KYBERNETIKA — VOLUME 34 (1998), NUMBER 4, PAGES 467 - 472

CONSTRUCTION OF A NONLINEAR DISCRIMINATION FUNCTION BASED ON THE MDL CRITERION

MANABU SATO, MINEICHI KUDO, JUN TOYAMA AND MASARU SHIMBO

Although a nonlinear discrimination function may be superior to linear or quadratic classifiers, it is difficult to construct such a function. In this paper, we propose a method to construct a nonlinear discrimination function using Legendre polynomials. The selection of an optimal set of Legendre polynomials is determined by the MDL (Minimum Description Length) criterion. Results using many real data show the effectiveness of this method.

1. INTRODUCTION

When sample distributions are normal distributions, the quadratic classifier is optimal in terms of the least expected error. In addition, as a special case, if these distributions have the same covariance matrix, the linear classifier becomes optimal. However such assumptions do not hold in many practical problems, and therefore linear or quadratic classifiers are not sufficient for a large variety of problems. On the other hand, k-NN or piecewise linear classifiers do not assume a statistical model in sample distributions and the decision rule is based on training samples. In these approaches, however, the decision rule is often too complex to work for unknown samples. Linear and quadratic classifier tend to be too simple, and k-NN and piecewise linear classifier tend to be too complex.

Although nonlinear discrimination functions have been shown to be effective theoretically, it is not clear how the function can be constructed. This paper presents a construction method of a nonlinear discrimination function with an appropriate complexity.

2. NONLINEAR DISCRIMINATION FUNCTION

We first consider two-class problems and then multi-class problems. The discrimination function is constructed in two stages: (1) transformation of the domain, and (2) construction by polynomials.

The idea is to construct a nonlinear discrimination function of a feature vector $\boldsymbol{x}$ as a linear function of $\boldsymbol{y}$ that is a multi-valued nonlinear function of $\boldsymbol{x}$. For example, it is known that the exclusive-or problem can be solved by a transformation of the domain. When two samples $(0,0)$ and $(1,1)$ are from ω_1 and the other

two samples $(1,0)$ and $(0,1)$ are from ω_2 in two-dimensional space, these samples can not be discriminated by a linear classifier. However we can construct a linear discrimination function $f(\boldsymbol{y}) = y_1 + y_2 - 2y_3 - 1/3$ by transformation of the domain: $y_1 = x_1,\ y_2 = x_2,\ y_3 = x_1 x_2$.

2.1. Transformation of domain

Let a transformation of the domain from $\boldsymbol{R}^n$ to $\boldsymbol{R}^{m+1}$ be

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T \quad \rightarrow \quad \boldsymbol{y}(\boldsymbol{x}) = (y_0(\boldsymbol{x}), y_1(\boldsymbol{x}), y_2(\boldsymbol{x}), \ldots, y_m(\boldsymbol{x}))^T,$$

where 'T' denotes the transposition, $y_i(\boldsymbol{x})$ is a function of $\boldsymbol{x}$, and $y_0(\boldsymbol{x})$ is a constant.

In our method, we express $y_i(\boldsymbol{x})$ by Legendre polynomials. The r-degree Legendre polynomials $P_r(x)$ is given by

$$P_r(x) \quad = \quad \frac{1}{2^r r!} \frac{\mathrm{d}^r}{\mathrm{d}x^r} (x^2 - 1)^r.$$

Normalized Legendre polynomials on $[-1,1]^n$ are

$$Q_r(x) \quad = \quad \frac{P_r(x)}{\|P_r(x)\|}, \quad \text{where} \quad \|P_r(x)\|^2 = \int_{-1}^{1} P_r^2(x)\,\mathrm{d}x.$$

These polynomials $\{Q_1(x), Q_2(x), \ldots, Q_r(x), \ldots\}$ make the most straight forward orthonormal system.

We define $y_i(\boldsymbol{x})$ $(i = 0, 1, \ldots, m)$

$$\begin{aligned} y_0(\boldsymbol{x}) &= 1/\sqrt{2} \\ y_1(\boldsymbol{x}) &= Q_1(x_1) \\ y_2(\boldsymbol{x}) &= Q_1(x_2) \\ &\vdots \\ y_{n+1}(\boldsymbol{x}) &= Q_1(x_1)\, Q_1(x_2) \\ &\vdots \end{aligned}$$

In general, $y_i(\boldsymbol{x})$ denoted by $y_i(\boldsymbol{x}) = Q_{r_1^i}(x_1)\, Q_{r_2^i}(x_2) \cdots Q_{r_n^i}(x_n)$ is specified by a sequence $\boldsymbol{r}^i = (r_1^i, r_2^i, \ldots, r_n^i)$, $r_j^i \in \{0, 1, 2, \ldots\}$. The degree of $y_i(\boldsymbol{x})$ is defined by

$$\deg(y_i(\boldsymbol{x})) = \sum_{j=1}^{n} r_j^i.$$

These functions $\{y_i(\boldsymbol{x})\}$ make an orthonomal system on $[-1,1]^n$, that is

$$\langle y_i(\boldsymbol{x}), y_j(\boldsymbol{x}) \rangle_{[-1,1]^n} \quad = \quad \int_{-1}^{1} \cdots \int_{-1}^{1} y_i(\boldsymbol{x})\, y_j(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \delta_{i,j},$$

where $\delta_{i,j}$ is Kronecker's delta.

2.2. Approximation of desired function by Legendre polynomials

We define a desired function in two classes ω_1 and ω_2 as

$$f(\boldsymbol{x}) = \begin{cases} 1 & (\boldsymbol{x} \in \omega_1) \\ -1 & (\boldsymbol{x} \in \omega_2) \end{cases}$$

and we approximate $f(\boldsymbol{x})$ by

$$g(\boldsymbol{y}(\boldsymbol{x})) = \boldsymbol{y}(\boldsymbol{x})^T \boldsymbol{a},$$

where $\boldsymbol{a} = (a_0, a_1, a_2, \ldots, a_m)^T$ is a coefficient vector. We determine $\boldsymbol{a}$ so as to minimize $\|f(\boldsymbol{x}) - g(\boldsymbol{y}(\boldsymbol{x}))\|^2$. We use $g(\boldsymbol{y}(\boldsymbol{x}))$ as a discrimination function.

Then, for N training samples $\{\boldsymbol{x}_i\}$ $(i = 1, 2, \ldots, N)$, we determine $\boldsymbol{a}$ by

$$\begin{aligned} \boldsymbol{a} &= Y^+ \boldsymbol{f}, \\ \boldsymbol{f} &= (f(\boldsymbol{x}_1), f(\boldsymbol{x}_2), \ldots, f(\boldsymbol{x}_N))^T, \\ Y &= (\boldsymbol{y}(\boldsymbol{x}_1), \boldsymbol{y}(\boldsymbol{x}_2), \ldots, \boldsymbol{y}(\boldsymbol{x}_N))^T, \end{aligned}$$

where Y^+ is the Moore–Penrose's general inverse matrix of Y.

2.3. Selection of features using the MDL criterion

The problem is how to select the optimal set $\{y_0, y_1, \ldots, y_m\}$. To identify f by g in every training sample, m must be $\geq N$. However, overfitting to training samples is known to be harmful for discrimination of unknown samples. Therefore, the problem is in how to select the optimal $\{y_i\}$ $(i = 0, 1, \ldots, m)$. We use the MDL criterion [3] to solve the problem.

At first, we prepare a candidature feature set S by

$$\begin{aligned} S &= \{y_i | y_i(\boldsymbol{x}) \text{ with } 0 \leq \deg(y_i(\boldsymbol{x})) \leq \ell\}, \\ m_0 &= |S| = \binom{n+\ell}{\ell}. \end{aligned}$$

We select ℓ as the least integer satisfying $m_0 \geq N$.

For a subset T $(ITI = m)$ of S, the MDL criterion value is calculated by

$$\mathrm{MDL}(T) = \frac{N}{2} \log_2 \frac{\varepsilon^2(T)}{N} + \frac{m}{2} \log_2 N,$$

where

$$\begin{aligned} \varepsilon^2(T) &= \sum_{i=1}^{N} (f(\boldsymbol{x}_i) - g(\boldsymbol{y}(\boldsymbol{x}_i)))^2, \\ \boldsymbol{y} &= (y_0, y_1, \ldots, y_m)^T, \quad y_i \in T. \end{aligned}$$

In the equation, the first term shows how well T explains the training samples and the second term shows how degree T is complicated. As a result, the criterion choose a smaller number of extended features and a lesser error for the training samples.

We select T so as to minimize MDL(T). To get a near-optimal T in a limited time, we adopt a sequential search. The algorithm is as follows.

1. Sort $\{y_i(\boldsymbol{x})\}$ $(i = 1, 2, \ldots, m_0)$ and rename them such that $\varepsilon^2(\{y_1(\boldsymbol{x})\}) \leq \varepsilon^2(\{y_2(\boldsymbol{x})\}) \leq \cdots \leq \varepsilon^2(\{y_{m_0}(\boldsymbol{x})\})$.
2. Let $T \leftarrow \{y_0(\boldsymbol{x})\}$ and $i \leftarrow 1$.
3. while (i $\leq m_0$) {
4. $\quad U \leftarrow T \cup \{y_i(\boldsymbol{x})\}$
5. $\quad \Delta\text{MDL} \leftarrow \frac{N}{2} \log_2 \frac{\varepsilon^2(U)}{\varepsilon^2(T)} + \frac{1}{2} \log_2 N$
6. $\quad$ if (Δ MDL < 0) then $T \leftarrow U$
7. $\quad i \leftarrow i + 1$
8. }

We make a margin in the domain $[-1, 1]^n$ so as to let training samples exist in $[-0.6, 0.6]^n$. In addition, for multi-class problems, we adopt a pair-wise discrimination in which we construct a discrimination function for every pair of classes, and we assign an unknown sample to a class by the majority rule.

3. EXPERIMENTS

We compared our method with Bayes linear, Bayes quadratic and the k-NN classifier ($k = 1, 5$) using many data. We used a d-fold ($d = 10$) cross-validation [4] (10-fold CV) to estimate the recognition rate.

A. Artificial data

First, we experimented on four sets of artificial data.

- Two-class normal distributions ("Norm1"): One class has mean $(0, 0)$ and covariance matrix I and the other has $(0, 0)$ and 4I.
- Two-class normal distributions ("Norm2"): One class has mean $(0, 0)$ and covariance matrix I and the other has $(3, 2)$ and $\begin{pmatrix} 4 & 0 \\ 0 & 2 \end{pmatrix}$.
- Two-class uniformly distributions ("Square"): One class is distributed within an inner square and the other is distributed between the inner and outer squares.
- Two-class arc-type data("Arc"): Both classes have a similar arc, and one class is shifted up by a half of the other.

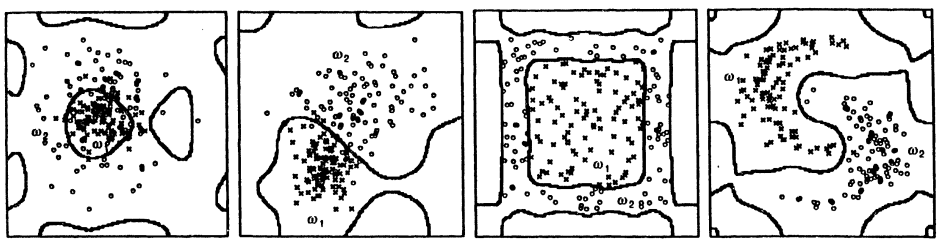
For all data, 100 of samples per class are used.

The results are shown in Table 1, and the training samples and the estimated boundary are shown in Figure 1. This figure shows that our approach succeeded in constructing an appropriate discriminate boundary.

Table 1. Results of artificial data.

Data Name	$n^{1)}$	$c^{2)}$	Discrimination rate [%] by 10-fold CV				
			Proposed method ($m^{3)}$)	Linear	Quad.	1-NN	5-NN
Norm1	2	2	73.7 (2.6)	58.9	**78.1**	64.1	73.6
Norm2	2	2	94.7 (10.2)	91.0	**96.3**	92.0	93.2
Square	2	2	**99.0** (16.0)	47.6	85.0	94.3	90.9
Arc	2	2	**99.5** (17.5)	95.3	95.7	98.6	98.6

1) n: Number of features. 2) c: Number of classes. 3) m: Average number of selected features.



(a) Norm1 (b) Norm2 (c) Square (d) Arc

Fig. 1. Discrimination boundaries by the proposed method.

B. Real data

Next, we dealt with many real data sets. Excepting ETL3, all data were taken from Machine Learning Databases [2]. ELT3 is a 26-class alphabetical character database taken from The ETL-3 database [1]. We used the first ten coefficients by Karhunen–Loéve expansion of the original 81 features in the data set.

The results are shown in Table 2. The rank in the discrimination rate is shown in Table 3. As can be seen in the tables our method is superior to others in most cases.

For the data set "glass", however, our method showed the lowest performance. The reason for this seems to be due to the MDL criterion, because the MDL criterion tends to select a simpler discrimination rule than the optimal rule when the number of training samples is small. Indeed, the ratios of the training samples to the number of original feature are from 1 to 8.4 depending on classes.

4. CONCLUSION

We proposed a method to construct a nonlinear discrimination function on the basis of Legendre polynomials and the MDL criterion. The proposed method showed better performance than other methods for many problems.

However, when there are only a few training samples, the proposed method cannot construct a good rule. Adoption of a more effective selection of features is a topic

Table 2. Results of real data.

Data Name	n[1]	c[2]	Number of samples [per Class]	Discrimination rate[%] by 10-fold CV				
				Proposed method (m[3])	Linear	Quad.	1-NN	5-NN
ETL3	10	26	all 100	**95.8** (46.5)	88.6	95.0	92.4	92.4
balance-scale	4	3	288,49,288	**93.6** (26.5)	70.4	91.6	87.1	87.6
breast-cancer-wisconsin	9	2	458,241	96.3 (46.8)	96.1	95.1	95.4	**96.9**
glass	9	6	70,17,76,13,9,29	65.1 (25.7)	67.7	66.5	**72.1**	68.4
heart-disease	14	4	303,29,123,200	**89.4** (24.4)	72.5	56.5	56.9	58.6
hepatitis	19	2	32,123	**76.2** (4.6)	64.7	76.1	66.6	71.8
ionosphere	34	2	225,126	**92.8** (26.5)	86.7	89.1	86.3	84.1
liver-disorders	6	2	145,200	**70.2** (8.7)	62.4	57.8	62.3	68.6
pima-indians-diabetes	8	2	500,268	**76.4** (13.5)	74.9	72.9	66.8	70.7
australian	14	2	383,307	**86.4** (16.7)	85.5	79.6	66.2	69.1
vowel	10	11	all 48	96.6 (61.5)	62.2	91.1	**98.4**	93.9

1) n: Number of features. 2) c: Number of classes. 3) m: Average number of selected features.

Table 3. Rank of discrimination rate.

Classifier	1st	2nd	3rd	4th	5th
Proposed method	8	2	0	0	1
Linear	0	3	4	0	4
Quadratic	0	4	2	2	3
1-NN	2	0	1	6	2
5-NN	1	2	4	3	1

for further study.

(Received December 18, 1997.)

REFERENCES

[1] ETL3. 1993. Japanese Technical Committee for Optical Character Recognition, ETL Character Database. Electrotechnical Laboratory, 1993.

[2] P. M. Murphy and D. W. Aha: UCI Repository of Machine Learning Databases [Machine-Readable Data Repository]. University of California, Department of Information and Computer Science, Irvine 1991.

[3] J. Rissanen: A universal prior for integers and estimation by minimum description length. Ann. Statist. *11* (1983), 416–431.

[4] M. Stone: Cross-Validatory choice and assessment of statistical predictions. J. Roy. Statist. Soc. *36* (1974), 111–147.

Manabu Sato, Mineichi Kudo, Jun Toyama, Masaru Shimbo, Division of Systems and Information Engineering, Graduate School of Engineering, Hokkaido University, Sapporo 060-8628. Japan.

e-mails: gaku, mine, jun, shimbo@main.eng.hokudai.ac.jp