All Training Samples Density Estimation Classifier

Jiřina, Marcel
2002

cAk

CENTRUM APLIKOVANÉ KYBERNETIKY

České vysoké učení technické v Praze - fakulta elektrotechnická

All Training Samples Density Estimation Classifier

Technical report

Marcel Jiřina

www@c-a-k.cz

2002

Institute of Computer Science
Academy of Sciences of the Czech Republic

All Training Samples Density Estimation Classifier

Marcel Jiřina

Technical Report No. 881

November 2002

Abstract

The method proposed is based on neighbours distances from a given (unknown) point, i.e. on a simple transformation $E_n \rightarrow E_1$. With this transformation the curse of dimensionality is straitforwardly eliminated in the first step. It is shown that the sum of reciprocals of (n-1)st power of these distances is convergent and can be used for Bayes ratio estimation. The classification quality was tested and compared with other methods using simulated multivariate data from gammma telescope. Essential advantage of the approach is the fact, that no tuning parameters exist. The amount of computation is proportional to the training set size, i.e. the dimensionality times the number of training samples.

Keywords:
Bayes ratio estimation, curse of dimensionality, multivariate data, classification quality

Introduction

Multivariate data classification is rather difficult task especially due to two main reasons. The first is often presence of large noise which makes the classes worse distinguishable. The second is so called curse of dimensionality which causes that the volume of computations grow fast, exponentially, with task dimension. One would understand that volume of computation would grow with amount of data (dimensionality times number of samples) quadratically or even with third power but exponential grow is too fast.
There is huge amout of literature dealing with this problem using several approaches:

- Bayes principle and statistical distribution density estimation [1], [2], [3] including fast but not too reliable naive approaches [6]
- neural networks of different kinds from single layer back propagation networks, radial basis function neural nets, neural nets with switching units etc. [4], [5], [7] sometimes optimalized by genetic algorithm [8]
- methods based on classification trees or random forests [9]
- and others [10].

All these approaches can be classified in two other classes, the methods suitable for tasks with small training data set, and methods which use or rely on large amount of data in the training set, i.e. large density of points in the corresponding space. The classifiers of the first kind rely on good approximation of probability density function over all region of interest, whereas the classifiers of the other kind can approximate the density function around each point of interest, the point to be classified, by a constant. In these methods it is supposed that with respect to volume density of points of the training set in the space the probability density function changes only slightly.
In this contribution a classifier of a second kind, i.e assuming large training set, is designed where the amount of computation grows linearly with the volume of the training data, i.e. proportionally to the data dimension and also proportionally to the number of samples in the training set. The classifier is based on rather naive approach reminding nearest neighbour or N-th nearest neighbour method and effectively uses suitable transformation $R_n \rightarrow R_1$, in fact $E_n \rightarrow E_1$, dependent on the dimension *n*.

The Task

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 called the signal (s) and by 0 (sometimes -1) we denote the background (b). The task is, given the vector *x* decide of what class it is, zero or one.
The optimal way to partition the feature space into signal and background regions is to choose the Bayes discriminat function. This function is simply ratio of the probability *P(s|x)* that a given sample is a signal event and the probability *P(b|x)* that it is a background sample. It is written in form [1], [2]

$$R(x) = \frac{P(s \mid x)}{P(b \mid x)} = \frac{P(x \mid s)P(s)}{P(x \mid b)P(b)}.$$

Each cut on the value of discriminant function correponds to a discriminating boundary in the feature space. The quantities $P(x|s)$ and $P(x|b)$ are likehood functions for signal and background, respectively. $P(s)$ and $P(b)$ are corresponding prior probabilities.
We have no other information than the training set of samples for which the class c of each individual sample is known. Thus we have the training set X_T of m_T n+1 dimensional samples $x_T = (x_1, x_2,... x_n, c) \in R_n \times \{0, 1\}$. From it we have to derive all the information we need. As we know nothing about the task set or testing set X of sample vectors we cannot use any concrete likehood functions for signal and background with exception that we consider both class equally possible and $P(x|s) = P(x|b)$. It corresponds to state of minimal information. On the other hand if the training set is large, we have rather good information about prior probabilities $P(s)$ and $P(b)$.

Probability density estimation

In our case the vectors x are points in R_n space, but let us consider it as an E_n space. Let us consider vectors - points of one class c only. The number n_v of these points in some small volume v in R_n recomputed to unit volume gives a "concentration" of samples of given class in that place of R_n space. If V is the volume of convex closure of all data points in the training set, m_T is the number of all training samples, then the estimation of the probability that the sample of class c is in v is given by the ratio

$$p_c(v) = \frac{n_v}{v} \Big/ \frac{m_T}{V}$$

In limit for $m_T \to \infty$ and $v \to 0$ (v shrinks to some fixed point x in P_n) it is just probability density $p_c(x)$ in point x of the multivariate distribution of samples of given class c. In fact, $p_c(x)$ for $c = 1$ and $c = 0$ are just prior probabilities $P(s)$ and $P(b)$, respectively, in the point x in R_n. For $x \in X_T$ we simply associate this sample to the class of the same sample from the X_T . The task is now, given $x \notin X_T$ find the prior probabilities $P(s)$ and $P(b)$.
Throughout this contribution let us assume that we deal with standardized data, i.e. the individual coordinates, columns of X_T, are standardized to zero mean and unit variance and the same standardizating constants (empirical mean and empirical variance) are applied to all other (testing) data.

All training samples approach

Let be given point $x \notin X_T$ and some points x_{Ti}, $i = 1,2...k$, $x_{Ti} \in X_T$ of class c nearest to the point x. The Eucleidian distance of these points let be $d_i = d(x, x_{Ti})$ and the largest of them be r. There is a ball of volume $V = \text{const}.r^n$ in R_n. We can conclude that probability density estimation in the point x can be proportional to k/V_k as in the k-th nearest neighbour method [6]. The greatest advantage of using either distance d_i or volume V_i is simple mapping $R_n \to R_1$ and thus no problem with curse of dimensionality arises. The particular value of k/V_k depends on k. For $k = 1$ the estimation will be very poor and dependent on random position of x to the nearest x_{Ti}. The larger k the better but on the other hand the ball is rather large and the density inside is approximated as it woud be homogenous and does not reflect more detailed structure of probability density function and more detailed structure of n-dimensional (n>1) Eucleidian space even if the probability density function is homogenous [11].
A better approach can use 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 x_{Ti} as its influence is most problematic. Then

$$\bar{p}_k = C\sum_{i=2}^{k} i/V_i \text{ , simply } \bar{p}_k = C(\frac{2}{V_2} + \frac{3}{V_3} + \cdots + \frac{k}{V_k}), \quad (1)$$

where C is proportionality constant. This can be used directly for probability density estimation but now we will continue in another way using rather ad hoc jump.
Having in mind nonequidistant (nonequivolumous) sizes of individual balls of volumes V_i, it seems more appropriate to use the true distance of the point No. *i* instead of some "weight" or "distance" expressed by numerator *i* in each fraction of (1). Thus

$$\bar{p}_k = C(\frac{d_2}{V_2} + \frac{d_3}{V_3} + \cdots + \frac{d_k}{V_k}) = C\sum_{i=2}^{k} \frac{d_i}{V_i} = C'\sum_{i=2}^{k} 1/d_i^{n-1} . \quad (2)$$

C and *C'* are constants. Under 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 training 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 (1).
The method is based on the

Theorem

Let for any $x \in E_n$, some *r*, $0 < r \leq n$ and any *k* exists $\varepsilon > 0$ such that $d_{k+1}^r / d_k^r > 1+\varepsilon$. Then $\sum_{k=1}^{\infty} \frac{1}{d_k^r}$ is convergent.

Proof is based on elementary use of D'Alembert convergency criterion.

Application

According to the theorem and with respect to the fact that for $r = n-1$ the ratio $d_{k+1}^{n-1} / d_k^{n-1} > 1+\varepsilon$ [11], the convergency of (2) is guaranteed. It can be found that only several nearest neighbours would suffice. On the other hand one must in any case search whole training set to find all nearest neighbours necessary.
Very essential advantage of the approach is the fact, that no tuning parameters exist. No neighbourhood size, no convergency coefficients etc. need to be set up in advance to assure convergency. In practical procedure we simply sum up all components $1/d_i^{n-1}$ and at the same time we store the largest one corresponding to the nearest point and 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 [1], [2] for particular point $x \in E_n$

$$R(x) = \frac{A_1}{A_0} .$$

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

Evaluation criteria

Remind that each cut on the value of discriminant function correponds to a discriminating boundary in the feature space. In practice it means that, depending on the cut value, some samples are denoted as signals, the others as background. In both of this classes there are usually well recognized samples (signal sample as signal, background sample as background). At the same time there are two sets of wrongly recognized samples.

Suppose that we have the testing set of samples of known classs so that this class can be compared with recognized class (associated by the classification algorithm). Now we can define:
Signal efficiency, SigEff, as a ratio of number of signal events after the classifier, i.e. number of properly recognized signal events, divided by the number of all signal events coming to the classifier.
Background error, BackErr, is a ratio of number of background events after the classifier, i.e. erroneously recognized backgrounds as a signals divided by number of all background events coming to the classifier.
Enrichment factor

$$E = SigEff / BackErr .$$

All these variables are functions of the cut value.
Another approach uses samples sorted according to value of *R(x)*, the value of response of the classifier. Let in the same order be ordered the values corresponding to background (0 or -1) and signal (+1). These values let us denote v_i, i is order number, $i = 1, 2, ... m$, where m is the number of samples. The estimation quality can be measured by value

$$Q_s = \sum_{i=1}^{m} iv_i ,$$

which is maximal for ideal ordering all backgrouds first, then all signals or minimal for reversed ordering. Modifications are possible not including all samples using different summation interval.

Results

The metod was tested on the same data as was used in study [13]. Also, the third and next lines of the Table 1 are cited from this source and then we do not describe the different methods in detail. The same is true for Fig. 1 and Fig. 2, where behaviour of new method is shown by full line.

TABLE 1

Method	loacc	hiacc	E{0.5}	σ{0.5}	σ{max}	SigEff for σ{max}
New method	0.452	0.778	15.7	8.3984	9.345	0.364
Random Forest	0.448	0.851	13.5	8.17	8.72	0.334
Nearest Nb.	0.448	0.816	13.2	8.03	9.12	0.317
Kernel	0.443	0.803	14.1	8.43	8.64	0.39
C5.0 CART	0.419	0.816	13	7.94	8.5	0.233
NNSU	0.472	0.731	17.5	9.74	9.82	0.483
NeuNet	0.405	0.84	12.7	7.82	8.08	0.58
MRS	0.348	0.779	11.4	7.16	7.31	0.431
MLP	0.3	0.767	10.9	6.93	7.22	0.576
GMDH	0.28	0.736	10.2	6.55	6.77	0.574
Comp. prob.	0.332	0.728	10.6	6.78	6.83	0.585
Direct Sel.	0.306	0.636	9	5.91	7.52	0.153
LDA	0.195	0.638	8.2	5.47	5.8	0.71
SVM	0.124	0.586	7.1	4.81	5.76	0.784

The table gives the quality numbers loacc, hiacc, significance σ, and enrichment factor E with the following meaning:
loacc is the average signal efficiency obtained by interpolating the curve at the points 0.01, 0.02, and 0.05 for background error;
hiacc is obtained in a similar way by averaging signal efficiency at the points 0.1 and 0.2 background error;
enrichment factor E is defined by E = SigEff/BckErr, the value given is that obtained at signal efficiency = 0.5;
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 the curve, along with the value of SigEff where it is found (in many cases this is at a low, unacceptable SigEff).
Results are also shown graphically in Fig 1, the Neyman-Pearson diagram or decision quality diagram [13].

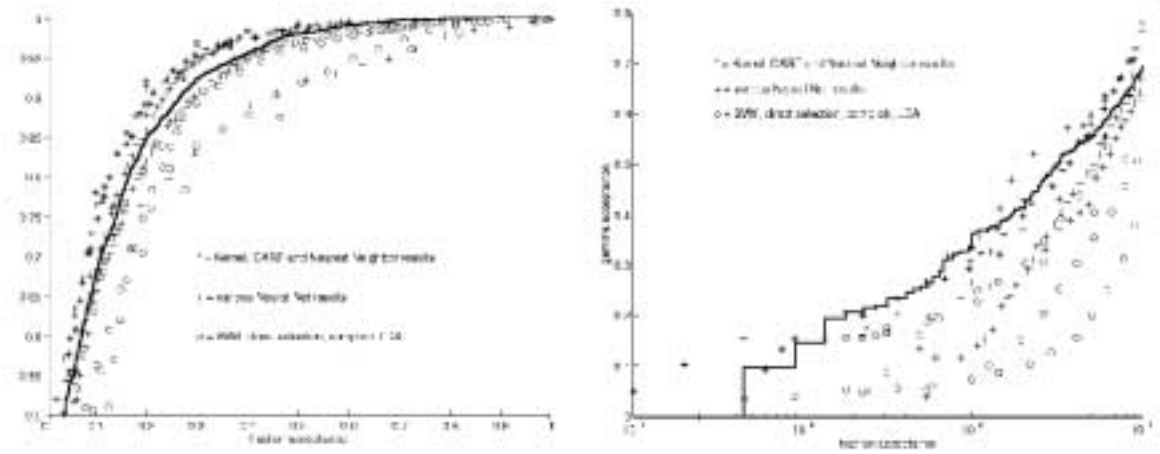

Fig 1. Decision quality diagram Signal efficiency (gamma acceptance) vs. Background error (hadron acceptance) for new method - full line. Left part for high signal efficiency, left for low background error. As background the diagrams from [13] were used.

Conclusion

The strongest and, at the same time, the weakest part of methods based on neighbours distances from a given (unknown) point is notion of distance, i.e. a simple transformation $E_n \rightarrow E_1$. With this transformation the curse of dimensionality is straitforwardly eliminated in the first step. The problem is what to do with two ordered set of distances, one for signal and the other for background. In this contribution it was shown that at least in statistical sense the (n-1)st power of these distances has stable behaviour and the sum of reciprocals of (n-1)st power of these distances is convergent.
The method proposed seems to be very good but not the best possible as seen in Table 1 and figures. Very essential advantage of the approach is the fact, that no tuning parameters exist. No neighbourhood size, no convergency coefficients etc. need to be set up in advance to assure convergency. 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 in the training set one element of sum (2) is computed. The amount of computation is thus proportional to training set size, i.e. the dimensionality times the number of training samples.

Acknowledgement

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

References

[1] Bhat, P.C.: Search for the Top Quark at D0 using Multivariate methods. CERN library hep-ex/9507007 V2, 15 Jul 1995.

[2] Lee, P.M.: Bayesian Statistics, an introduction. Oxford University Press New York, 1988.

[3] Heckermann D.: A Tutorial on Learning with Bayesian Networks. Technical Report MSRT-TR-95-06, Microsoft Research, Advanced Technology Division, Microsoft Corp., USA, 1995. ftp://ftp.research.microsoft.com/pub/dtg/david/

[4] Refregier, P., Vallet, F.: Probabilistic Approach for Multiclass Classification with Neural Networks. In: Artificial Neural Networks – Proc. of the 1991 Int. Conf. on Artificial Neural Networks (ICANN-91), Espoo, Finnland 24-28 June, 1991, Ed. T. Kohonen et al., North-Holland Amsterodam, 1991, Vol 2, 1003-1006.

[5] Vallet, F.: A Global Approach to Classification: Probabilistic Aspects. In: Artificial Neural Networks – Proc. of the 1991 Int. Conf. on Artificial Neural Networks (ICANN-91), Espoo, Finnland 24-28 June, 1991, Ed. T. Kohonen et al., North-Holland Amsterodam, 1991, Vol 2, pp. 1049-1052.

[6] Silverman, B. W.: Density Estimation for Statistics and data Analysis. Chapman and Hall, London, 1986.

[7] Hakl František, Jiřina Marcel 1999 Using GMDH Neural Net and Neural Net with Switching Units to Find Rare Particles. In: Artificial Neural Nets and Genetic Algorithms. (Ed.: Dobnikar A., Steele N. C., Pearson D. W., Albrecht R.) - Wien, Springer-Verlag 1999, pp. 52-58 (ISBN: 3-211-83364-1) Held: ICANNGA99, Portoroz, SI, 99.04.06-99.04.09

[8] Hakl František, 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

[9] An Overview of the CART Methodology. Salford Systems White Paper Series, http://www.salford-systems.com/, 2002

[10] Hájek, P.: Logics for Data Mining (GUHA rediviva). Neural Network World, Vol. 10 (2000) No. 3, pp. 299-300.

[11] Jiřina, M.: Nearest Neighbour Distance Statistics Estimation. Technical report ICS AS CR No. 878, 2002, pp.12.

[12] ATLAS Technical proposal for a General-Purpose pp Experiment at the Large hjadron Collider at CERN. CERN/LHCC/94-43 LHCC/P2 15 December 1994.

[13] Bock, R. K. et al.: Methods for multidimensional event classification: a case study using images from a Cherenkov gamma-ray telescope. To be published as Magic Internal Note in CERN, 200