Probability Density Estimation Classifier Based on Sample Distances

Jiřina, Marcel
2003

cAk

CENTRUM APLIKOVANÉ KYBERNETIKY

České vysoké učení technické v Praze - fakulta elektrotechnická

Probability Density Estimation Classifier Based on Sample Distances

Technical report

Marcel Jiřina

www@c-a-k.cz

2003

Institute of Computer Science
Academy of Sciences of the Czech Republic

Probability Density Estimation Classifier Based on Sample Distances

Marcel Jiřina

Technical Report No. V-902

January 2003

Abstract

The method proposed is very close to popular and often efficienct methods of the nearest neighbors. Standard methods of probability density estimate for classification which are based on the nearest neighbors approach solve the problem of classification by an estimate of the probability density in the point x of the data space by ratio i/V. i is the number points of a given class of the training set in a suitable ball of volume V with center at the point x. The new method is based on distances of all points of a given class of the training set from a given (unknown) point x. It is shown that the sum of reciprocals of (n-1)-st power of these distances is convergent and can be used as the probability density estimate. The speed of convergence is the better the higher dimensionality. The classification quality was tested and compared with other methods.

Keywords:
Bayes ratio estimation, multivariate data, classification, curse of dimensionality, classification speed, nearest neighbor

Introduction

The methods of classification based on the nearest neighbors estimate the probability density in the point x of the data space by ratio i/V_i . i is anumber of points of a given class in a suitable ball of volume V_i with center in the point x [1]. These methods need to optimize the best size of the neighborhood, i.e. the number of points i in the neighborhood of the point x or size of volume V_i. The probability density in the feature (data) space is given by training data. Optimal neighborhood size depends on training data set, i.e. on character of data as well as on the number of samples of a given class in the training set.
The method proposed is based on distances of the training set samples x_i, i = 1, 2, … k from the point x. We will show its fast convergence, i.e. small influence of distant samples in multidimensional Euclidean space.
Using distances, i.e. a simple transformation from E_n to E_1, and no iterations the curse of dimensionality is straightforwardly eliminated. The method can be also considered as a variant of kernel method, based on a probability density estimator, but using a much simpler metric.
Throughout this paper let us assume that we deal with standardized data, i.e. the individual coordinates of the samples of the learning set are standardized to zero mean and unit variance and the same standardization constants (empirical mean and empirical variance) are applied to all other (testing and of unknown class) data.

All learning samples approach

Let be given the learning set of total m_T samples in form of a matrix X_T with m_T rows and n columns. Each sample x_i = (x_{i1}, x_{i2}, … x_{in}) $\notin X_T$, i = 1, 2, ... m_T corresponds to one row of X_T and, at the same time, corresponds to a point in n-dimensional Euclidian space E_n. The learning set consists of points (rows) of two classes $c \in \{0, 1\}$, i.e. each row (point or sample) corresponds to one class. We use standardized data, i.e. each variable x_{ij} (j fixed, i = 1, 2, ... m_T , corresponds to j-th column of the matrix X_T) has zero mean and unit variance.
Let there be a point $x \in E_n$ different from samples (rows) of the learning set X_T. In the learning set there exist points x_{Ti}, i = 1, 2, ... k, $k \leq m_T$, $x_{Ti} \in X_T$ of class c nearest to the point x_i . x_{T1} is the nearest point to x, x_{T2} is the second nearest point to x, etc. The Euclidean distance of these points from the point x let be $d_i = d(x, x_{Ti})$. There is a ball with center at the point x and radius sufficiently large to contain just i points nearest to the point x. The volume of the ball is V_i = const. $d_i^{\,n}$ in E_n. For each ball with index i and having just i points inside it, the probability density estimate can be given by formula (C is a constant)

$$p(x,i) = C\frac{i}{V_i} \cdot$$

For probability density estimation in the point x we take average values of i/V_i for several i´s. Let us use i = 2, 3, ...k , excluding, in fact, the influence of the nearest neighbor because its influence is most unreliable. Having in mind no equidistant (no equivolumous) sizes of individual balls of volumes V_i, it seems more appropriate to use the true distance d_i of the point i from the point x instead of some "weight" expressed by numerator i in each fraction i/V_i. Thus if C‘ is a constant independent of class the probability estimate that x belongs to the class c is

$$p_c(x) = \frac{C'}{k-1}\sum_{i=2}^{k}\frac{d_i}{V_i} = \frac{C'}{k-1}\sum_{i=2}^{k} 1/\,d_i^{\,n-1} \cdot \qquad (1)$$

Under the assumption that the series $1/d_i^{\,n-1}$ converges with size of d_i for $n > 1$ we have no reason to limit ourselves to nearest k points and we can use all points in the learning set using $k = m_T$. At the same time the ordering of individual components is not essential and we need not sort the samples of X_T with respect to their d_i as when using nearest neighbor approach.
In practical procedure we simply sum up all components $1/d_i^{\,n-1}$ and at the same time we store the largest component which corresponds to the nearest neighbor of the point x which has the smallest $d_i^{\,n-1}$. In the end we subtract it thus excluding the nearest point. This is made for both classes simultaneously getting numbers A_0 and A_1 for both classes. Their ratio gives value of discriminant function, here the Bayes ratio or the probability estimation that the point x$\in E_n$ is of the class 1

$$R(x) = \frac{A_1}{A_0} \quad \text{or} \quad p_1(x) = \frac{A_1}{A_1 + A_0} \cdot$$

Then for a threshold (cut) θ chosen, if $R(x) > \theta$ or $p_1(x) > \theta$ then x belongs to class 1 else to class 0.

Using distances, i.e. a simple transformation $E_n \rightarrow E_1$ and no iterations the curse of dimensionality is straightforwardly eliminated. The method needs no tuning parameters: No neighborhood size, no convergence coefficients etc. need to be set up in advance to assure convergence. The speed is high. In the learning phase only standardization constants are computed. In the recall phase for each sample to be classified the learning set is searched once and for each sample of the learning set one element of sum (1) is computed. The amount of computation is thus proportional to learning set size, i.e. the dimensionality times the number of learning samples.
The method is very close to the nearest neighbor as well as kernel methods. The procedure described in the text above Eq. (1) is nothing else than the nearest neighbor method. Simply an average of several neighborhoods is taken, but the number of points inside ball is changed to distances. From the point of view of kernel methods, the kernel is or would be $K(x) = \| x - x_i \|^{-(n-1)}$ with Euclidean norm $\|.\|$ in E_n. There is no smoothing (bandwidth) parameter. The problem is that this kernel is difficult to consider as a probability function according to the definition of a kernel [1]. Taking $\|x\text{-}x_i\| = r$ we have $K(r) = r^{-(n-1)}$ and integrals $\int_{-\infty}^{\infty} K(r)dr$ or $\int_{0}^{\infty} K(r)dr$ are not convergent; they should be equal to 1 or at least finite.

Probability Density Estimation

Let us look at the problem what is the relation of the part D_k of the space E_n which falls on k nearest neighbors of the given point x. We will assume the following:
Assumption 1
Let there be points in the Euclidian space E_n distributed randomly and homogenously in the sense that the distribution of each of n coordinates is uniform. Let k be the order number of the k-th nearest neighbor to the point x. Let r_k be the distance of the k-th nearest neighbor of the given point $x \in E_n$ from the point x_k. Let D be a constant, and $\overline{D}_k$ be the mean value of the variable r_k^{n-1}, and let it holds

$$\overline{D}_k = kD \ .$$

Comment
„The part D_k of the space E_n“ is not a volume of a ball with the center in the point x and radius r_k but, in fact (except for a multiplicative constant), the ball of the same center and radius but in the space of dimension by one lower, i.e. in the E_{n-1}. By simulation one can find that the relation $\overline{V}_k = kV$ where V is a constant does not hold but it holds $\overline{D}_k = \overline{r}_k^{n-1} = kD$ where k is the number of the k-th nearest neighbor of the point $x \in E_n$ and D is a constant. It can be found that the mean value of the n-th power of r_k grows faster than linearly and the (n-2)-nd power grows slower than linearly. It is demonstrated in Fig. 1.

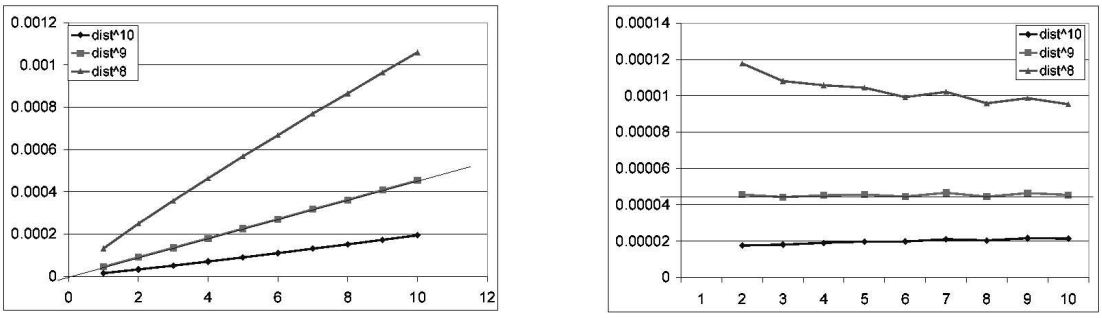

Fig. 1. Dependence of three different powers of the distance of ten nearest neighbors on the order of the nearest neighbor (left) and corresponding differences (right) in E_{10}. Each point of the left hand graph is the average of 3390 entries. The straight line shows true linear dependence.

One can look at the problem also differently. There is a space of a series of nearest neighbors for some arbitrary point x. The space where the nearest neighbors lie is the E_n but their placement is limited by the position of the point x and their distances from this point. The mean value of the k-th nearest neighbor distance from the point x

is fixed and depends on the point *x* and the probability density of the presence of the points of the given class in corresponding neighborhood of the point *x*. If the point *x* and the mean distance r_k^{n-1} are given then the position of the *k*-th nearest neighbor has *n*-1 degrees of freedom. It is, really, a point of a ball surface of radius r_k^{n-1} and with center in the point *x*. Then the space of nearest neighbors has the dimension *n*-1, not *n*.
The Assumption 1 is supported by the following lemma:
Lemma [6]
The sum of *k* independent exponentially distributed random variables with parameter λ is equal to an Erlang (gamma) distributed random variable with parameters λ and *k*, i.e. let $Y_i \sim \exp(\lambda)$ then $X \sim \mathrm{Erl}(\lambda, k)$ where $X = \sum_{i=1}^{k} Y_k$. □

Theorem 1
Let $\overline{D}_i$ be mean value of $D_i = r_i^{n-1} - r_{i-1}^{n-1}$, $\overline{D}_k$ be mean value of $D_k = r_k^{n-1}$, $\overline{V}_k$ be mean value of $V_k = cr_k^n$ where *c* is a constant, and the Assumption 1 be valid. Moreover let exist a constant *C* such that $p(\overline{D}_i) = \frac{C}{\overline{D}_i}$.
Then for the probability density $p(k) = C'k/\overline{V}_k$ of points in the neighborhood of point *x* it holds $p(\overline{D}_i) = p(\overline{D}_k) = p(k)$, where $p(\overline{D}_k) = \frac{kC}{\overline{D}_k}$.

Proof
Under the assumption $p(k)$ is probability density and at the same time due to Assumption 1 $1/\overline{D}_k$ is proportional to $p(k)$. Then there is a constant *C* that $p(D_k) = p(k)$. Under the Assumption 1 there is $\overline{D}_k = k\overline{D}_i$ and then $p(\overline{D}_k) = p(\overline{D}_i)$. □

The Proof of Convergence

Let be given n dimensional data, each sample in form of a row vector $x = (x_1, x_2, \ldots x_n) \in R_n$. All these data form the feature space. These data come from two sources, then these data are of two classes. The class $c = 1$ is usually denoted as the signal (*s*) and class = 0 (sometimes -1) is usually denoted the background (*b*). The part of data where the relation of each sample to the class is known and is used as a basis for probability density estimation is called the learning set. The other data where the relation of each sample to the class is known can be used as the testing set for evaluation of behavior of the classifier. These notions are commonly used; sometimes the learning set is called the training set.
Notation
Let the learning set $U=U_1 \cup U_2$, $U_1 \cap U_2 = \varnothing$, $U_c = \{x_{ci}\}$, $i=1, 2, \ldots N_c$, $c=\{0,1\}$ be given. N_c is the number of samples of the class *c*, $x_{ci}=\{x_{ci1}, x_{ci2}, \ldots x_{cin}\}$ is the data sample, where *n* is the sample space dimension. Let point $x = \{x_1, x_2, \ldots x_n\} \notin U$ be given and let points x_{ci} of each class U_1, U_2 be sorted so, that index $i = 1$ corresponds to the nearest neighbor, the index $i = 2$ to the second nearest neighbor, etc. In the Euclidian metrics, $r_i = \|x, x_{ci}\|$ is the distance of the *i*-th nearest neighbor of the class *c* from the point *x*.
Theorem 2
Let exist a mapping of probability density distribution of points of the class *c* in E_n, $E_n \to E_1$: $p(x_{ci}) = p(r_{ci}^{n-1})$ so that

$$K/r_{c1}^{n-1} = p(x_{c1}),\ K/(r_{c2}^{n-1} - r_{c1}^{n-1}) = p(x_{c2}),\ \cdots K/(r_{cNc}^{n-1} - r_{c(Nc-1)}^{n-1}) = p(x_{cNc}), \qquad (2)$$

where *K* is a fixed constant that has the same value for both classes.
Let exist a constant $\varepsilon > 0$ and index $k > 2$ so that for each $j > k$ it holds

$$p(x_{cj}) \le \frac{p(x_{c2})}{(1+(j-k)\varepsilon)^{j-k}}\ . \qquad (3)$$

Then

$$S = \sum_{j=2}^{N_c} \frac{1}{r_{cj}^{n-1}} = p(x_{c2}) K (1 + C_c), \tag{4}$$

where K and C_c are finite constants.

Proof.
First we arrange (4) in form

$$S = \sum_{j=2}^{N_c} \frac{1}{r_{cj}^{n-1}} = \frac{1}{r_{c2}^{n-1}} + \sum_{j=3}^{N_c} \frac{1}{r_{c2}^{n-1} + \Delta_{c3} + \Delta_{c4} + \mathrm{K} + \Delta_{cNc}} \quad .$$

Then using the mapping (2) introduced we get

$$S = K p_{c2} + K \sum_{j=3}^{Nc} \frac{1}{\frac{1}{p_{c2}} + \frac{1}{p_{c3}} + \mathrm{K} + \frac{1}{p_{cNc}}} = p_{c2} K (1 + \sum_{j=3}^{Nc} \frac{1}{1 + \frac{p_{c2}}{p_{c3}} + \mathrm{K} + \frac{p_{c2}}{p_{cj}}}) \equiv p_{c2} K (1 + \sum_{j=3}^{Nc} P_j) \tag{5}$$

For individual elements p_{c2} / p_{cj} in denominators of fractions in the sum it holds

$$\frac{p_{c2}}{p_{cj}} = \frac{p_{c2} (1 + (j-k)\varepsilon)^{j-k}}{p_{c2}} = (1 + (j-k)\varepsilon)^{j-k} .$$

Using the condition (3) the summed elements P_k, P_{k+1}, … in (5) since the k-th have form

$$P_k = \frac{1}{C}, \quad P_{k+1} = \frac{1}{C + 1 + \varepsilon}, \quad P_{k+2} = \frac{1}{C + 1 + \varepsilon + (1+\varepsilon)^2}, \quad \mathrm{K} \ ,$$

$$P_{k+i} = 1 / [C + (1+\varepsilon) + (1+2\varepsilon)^2 + \ldots + (1+i\varepsilon)^i] \ .$$

Then according to d'Alembert's criterion

$$\frac{P_{k+i+1}}{P_{k+i}} = \frac{C + (1+\varepsilon) + (1+2\varepsilon)^2 + \ldots + (1+i\varepsilon)^i}{C + (1+\varepsilon) + (1+2\varepsilon)^2 + \ldots + (1+i\varepsilon)^i + (1+(i+1)\varepsilon)^{i+1}}$$

and after a little algebra

$$\frac{P_{k+i+1}}{P_{k+i}} \le \frac{C/(1+i\varepsilon)^i + i}{C/(1+i\varepsilon)^i + i + (1+(i+1)\varepsilon)^{i+1}/(1+i\varepsilon)^i} < \frac{C/(1+i\varepsilon)^i + i}{C/(1+i\varepsilon)^i + i + (1+i\varepsilon)} < 1$$

$\forall i > 0$ and $\forall \varepsilon > 0$. Then the series is convergent.

Notes
a) In the statement of the theorem the sum need not start just by index $j = 2$. One can start with the nearest neighbor ($j = 1$) or other neighbor ($j > 2$). The value $j = 2$ is given by compromise between the error caused by small value and large variability of $\Delta_{c1} = r_{c1}$ and inaccuracy caused by larger distance from the point x for $j > 2$.
b) The last condition (3) defines the speed of diminishing of the tail of the distribution; probably a condition that the distribution should have the mean would suffice.

Discussions

From the formula (5) it is seen that for „smooth" form of distribution function around the point x and for large density of points for both classes the ratios p_{c2} / p_{cj} are very close to 1 for rather large values of j (e.g. 100, but let us take 11 here). For both classes are the elements of sum in (5) $\frac{1}{2}, \frac{1}{3}, \mathrm{K} \frac{1}{11}$ and their sum is 2.01987 here and the other elements have form $\frac{1}{11 + (j-11)(1+\delta)}$, where since the index k it is $\delta \ge \varepsilon$. (The index k can be different for both classes.) It is then probable that values of sums in (5) will be very close for both classes and ratio of (5) for one and the other class will be close to Bayes ratio $p_1(x_{c2}) / p_0(x_{c2}) = S_1 / S_0$. In such a case one can also estimate the probability that the sample x belongs among signals:

$$p_1(x) \approx p_1(x_{c2}) \approx \frac{S_1}{S_1 + S_0} \ .$$

Blessed Dimensionality - the Speed of Convergence Estimation

Remind that the samples of the learning set are standardized to zero mean and unit variance for each variable. Assume that all thus arising marginal distributions are approximately normal. Assume also that our point x has an unknown class or unknown probabilities $p_1(x)$ and $p_0(x)$ and lies not too far from the point (0, 0, … 0). For the point x one can introduce different neighborhoods, now let us use three only:

- Till the distance of one sigma,
- From the distance of one sigma to the distance of two sigma,
- Since the distance of two sigmas further all in each dimension.

Due to the standardization of all variables in each dimension approximately 68 % points of the learning set lie inside A, 95 % points lie inside A and B, i.e. 27 % in B, and 5 % in C. The results of some computations for dimensionality n = 2 to 50 shows the Table 1.

layer→	A	B	C
	≤ 1 sigma	Between 1 and 2 sigma	>2 sigma
	Average distance in one dimension		
	0,5	1,5	3
n		**Total points inside layer**	
2	46,24%	44,01%	9,75%
3	31,44%	54,29%	14,26%
4	21,38%	60,07%	18,55%
5	14,54%	62,84%	22,62%
7	6,72%	63,11%	30,17%
10	2,11%	57,76%	40,13%
20	0,044687%	35,80%	64,15%
30	0,000945%	21,46%	78,54%
50	4,22129E-09	7,69%	92,31%
n		Benefits to the total sum	
2	73,94%	23,46%	2,60%
3	83,02%	15,93%	1,05%
4	90,25%	9,39%	0,36%
5	94,83%	5,06%	0,11%
7	98,72%	1,27%	0,0095%
10	99,86%	0,14%	0,00019%
20	99,999931%	0,000069%	2,35588E-12
30	99,999999967%	0,000000033%	2,2564E-18
50	99,9999999999%	7,61716E-17	1,62319E-30

Table 1. Total number of points of the learning set inside layers A, B, C and their benefits to the total sum.

The benefit to the total sum was estimated from average distance in each dimension in corresponding layer (A, B or C). These estimations show that due to the geometry of multidimensional Euclidian space the share of points corresponding to A with respect to total number of points lessens essentially with dimension. At the same time, their benefit to the total sum is closer to 100 %. This is because the parts A, B, C are, in fact, not cubes but n-dimensional balls of radii computed from an average distance in one dimension as stated in the Table 1. From it also follows that the share of the part C to the total sum is negligible since the dimension 6. With growing dimension also the convergence of the sum is much faster as the points of the learning set near to point x gave practically whole value of the sum. The larger dimension, the lesser percentage of points from the learning set influences the result. On the other hand for low dimensionality, especially 2 and 3 even the farthest points influence the result.

Testing the Convergence on Examples

The course of convergence for dimensions 2, 3, and 10 show Figs. 2 till 9. For Figs. 2 till 7 artificial tasks were used. In these tasks the signal has distribution in form of diamond and the background in the form of top hat in

all dimensions. We used 250 samples in each dimension in each class. Small number of samples causes not too smooth curves but on the other hand demonstrates usefulness of the method for small learning set. Figs. 8 and 9 show results with practical data [2] and it is seen that this task converges faster than the artificial task of the same dimension.

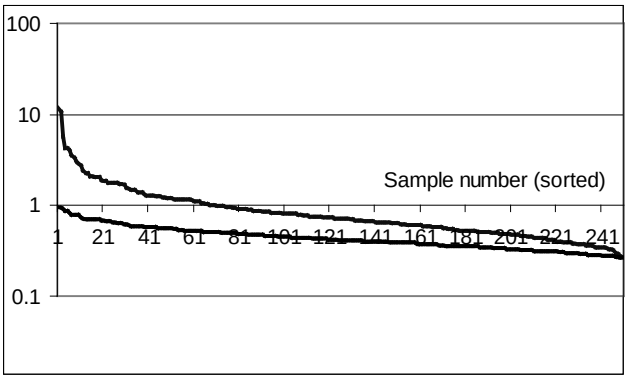

Fig.2. Sample contribution to the total sum for signal and background sorted according to size, two-dimensional artificial data.

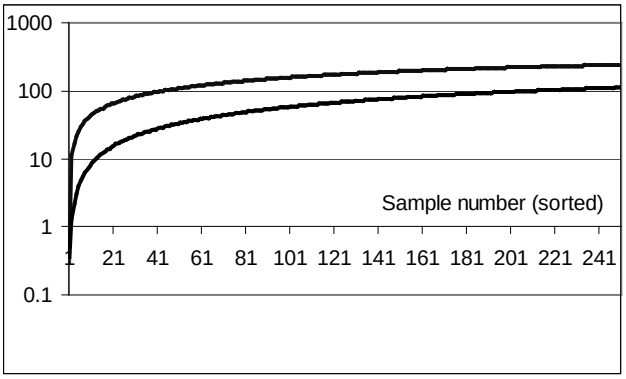

Fig. 3. Fig.2. Size of the total sum for signal and background sorted according to size of sample contribution, two-dimensional artificial data

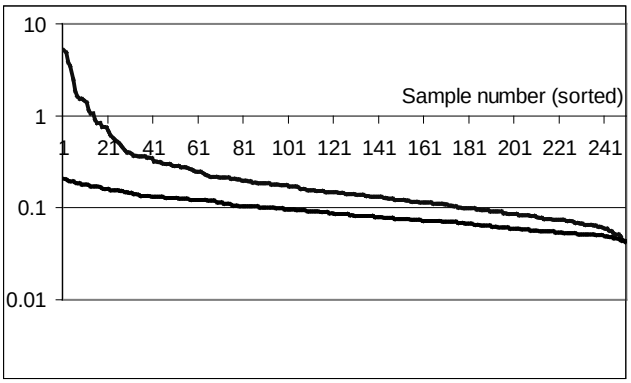

Fig. 4. Sample contribution to the total sum for signal and background sorted according to size, three-dimensional artificial data.

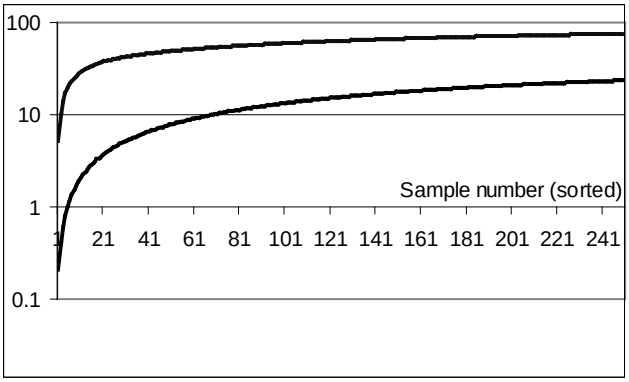

Fig. 5 Size of the total sum for signal and background sorted according to size of sample contribution, three-dimensional artificial data.

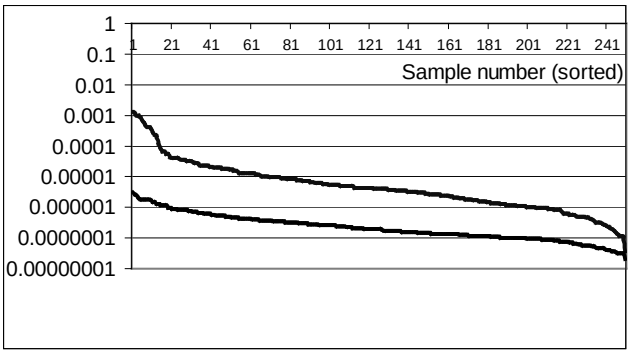

Fig. 6. Sample contribution to the total sum for signal and background sorted according to size, ten-dimensional artificial data.

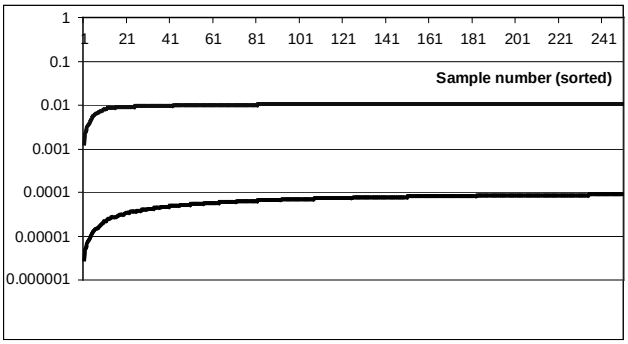

Fig. 7. Size of the total sum for signal and background sorted according to size of sample contribution, ten-dimensional artificial data

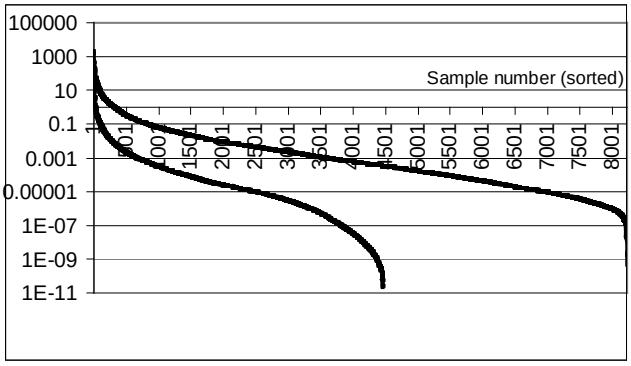

Fig. 8. Sample contribution to the total sum for signal and background sorted according to size, ten-dimensional practical data [2]. There are different numbers of

signal samples and background samples in the learning set.

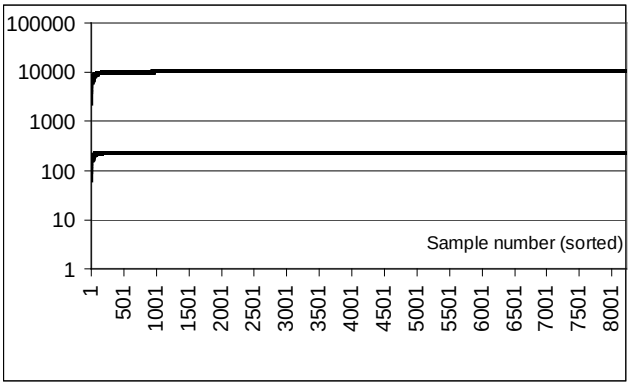

Fig. 9. Size of the total sum for signal and background sorted according to size of sample contribution, ten-dimensional practical data [2].

Classification Ability

The metod was tested on the same data as was used in study [2]. Also, the third and next lines of the Table 2 are cited from this source and then we do not describe the different methods in detail.

TABLE 2

Method	loacc	hiacc	σ{0.5}	σ{max}	SigEff for σ{max}
New method	0.452	0.778	8.40	9.35	0.364
C5.0	0.441	0.830	8.14	8.74	0.408
CART	0.414	0.810	7.94	8.03	0.538
NearestNeighb	0.443	0.816	8.03	9.12	0.317
Kernel	0.443	0.803	8.43	8.64	0.390
NNSU	0.472	0.731	9.74	9.82	0.483
NeuNet	0.445	0.839	8.73	8.75	0.483
MLP	0.300	0.767	6.93	7.22	0.576
GMDH	0.280	0.736	6.55	6.77	0.574

The table gives the quality numbers loacc, hiacc, and significance σ with the following meaning: loacc is the average signal efficiency obtained by interpolating values of signal efficiency SigEff at the points 0.01, 0.02, and 0.05 for background error BckErr; hiacc is obtained in a similar way by averaging signal efficiency at the points 0.1 and 0.2 background error; significance σ is defined by $\sigma = S/\sqrt{2B+S}$, where S = SigEff.N_s and B = BckErr.N_b ; N_s and N_b are the number of signal and background events that would be obtained by selecting events in samples with N_b = 10 000 and N_s = 500; we give the value of σ obtained at SigEff = 0.5, and the maximum value along with the value of SigEff where it is found.
The results are also compared in Fig. 10.

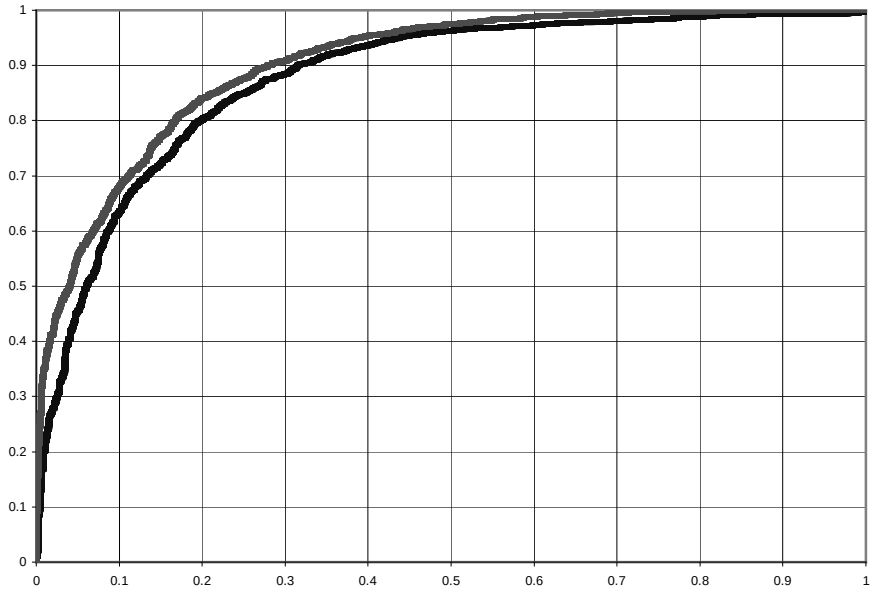

Fig. 10. Dependence of Signal Efficiency vs. Background Error for new method (upper line) and for quadratic GMDH MIA method (bottom line).

Conclusions

The method presented is based on two ideas, a simple transformation $E_n \rightarrow E_1$ and differences of volumes of multidimensional cube and multidimensional ball in Euclidean space.
Using distances, i.e. a simple transformation $E_n \rightarrow E_1$ and no iterations the curse of dimensionality is straightforwardly eliminated.
The theorem on convergence was formulated and proved and convergence estimation was given. It was shown that the higher dimensionality, the better.
The method needs no tuning parameters: No neighborhood size, no convergence coefficients etc. need to be set up in advance to assure convergence. The other advantage is the speed. In the learning phase only standardization constants are computed. In the recall phase for each sample to be classified the learning set is searched once and for each sample of the learning set one element of sum (1) is computed. The amount of computation is thus proportional to learning set size, i.e. the dimensionality times the number of learning samples. With approximately the same quality, the method gives results in orders of magnitude shorter time than much sophisticated approaches [2], [5].

Acknowledgement
This work was supported by the Ministry of Education of the Czech Republic under project No. LN00B096.

References

[1] Silverman, B. W.: Density Estimation for Statistics and data Analysis. Chapman and Hall, London, 1986.
[2] Bock, R. K. et al.: Methods for multidimensional event classification: a case study. To be published as Internal Note in CERN, 2003.
[3] Jiřina, M.: All Training Samples Density Estimation Classifier. Technical Report No. 881, ICS AS CR Prague, November 2002.
[4] Jiřina, M.: Nearest Neighbor Distance Statistics Estimation. Technical report No. 878, ICS AS CR Prague, November 2002, pp.12.
[5] Hakl F., Hlaváček M., Kalous R. 2002 Application of Neural Networks Optimized by Genetic Algorithms to Higgs Bosson Search. In: The 6th World Multi-Conference on Systemics, Cybernetics and Informatics. Proceedings. (Ed.: Callaos N., Margenstern M., Sanchez B.) Vol. : 11. Computer

Science II. - ISSS, Orlando 2002, pp. 55-59 (ISBN: 980-07-8150-1) Held: ISAS SCI 2002 /6./, Orlando, US, 02.07.14-02.07.18
[6] Kleinrock, L.: Queuing Systems. Vol. I - Theory. (Chap. 4.2) John Wiley and Sons, New York, 1975.
[7] Farlow, S.J.: Self-Organizing Methods in Modelling. GMDH Type Algorithms. Marcel Dekker, inc., Mew York, 1984.