INFORMATICA, 2000, Vol. 11, No. 2, 115–124

Influence of Projection Pursuit on Classification Errors: Computer Simulation Results

Gintautas JAKIMAUSKAS
Institute of Mathematics and Informatics
Akademijos 4, 2600 Vilnius, Lithuania
e-mail: gnt@ktl.mii.lt

Ričardas KRIKŠTOLAITIS
Vytautas Magnus University
Vileikos 8, 3000 Kaunas, Lithuania
e-mail: ricardas_krikstolaitis@fc.vdu.lt

Received: January 2000

Abstract. Influence of projection pursuit on classification errors and estimates of a posteriori probabilities from the sample is considered. Observed random variable is supposed to satisfy a multidimensional Gaussian mixture model. Presented computer simulation results show that for comparatively small sample size classification using projection pursuit algorithm gives better accuracy of estimates of a posteriori probabilities and less classification error.

Key words: Gaussian mixture model, projection pursuit, classification.

1. Introduction

The most common method of estimating a posteriori probabilities in classification is to replace unknown parameters by the maximum likelihood estimates (MLE). One of methods used to reduce variance of the estimates is to reduce dimension of observations by projecting them to a subspace of lower dimension and to calculate MLE in the subspace.

Theoretical background of this problem is given, e.g., in (Aivazyan *et al.*, 1989) and (Friedman and Tukey, 1974). Some computer simulation results are presented in (Jakimauskas and Krikštolaitis, 1999). In this paper we present more computer simulation results that show that for comparatively small sample size classification using projection pursuit algorithm gives better accuracy of estimates of a posteriori probabilities and less classification error. We are thankful to prof. R. Rudzkis who gave the idea of these papers and many constructive and valuable remarks.

The introduction presents already known methods. Description of projection pursuit algorithm see, e.g., in (Rudzkis and Radavičius, 1997). Further studies are on the way, which will help to make practical decision (probably using bootstrap methods) in which situations use of projection pursuit algorithm is preferable (including computational costs of finding discriminant subspace).

Main definitions

Let we have q independent d-dimensional Gaussian random variables Y_i with different distribution densities $\varphi(\cdot\,; M_i, R_i) \stackrel{\text{def}}{=} \varphi_i$, where means M_i and covariance matrices R_i, $i = 1, 2, \ldots, q$, are unknown. Let ν be random variable (r.v.) independent of Y_i, $i = 1, 2, \ldots, q$, and taking on values $1, 2, \ldots, q$ with unknown probabilities $p_i > 0$, $i = 1, 2, \ldots, q$, respectively. In this paper we assume that number of classes q is known. We observe d-dimensional r.v. $X = Y_\nu$. Each observation belongs to one of q classes depending on r.v. ν. Distribution density of r.v. X is therefore a Gaussian mixture density

$$f(x) = \sum_{i=1}^{q} p_i \varphi_i(x) \stackrel{\text{def}}{=} f(x, \theta), \quad x \in \mathbf{R}^d, \tag{1}$$

where $\theta = (p_i, M_i, R_i,\ i = 1, 2, \ldots, q)$ is an unknown multidimensional parameter. Probabilities $p_i = \mathbf{P}\{\nu = i\}$ are *a priori* probabilities for r.v. X to belong to ith class.

We will consider the general classification problem of estimating *a posteriori* probabilities $\pi(i, x) = P\{\nu = i | X = x\}$ from the sample $\{X_1, X_2, \ldots, X_N\} \stackrel{\text{def}}{=} X^N$ of i.i.d. random variables with distribution density (1). Under assumptions above

$$\pi(i, x) = \pi_\theta(i, x) = \frac{p_i \varphi_i(x)}{f(x, \theta)}, \quad i = 1, 2, \ldots, q,\ \ x \in \mathbf{R}^d. \tag{2}$$

The problem is to estimate the unknown multidimensional parameter θ.

The EM algorithm

If number of classes q is known, then the maximum likelihood estimate θ^* is an efficient estimate of θ. The most common method for calculating the MLE for Gaussian mixtures is so-called EM (Expectation Maximization) algorithm. Let $\pi^*(\cdot\,, \cdot\,) = \pi_{\theta^*}(\cdot\,, \cdot\,)$, i.e., for given θ^* corresponding a posteriori probabilities π^* are obtained using (2). Reversely, θ^* can be obtained from the given $\pi = \pi^*$ using the following equalities:

$$p_i = \frac{1}{N} \sum_{j=1}^{N} \pi(i, X_j), \quad i = 1, 2, \ldots, q, \tag{3a}$$

$$M_i = \frac{1}{N} \sum_{j=1}^{N} \frac{\pi(i, X_j)}{p_i} X_j, \quad i = 1, 2, \ldots, q, \tag{3b}$$

$$R_i = \frac{1}{N} \sum_{j=1}^{N} \frac{\pi(i, X_j)}{p_i} (X_j - M_i)(X_j - M_i)^{\mathrm{T}}, \quad i = 1, 2, \ldots, q. \tag{3c}$$

Let $\pi^N = \{\pi(i, X),\ i = 1, 2, \ldots, q,\ X \in X^N\}$ be any given a posteriori probabilities for sample data points X^N. For given π^N, the parameter $\theta = (p_i, M_i, R_i,\ i =$

$1, 2, \ldots, q)$ is obtained using (3). For given θ probabilities π^N are calculated using formula (2). The EM algorithm is an iterative procedure which starts either from a given parameter θ or given probabilities π^N applying in turn formulae (3) and (2). The EM algorithm usually ends after some predefined number of iterations. Parameter θ in the EM algorithm converges to MLE if starting parameter θ^0 is sufficiently close to θ^*.

Note that starting parameter θ^0 can be obtained from starting values of a posteriori probabilities. In higher dimensions increasing number of parameters leads to unstability of EM algorithm. Also the problem of finding the starting parameter θ^0 becomes more and more complicated.

For mixture distributions the EM algorithm was proposed independently by Schlesinger (1965), Hasselblad (1966), and Behboodian (1970). By now the properties of the EM algorithm have been studied well enough. On the convergence properties of the EM algorithm see (Wu, 1983). Also see, e.g., monographs (Aivazyan *et al.*, 1989; Everitt and Hand, 1981; McLacklan and Basford, 1988; Titterington *et al.*, 1985). For further references see (Rudzkis and Radavičius, 1995). The popularity of the EM algorithm is explained by computational stability and simplicity of implementation on a computer.

Discriminant space

Let $V = \operatorname{cov}(X, X)$ be the covariance matrix of r.v. X and suppose for simplicity $\mathbf{E}X = 0$. Define the scalar product of arbitrary vectors $u, h \in \mathbf{R}^d$ as $(u, h) = u^{\mathrm{T}} V^{-1} h$ and denote by u_L the projection of arbitrary vector $u \in \mathbf{R}^d$ to a linear subspace $L \subset R^d$. Discriminant space H is defined as a linear subspace $H \subset \mathbf{R}^d$ with the property $\mathbf{P}\{\nu = i | X = x\} = \mathbf{P}\{\nu = i | X_H = x_H\},\ i = 1, 2, \ldots, q,\ x \in \mathbf{R}^d$, and the minimal dimension. It is known that for Gaussian mixture densities (1) with equal covariance matrices we have $\dim H \leqslant q - 1$ and for most natural cases $\dim H = q - 1$.

Let $k = \dim H$ and vectors $u_1, u_2, \ldots, u_k$ be a basis in the discriminant space H. Denote $U = (V^{-1}u_1, V^{-1}u_2, \ldots, V^{-1}u_k)^{\mathrm{T}}$. Then $\pi(i, x) = \mathbf{P}\{\nu = i | UX = Ux\}$, $i = 1, 2, \ldots, q$, $x \in \mathbf{R}^d$. This means that projected sample $\{UX_1, UX_2, \ldots, UX_N\}$ is a sufficient statistics for estimating a posteriori probabilities. The distribution density of r.v. UX is a Gaussian mixture density

$$f^H(z) = \sum_{i=1}^{q} p_i \varphi_i^H(z) \stackrel{\mathrm{def}}{=} f_q^H(z, \theta_H), \quad z \in \mathbf{R}^k, \tag{4}$$

where $\varphi_i^H = \varphi(\cdot\,, M_i^H, R_i^H)$, $i = 1, 2, \ldots, q$, are k-dimensional Gaussian distribution densities with means $M_i^H = UM_i$ and covariance matrices $R_i^H = U^{\mathrm{T}} R_i U$, $\theta_H = (p_i, M_i^H, R_i^H,\ i = 1, 2, \ldots, q)$ is the multidimensional parameter.

Projection pursuit algorithm

One of methods to find discriminant space is projection pursuit algorithm. It is step-by-step procedure to find the basic vectors of discriminant space. Projection pursuit

method was introduced by Friedman and Tukey (1974). Properties of the projection pursuit method also have been studied well enough, see, e.g., (Friedman, 1987). See also (Aivazyan *et al.*, 1989) and (Rudzkis and Radavičius, 1999). Description below is based on (Rudzkis and Radavičius, 1997).

Let $\boldsymbol{F}$ be the set of one-dimensional Gaussian mixture distribution functions, $\rho = \rho(G_1, G_2)$, $G_1, G_2 \in \boldsymbol{F}$, be some functional satisfying the following conditions: $\rho(G_1, G_1) = 0$ and $\rho(G_1, G_2) > 0$, if $G_1 \neq G_2$. For arbitrary non-zero $u \in \mathbf{R}^d$ define a projection index $Q(u) = \rho(F_u, \Phi)$, where F_u is the distribution function of the standardized r.v. $u^{\mathrm{T}}X$, Φ is the standard Gaussian distribution function.

Let orthonormal vectors $u_1, u_2, \ldots, u_k$ be found step-by-step as follows: set $U_0 = \{0\}$, for $i = 1, 2, \ldots, d$, calculate $u_i = \arg\max_u\{Q(u),\ u \in U_{i-1}^{\perp},\ \|u\| = 1\}$, $U_i = span\{u_1, u_2, \ldots, u_i\}$ and stop when $Q(u_i) = 0$. We set $k = \min\{i :\ Q(u_{i+1}) = 0\}$ (by definition $Q(u_{d+1}) = 0$).

Assume that the covariance matrices of components of X are equal. If the functional ρ is shift- and scale-invariant and $\rho(G_1 * \Phi, G_2 * \Phi) < \rho(G_1, G_2)$ for any Gaussian G_2 and $G_1 \neq G_2$, $G_2 \in \boldsymbol{F}$, then the vectors $u_1, u_2, \ldots, u_k$ form a basis of the discriminant space. Note that if covariance matrices R_i are non-equal (see (Rudzkis and Radavičius, 1997)) then more complicated algorithm must be used.

Practical problems

In real calculations we use projection index estimate $\widehat{Q}(u) = \widehat{Q}(u, X^N)$ based on the sample X^N and use stopping condition $\widehat{Q}(u_i) < \varepsilon_i$, where $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_d$ is some sequence of small positive numbers satisfying certain conditions. In this paper we do not analyse the problem when to stop estimation of the discriminant space (we assume that $\dim H$ is known).

We can mention two other practical problems:

1. How much better estimates of a posteriori probabilities can be obtained when projection pursuit is used in the case $\dim H < d$;
2. Will we get better estimates (if so, how much better estimates) of a posteriori probabilities applying projection pursuit but intentionally projecting data to the lower dimension subspace (supposing that we do not lose many information about cluster structure).

Analysis of these problems can be done using computer simulation. In this paper we used computer simulation trying to get answers to the problems mentioned above.

2. Computer Simulation Results

We performed four tests intended to determine the dependency of the estimates of a posteriori probabilities and number of classification errors on the dimension d, sample size N, number of classes q, projection to discriminant subspace versus projection to lower dimension subspace. We intentionally selected simple mixture models in order to achieve maximum available differences in estimation.

We assume that we have sufficiently good starting parameters for the EM algorithm – we start from parameters that were used for simulation of the sample X^N. For projected sample we use corresponding theoretical density (4). Also we use theoretical basic vectors of the discriminant space. So presented results do not contain errors due to selection of starting parameters for the EM algorithm and errors due to finding basic vectors of the discriminant space. We performed additional test projecting data to direction estimated from the sample (using the projection pursuit algorithm based on generalized Ω^2 metrics for finding the basic vectors of the discriminant space). This test showed that for the selected mixture models the errors due to finding discriminant space are very small and insignificant.

In all performed tests we have studied dimensions d in the range 5..10. Dimension of the discriminant space was in the range 1..2. The sample size N varied from 100 to 400. Number of classes q was in the range 2..5. Covariance matrices of all partial distribution densities were unit.

We studied accuracy of estimation of a posteriori probabilities, number of Bayesian classification errors (i.e., classification using estimated parameters vs. classification using theoretical parameters) and true classification errors (i.e., Bayesian classification using estimated or theoretical parameters vs. known true class numbers of the sample). Accuracy of estimation of a posteriori probabilities is measured as mean absolute distance $l(\hat{\pi}^N, \pi^N)$ between the estimated a posteriori probabilities $\hat{\pi}^N$ and the theoretical a posteriori probabilities π^N, i.e.,

$$l(\hat{\pi}^N, \pi^N) = \frac{1}{Nq} \sum_{i=1}^{q} \sum_{j=1}^{N} |\hat{\pi}(i, X_j) - \pi(i, X_j)| \,. \tag{5}$$

We compare distance $l(\hat{\pi}^N, \pi^N)$ and $l(\hat{\pi}_H^N, \pi^N)$ where $\hat{\pi}^N$ are obtained from MLE in the initial space and $\hat{\pi}_H^N$ are obtained from MLE in the discriminant subspace H. Number of Bayesian classification errors is measured as percentage of differences in Bayesian classification comparing classification using known theoretical parameter versus classification using estimated parameter. Recall, that Bayes rule of classification assigns an observation $X \in X^N$ to the ith class if $i = \arg\max_{k=1,2,\ldots,q} p_k \varphi_k(X)$.

Test 1 – dependence on dimension d

For the first test we selected Gaussian mixture model with five clusters with means $(-2r, -2r, 0, 0, \ldots)$, $(-r, -r, 0, 0, \ldots)$, $(0, 0, 0, 0, \ldots)$, $(r, r, 0, 0, \ldots)$, $(2r, 2r, 0, 0, \ldots)$. Calculations were done for three sample sizes: $N = 100, 200, 400$, varying distance between neighbour clusters, separately for $d = 5, 6, 7, 8, 9, 10$. Note that in this test r corresponds to distance $r\sqrt{2}$.

Here we present results for $N = 400$, comparing the case $d = 5$ with the case $d = 10$. In Figs. 1–3 on x axis we have parameter r. We observe weak dependence on dimension d, despite that number of parameters in case $d = 10$ is about 4 times bigger that in case $d = 5$. This is true also for all calculations in this test. This allows us to select

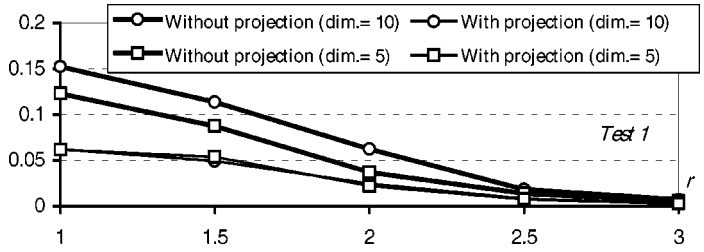

Fig. 1. Mean absolute error (average of 10 realizations).

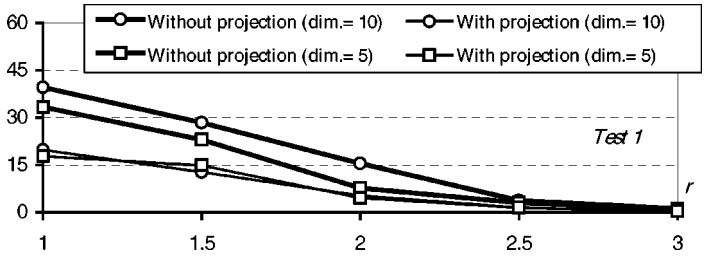

Fig. 2. Number of classification errors (average of 10 realizations).

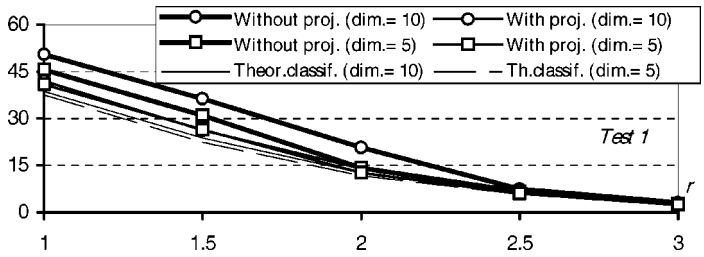

Fig. 3. True classification errors (average of 10 realizations).

one particular dimension for the next tests, thus saving much time for calculations. It is interesting to note that behaviour of number of classification errors (or true classification errors) is similar to that of mean absolute error.

Test 2 – dependence on sample size N

For this test we selected 5-dimensional Gaussian mixture model with five clusters with means $(-2r, 0, 0, 0, \ldots), (-r, 0, 0, 0, \ldots), (0, 0, 0, 0, \ldots), (r, 0, 0, 0, \ldots), (2r, 0, 0, 0, \ldots)$. Calculations were done for sample sizes: $N = 100, 110, 120, \ldots, 400$, varying distance between neighbour clusters.

Here we present results for $r = 3.0,\ 4.0,\ 5.0$ for mean absolute error (behaviour of number of classification errors is very similar). In Fig. 4 on x axis we have sample size N. We observe weak dependence on sample size N over all considered interval. This is

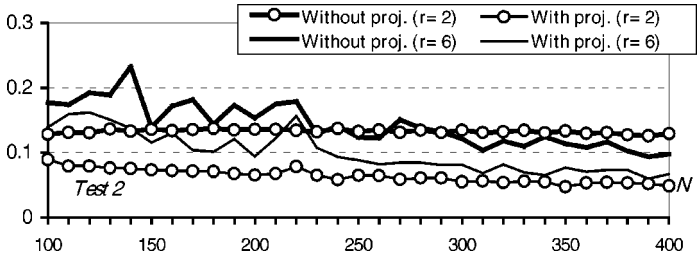

Fig. 4. Mean absolute error (average of 100 realizations, values for $r = 6$ are multiplied by 100).

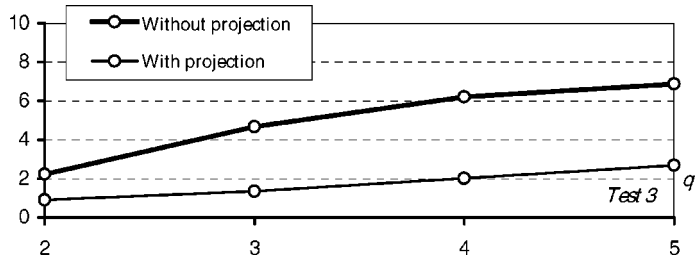

Fig. 5. Number of classification errors ($N = 300$, avg. of 100 realizations).

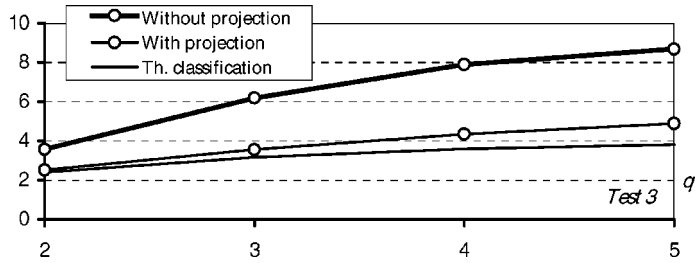

Fig. 6. True classification errors (average of 100 realizations).

true also for all calculations in this test. This also allows us to select one particular sample size for the next tests.

Test 3 – dependence on number of clusters q

For this test we selected 5-dimensional and 10-dimensional Gaussian mixture models with $q = 2, 3, 4, 5$ clusters with means placed in line with distance $r = 4.0$ in similar way as in previous test. Calculations were done for sample sizes: $N = 100, 110, 120, \ldots, 400$.

We present results for $d = 10$ and $N = 300$ for number of classification errors and true classification errors which show most advantage of using projection. In Figs. 5, 6 on x axis we have number of clusters q. Dependence on number of clusters is evident and for all values of q we observe advantage of using projection to discriminant subspace.

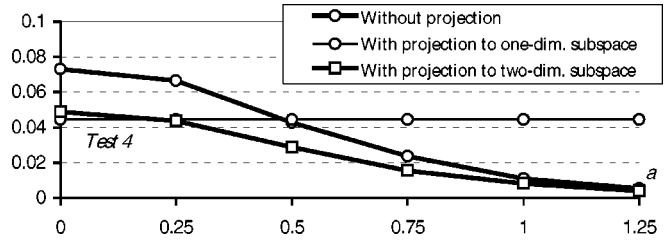

Fig. 7. Mean absolute error (average of 100 realizations).

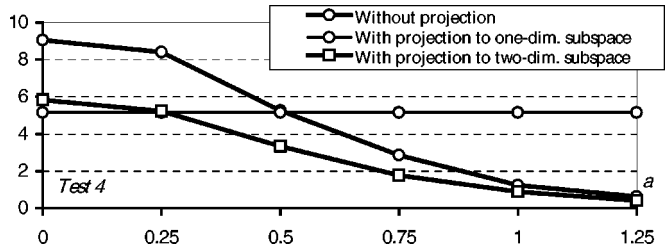

Fig. 8. Number of classification errors (average of 100 realizations).

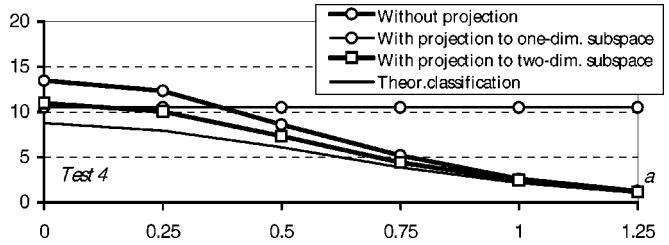

Fig. 9. True classification errors (average of 100 realizations).

Test 4 – projecting to discriminant subspace versus projection to lower dimension subspace

For this test we selected 5-dimensional Gaussian mixture model with three clusters with means $(-r, -a, 0, 0, \ldots)$, $(0, 2a, 0, 0, \ldots)$, $(r, a, 0, 0, \ldots)$, where $r = 3$ and a is a parameter. Calculations were done for sample sizes: $N = 100, 110, 120, \ldots, 400$.

We present results for $N = 400$. In Figs. 7–9 on x axis we have parameter a. Dependence on parameter a is evident and for all values of a we observe advantage of using projection to discriminant subspace. For bigger a starting from $a > 0.25$ projection to one-dimensional subspace gives worse results.

Conclusions

Performed tests show that the projection to the discriminant subspace is definitely recommended if possible. Much better estimates of a posteriori probabilities can be obtained when projection pursuit is used in the case $\dim H < d$. In some cases we get better estimates of a posteriori probabilities applying projection pursuit but intentionally projecting data to the lower dimension subspace. The advantages do not depend much on the sample size. Note that for Gaussian mixture models presented in the examples finding the discriminant subspace is quite simple. In general, finding the discriminant subspace in the higher dimensions is a very time consuming procedure and can significantly reduce advantages of using projection pursuit.

References

Aivazyan, S.A. (1996). Mixture approach to clustering via maximum likelihood, criteria of model complexity and projection pursuit. In *Data Science, Classification and Related Methods, Abstracts of 5th IFCS Conference*. IFCS, Cobe, **1**, 36.

Aivazyan, S.A., V.M. Buchstaber, I.S. Yenyukov and L.D. Meshalkin (1989). *Applied Statistics. Classification and Reduction of Dimensionality*. Finansy i Statistika, Moscow (in Russian).

Behboodian, J. (1970). On a mixture of normal distributions. *Biometrika*, **57**, 215–217.

Everitt, B.S., and D.J. Hand (1981). *Finite Mixture Distributions*. Chapman and Hall, London.

Friedman, J.H. (1987). Exploratory projection pursuit. *J. Amer. Statist. Assoc.*, **82**, 249–266.

Friedman, J.H., and J.W. Tukey (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. Comput.*, **C–21**, 881–889.

Hasselblad, V. (1966). Estimation of parameters for a mixture of normal distributions. *Technometrics*, **8**, 431–444.

Jakimauskas, G., and R. Krikštolaitis (1999). Appropriateness of projection pursuit in classification. *LMD mokslo darbai*. MII, Vilnius, **3**, 370–375.

McLacklan, G.J., and K.E. Basford (1988). *Mixture Models. Inference and Applications to Clustering*. Marcel Dekker, New York.

Rudzkis, R., and M. Radavičius (1995). Statistical estimation of a mixture of Gaussian distributions. *Acta Applicandae Mathematicae*, **38**, 37–54.

Rudzkis, R., and M. Radavičius (1997). Projection pursuit in Gaussian mixture models preserving information about cluster structure. *Lithuanian Math. J.*, **4**(37), 416–425.

Rudzkis, R., and M. Radavičius (1999). Characterization and statistical estimation of a discriminant space for Gaussian mixtures. *Acta Applicandae Mathematicae*, **58**, 279–290.

Schlesinger, M.I. (1965). On spontaneous discrimination of images. In *Reading Automata*. Naukova Dumka, Kiev, pp. 38–45 (in Russian).

Titterington, D.M., A.F.M. Smith and U.E. Makov (1985). *Statistical Analysis of Finite Mixture Distributions*. Wiley, New York.

Wu, C.F.J. (1983). On the convergence properties of the EM algorithm. *Ann. Statist.*, **11**, 95–103.

G. Jakimauskas was born in 1956. He graduated the Faculty of Mathematics in the Vilnius University in 1979. He is a researcher at the Institute of Mathematics and Informatics.

R. Krikštolaitis was born in 1972. He graduated the faculty of Fundamental Sciences in the Kaunas University of Technology in 1996. He is a Ph. D. student at the Vytautas Magnus University.

Tikslinio projektavimo įtaka klasifikavimo paklaidoms: kompiuterinio modeliavimo rezultatai

Gintautas JAKIMAUSKAS, Ričardas KRIKŠTOLAITIS

Nagrinėjama tikslinio projektavimo įtaka klasifikavimo paklaidoms ir aposteriorinių tikimybių įverčiams, gautiems iš imties. Laikoma, kad stebimas atsitiktinis dydis tenkina daugiamačio Gauso mišinio modelį. Pateikti kompiuterinio modeliavimo rezultatai rodo, kad, esant palyginti nedideliam imties tūriui, klasifikavimas, panaudojant tikslinio projektavimo algoritmą, duoda didesnį aposteriorinių tikimybių įverčių tikslumą ir mažesnę klasifikavimo paklaidą.