifiers) and second-best in two, but it never achieves minimum LOO for data with moderate *nv* or for those data sets that have the greatest number of non-empty cells irrespective of *nv*—Pulmonary (Table 2), Psychology and Pulmonary (Table 3), Psychology and Epilepsy (Table 4). The EIR values show it to be very effective for (a) large *nv* irrespective of the sample size and (b) small *nv* for a large sample size (EIR $> 70\%$).

kNN-Hills. For this classifier, the LOO is minimum only for the moderate *nv* Psychology data (equal prior probabilities). Judging by the EIR values this classifier is among the three worst for every *nv* and it is extremely ineffective for the large sample size data. We do not think it should be taken into account in future applications.

kNN-Hall, Kernel (*with Hall's smoothing parameter*) *and Fourier classifiers.* We group these classifiers in our discussion since they showed a common property. It is well known that the resubstitution error (RES—the probability of misclassification obtained when the design set is reclassified) is an unreliable estimate of the classification performance and this is the reason we have not included it in our study. However, it should be noted that the minimum RES is always obtained by the kNN-Hall and in most cases by the Kernel and Fourier classifiers, the RES of the other classifiers being substantially greater. This minimum is also significantly less than their corresponding LOO for all data sets with the exception of the Aneurysm data. *It is evident that these three classifiers are very flexible and have a tendency to overfit the data.* Hand (1983) came to the same conclusion about the Kernel estimator on the basis of the experimental study of six multivariate binary data sets.

The experimental results allow us to draw the following conclusions about these non-parametric classifiers:

1. kNN-Hall and Kernel give very similar results. The kNN-Hall LOO is the best for six experiments and second-best for four experiments, while the Kernel LOO is three times the best and seven times the second-best. General EIR values (between 50% and 60%) show both classifiers to be of average quality, but as regards the sample size EIR values we are able to say that:

- kNN-Hall is very effective in the cases of large and moderate *nv* for small sample size and is definitely not effective in the cases of large and moderate *nv* for large sample size;
- the Kernel classifier is effective for moderate and small *nv* if the sample size is small.

2. The Fourier classifier gives minimum LOO once for moderate *nv* (Thrombosis data) and twice for small *nv* (Aneurysm data). If we add to them one second-best result for large *nv* (Thrombosis data) we see that these good results are always obtained for data with the smallest number of non-empty cells. In all other

experiments its performance is much poorer, and in eight of the 30 experiments it is the worst. The EIR values of the Fourier classifier are among the poorest three. It seems that this procedure has not yet been adapted to cope with high-dimensional data (Asparoukhov and Andreev, 1995; Titterington et al., 1981).

Multilayer perceptron neural networks. MLP features in 15 experiments, in one of which its result is the best and in five it is second-best, but in all other nine experiments its LOO is close to the smallest one irrespective of the number of non-empty multinomial cells. The EIR values show that: (a) it is among the best four classifiers with EIR very close to or over 60%; (b) in the case of small sample size it is very effective, with the greatest EIR over 80%, for moderate and large *nv*, and effective, with the second-best EIR values close to 70%, for small *nv*; (c) it is more effective than LVQ for large *nv*. However, it should be noted that application of MLP to concrete data involves several decisions—choice of number of hidden layers, number of neurons in these layers, and procedure for least squares minimization. The choice of these factors may decrease or increase (sometimes dramatically) the MLP's accuracy.

Learning vector quantization neural networks. LVQ also features in 15 experiments, in two of which its result is the best and in three of which it is second-best. In all other 10 experiments its LOO is close to the minimum irrespective of the number of non-empty multinomial cells. From the EIR values we see that: (a) it is one of the two best classifiers with EIR greater than 60% for moderate and small *nv*; (b) in the case of small sample size it is extremely effective, with the greatest EIR over 95%, for small *nv* and among the best four classifiers for moderate and small *nv*; (c) it is definitely not effective for large sample size.

6.3. Summary

Let us divide our classifiers into two groups—the group of non-traditional (recent) classifiers (MIP, MLP, LVQ) and the group of traditional classifiers (all others). One notable feature is that the average EIR values of the traditional classifiers decrease with decreasing number of variables, while the average EIR values of the non-traditional classifiers increase as this number decreases. Careful survey of the sample size EIR values also allows us to draw the following conclusions:

(a) All the very effective (EIR $>70\%$) classifiers for large samples are traditional ones (IBM, Bahadur (2), LDF and LLM(2));

(b) Most of the very effective classifiers for small samples (MLP, LVQ, MIP, Hall, Bahadur (2)) are non-traditional ones; the best classifiers (EIR $>80\%$) are always the non-traditional.

A challenging, and maybe heretical, generalization of the above two conclusions could be that most of the traditional statistical classifiers do not cope with cases of small sample (sparse) binary data while the non-traditional (MLP, LVQ, MIP) classifiers are much better in these circumstances. Of course we do not consider our experimental study exhaustive enough to draw such a general conclusion; however there are apparently good reasons to believe in the existence of such a trend.

Of course the well-developed variable selection procedures that are implemented in the known packages constitute a major advantage for some traditional statistical classifiers (LDF, LD, LLM).

7. Conclusions

We have studied a variety of classifiers in the presence of large, moderate and small number of variables for five real binary data sets. While these sets were all ones that arose in the course of day-to-day analysis of medical data, certain features militate against the drawing of widespread conclusions about the classifiers. First, the numbers of individuals were very similar in four of the five studies, and overall the range of sample sizes was not as great as might be met in practice. Second, the sample sizes were not sufficiently large to permit a training/test split in each study and recourse had to be made to the leave-one-out measurement of error rates. This could be important in those situations where several variants of one method were used and a particular one was then selected for presentation. Finally, strong conclusions about which methods are to be preferred would require a comparison of strategies which included variable selection, and this would again require a training/test split. Nevertheless, despite these reservations, the results of the experiments lead us to the following tentative conclusions:

1. The kNN-Hall, Kernel and Fourier classifiers are very flexible but seem to have a tendency to overfit the data. Their markedly superior RES error rate can lead to a dangerous optimism for moderate/large numbers of variables. kNN-Hall seems to be very effective in the cases of large and moderate number of variables if the sample size is small, but not effective if the sample size is large. The Kernel classifier appears to be effective for moderate and small numbers of variables if the sample size is small, while the Fourier classifier is effective only for data with small numbers of non-empty multinomial cells.

2. The use of either LDF or IBM is not recommended when there are high correlations between binary variables. LDF appears to be very effective in the case of moderate/large numbers of variables if the sample size is large, but small numbers of variables and/or sample sizes can decrease its accuracy. IBM seems to be suited to generally uncorrelated or weakly correlated variables, or to cases when there are many variables. The performance of the LDF is expected to be better than that of LD, especially for small sample size data except when the number of non-empty multinomial cells is very small.

3. The second-order Bahadur procedure performs well for large numbers of variables irrespective of the sample size, and for small numbers of variables when the sample size is large. The second-order log-linear models perform well for small numbers of variables or large sample sizes. kNN-Hills and QDF do not seem to be effective on binary data.

4. The MIP classifier is clearly competitive against the other linear classifiers. However, the lack of mixed integer programming classification “branch and bound” algorithms and software that take into account the specificity of the binary variables

strongly limits the application of this promising approach. It seems that the traditional statistical classifiers are not well able to cope with small sample (sparse) binary data, but the non-traditional (MLP, LVQ, MIP) classifiers may be much better in these circumstances.

Acknowledgements

We would like to thank Professor David Hand for useful discussions and help during this investigation. The research was assisted by a Royal Society travel grant, which allowed Dr Asparoukhov to visit Exeter.

References

Agresti, A., 1990. Categorical Data Analysis. Wiley, New York.

Aitchison, J., Aitken, C.G.G., 1976. Multivariate binary discrimination by the kernel method. Biometrika 63, 413–420.

Anderson, J.A., 1972. Separate sample logistic discrimination. Biometrika 59, 19–35.

Anderson, T.W., 1984. An Introduction to Multivariate Statistical Analysis, 2nd Edition. Wiley, New York.

Asparoukhov, O.K., 1985. Microprocessor system for investigation of thromboembolic complications. Unpublished Ph.D. Dissertation, Technical University, Sofia.

Asparoukhov, O., Andreev, Tz., 1995. Comparison of one-stage classifiers for assessment of the ability of children to construct grammatical structures consciously. In: Panchev, I. (Ed.), Multivariate Analysis in the Behavioral Sciences. Philosophic to Technical. Academic Publishing House "Prof. Marin Drinov", Sofia, pp. 1–13.

Asparoukhov, O., Danchev, S., 1997. Discrimination and classification in the presence of binary variables. Biocybernet. Biomed. Eng., Pol. Acad. Sci. 17 (1/2), 25–39.

Asparoukhov, O., Stam, A., 1997. Mathematical programming formulations for two group classification with binary variables. Ann. Oper. Res. 74, 89–112.

Bahadur, R.R., 1961. A representation of the joint distribution of response to n dichotomous items. In: Solomon, H. (Ed.), Studies in Item Analysis and Prediction. Stanford University Press, Palo Alto, CA, pp. 158–168.

Bailey, N.T.J., 1964. Probability methods of diagnosis based on small samples. Mathematics and Computer Science in Biology and Medicine. H.M.S.O., London, pp. 103–107.

Boyle, J.A., Greig, W.R., Franklin, D.A., Harden, R.McG., Buchanan, W.W., McGirr, E.M., 1966. Construction of a model for computer-assisted diagnosis: application to the problem of non-toxic goitre. Quart. J. Med. 35, 565–588.

Cox, D.R., 1966. Some procedures connected with the logistic qualitative response curve. In: David, F.N. (Ed.), Research Papers in Statistics: Festschrift for J. Neyman. Wiley, London, pp. 55–71.

Day, N.E., Kerridge, D.F., 1967. A general maximum likelihood discriminant. Biometrics 23, 313–323.

Dillon, W.R., Goldstein, M., 1978. On the performance of some multinomial classification rules. J. Amer. Statist. Assoc. 73, 305–313.

Duarte Silva, A.P., 1995. Minimizing misclassification costs in two-group classification analysis. Unpublished Ph.D. Dissertation, The University of Georgia.

Duarte Silva, A.P., Stam, A., 1997. A mixed integer programming algorithm for minimizing the training sample misclassification cost in two-group classification. Ann. Oper. Res. 74, 129–157.

Fisher, R.A., 1936. The use of multiple measurements in taxonomic problems. Ann. Eugen. 7, 179–188.

Ganeshanandam, S., Krzanowski, W.J., 1989. On selecting variables and assessing their performance in linear discriminant analysis. Aust. J. Satist. 31 (3), 433–447.

Ganeshanandam, S., Krzanowski, W.J., 1990. Error-rate estimation in two-group discriminant analysis using the linear discriminant function. J. Statist. Comput. Simulation 36, 157–175.

Gilbert, E.S., 1968. On discrimination using qualitative variables. J. Amer. Statist. Assoc. 63, 1399–1412.

Gill, P.E., Murray, W., Wright, M.H., 1981. Practical Optimization. Academic Press, London.

Goldstein, M., Dillon, W.R., 1978. Discrete Discriminant Analysis. Wiley, New York.

Hall, P., 1981a. On nonparametric multivariate binary discrimination. Biometrika 68 (1), 287–294.

Hall, P., 1981b. Optimal near neighbour for use in discriminant analysis. Biometrika 68 (2), 572–575.

Hand, D.J., 1983. A comparison of two methods of discriminant analysis applied to binary data. Biometrics 39, 683–694.

Hand, D.J., 1993. Discriminant analysis for categorical data. The Lecture Notes and Program of the Fourth European Courses in Advanced Statistics Program, "Analysis of Categorical Data Theory and Application". Leiden, The Netherlands, pp. 135–174.

Hand, D.J., 1997. Construction and Assessment of Classification Rules. Wiley, Chichester.

Hertz, J., Krogh, A., Palmer, R.G., 1991. Introduction to the Theory of Neural Computation. Addison-Wesley, Redwood City.

Hills, M., 1967. Discrimination and allocation with discrete data. Appl. Statist. 16, 237–250.

Joachimsthaler, E.A., Stam, A., 1988. Four approaches to the classification problem in discriminant analysis: an experimental study. Decision Sci. 19, 322–333.

Joachimsthaler, E.A., Stam, A., 1990. Mathematical programming approaches for the classification problem in two-group discriminant analysis. Multivar. Behav. Res. 25, 427–454.

Koehler, G.J., Erenguc, S.S., 1990. Minimizing misclassifications in linear discriminant analysis. Decision Sci. 21, 63–85.

Kohonen, T., 1990. The self-organizing map. Proceeding of the IEEE 78, 1990, 1464–1480.

Krzanowski, W.J., 1977. The performance of Fisher's linear discriminant function under non-optimal conditions. Technometrics 19, 191–200.

Krzanowski, W.J., 1988. Principles of Multivariate Analysis: a User's Perspective. Oxford University Press, New York.

Kullback, S., 1952. Information Theory and Statistics. Wiley, New York.

Lachenbruch, P.A., Mickey, M.R., 1968. Estimation of error rates in discriminant analysis. Technometrics 10, 1–11.

McCabe Jr., G.P., 1975. Computations for variable selection in discriminant analysis. Technometrics 17, 103–109.

McLachlan, G.J., 1992. Discriminant Analysis and Statistical Pattern Recognition. Wiley, New York.

Moore, D.H., 1973. Evaluation of five discriminant procedures for binary variables. J. Amer. Statist. Assoc. 68, 399–404.

Nordyke, R.A., Kulikowski, C.A., Kulikowski, C.W., 1971. A comparison of methods for the automated diagnosis of thyroid disfunction. Comput. Biomed. Res. 4, 374–389.

Ott, J., Kronmal, R.A., 1976. Some classification procedures for multivariate binary data using orthogonal functions. J. Amer. Statist. Assoc. 71, 391–399.

Ripley, B., 1994. Neural networks and related methods for classification. J. Roy. Statist. Soc. Ser. B 56 (3), 409–456.

Rubin, P.A., 1990. Heuristic solution procedures for a mixed-integer programming discriminant model. Managerial Decision Econom. 11, 255–266.

Rubin, P., 1997. Solving mixed integer classification problems by decomposition. Ann. Oper. Res. 74, 51–64.

Schrage, L., 1991. LINDO: User's Manual. Release 5.0. The Scientific Press, South San Francisco, CA.

Smith, C.A.B., 1947. Some examples of discrimination. Ann. Eugen. 13, 272–282.

Stam, A., 1997. MP approaches to classification: Issues and Trends. Ann. Oper. Res. 74, 1–36.

Titterington, D.M., Murray, G.D., Murray, L.S., Spiegelhalter, D.J., Skene, A.M., Habbema, J.D.F., Gelpke, G.J., 1981. Comparison of discriminant techniques applied to a complex data set of head injured patients (with discussion). J. Roy. Statist. Soc. Ser. A 144, 145–175.

Trampisch, H.J., 1978. Classical discriminant analysis and Lancaster models for qualitative data. Compstat 1978, Proceedings in Computational Statistics. Physica-Verlag, Vienna, pp. 205–211.

Vladimirova, G., Asparoukhov, O., Nikova, V., Andreev, Tz., Daskalov, Hinkov, Chr., 1995. Follow-up, analysis and prognostication of the craniocerebral trauma outcome in children aged up to 14. Proceedings of the First Bulgarian International Symposium on Cardiovascular Diseases and Pediatric Trauma. Sofia, Bulgaria, 1995, pp. 283–296.

Wasserman, P.D., 1989. Neural Computing: Theory & Practice. Van Nostrand Reinhold, New York.

Young, T.Y., Liu, P.S., Rondon, R.J., 1981. Statistical pattern classification with binary variables. IEEE Trans. Pattern Anal. Mach. Intell. 2, 155–163.