Complementary Discriminant Analysis for Classification of Double Attributes

Kazuyuki Hiraoka and Taketoshi Mishima
Department of Information and Computer Sciences, Saitama University,
255 Shimo-Okubo, Saitama, 338-8570, Japan
Tel. +81-48-858-3723, Fax +81-48-858-3723
Email: hira@ics.saitama-u.ac.jp

Abstract:

Real-world objects often have two or more significant attributes. For example, face images have attributes of persons, expressions, and so on. Even if we are interested in only one of those attributes, additional informations on auxiliary attributes can help recognition of the main one.

In the present paper, the authors propose a method for pattern recognition with double attributes. A pair of classifiers are combined: each classifier makes a guess of its corresponding attribute, and it tells the guess to the other as a hint. Equilibrium point of this iteration can be calculated directly without iterative procedures.

1. Introduction

Pattern recognition on one attribute has been studied widely[2]. However, real-world problems often have two or more attributes. For example, face images have attributes of persons, expressions, and so on. Even if we are interested in only one of those attributes, additional informations on auxiliary attributes can help recognition of the main one. In the present paper, the authors propose a method for pattern recognition with double attributes.

1.1 Task

As training samples, n vector data $\boldsymbol{x}(1), \cdots, \boldsymbol{x}(n)$ are presented. In addition, double attributes (s, c) for each $\boldsymbol{x}$ are also presented:

$$\boldsymbol{x}(t) = (x_1(t), \ldots, x_N(t))^T \in R^N, \quad (1)$$
$$s(t) \in \mathcal{S} = \{1, \cdots, S\}, \quad (2)$$
$$c(t) \in \mathcal{C} = \{1, \cdots, C\}, \quad (3)$$
$$(t = 1, \cdots, n), \quad (4)$$

where T denotes matrix transposition. Then, a new datum $\boldsymbol{x}$ is presented and estimation of its attributes (s, c) is required.

A solution for this task has been proposed in [1] for the case that the whole data can be approximated by the bilinear model. In the present paper, different approach is proposed for general cases.

1.2 Naive methods

Naive methods for this problem are shown in Fig. 1:

- Combined approach: Consider the pair (s, c) as one complex label, and construct a classifier for SC classes.

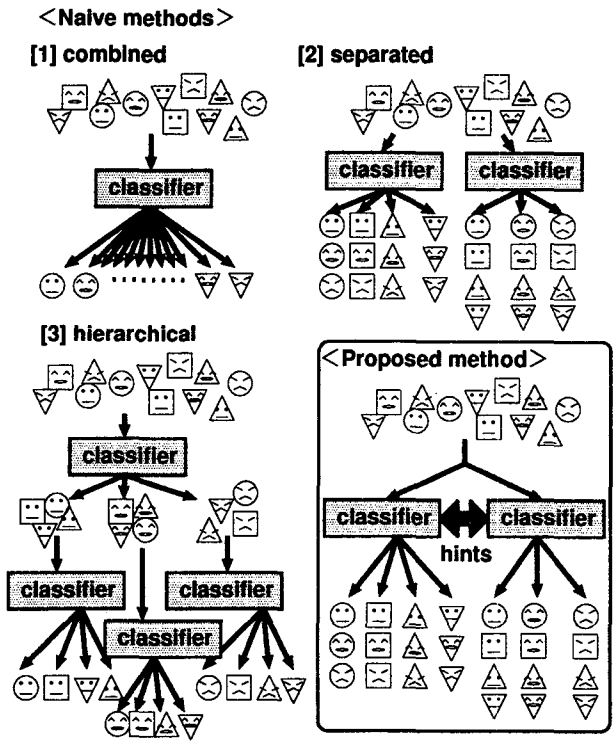

Figure 1. Comparison among naive methods and the proposed method

- Separated approach: Construct two independent classifiers, one for s and one for c.
- Hierarchical approach: First, construct a classifier for s, and then construct "expert" classifiers for c. Each expert is specialized to one s.

They have a drawback that they do not use the structure of the task efficiently.

Combined approach requires many samples since the number SC of classes can be large and sufficient samples are required for each class. This is caused from lack of use of the fact that samples of classes (s, c') and (s', c) often have some information on the class (s, c).

Separated approach is opposite in the sense that it ignores given label c for classification of s and vice versa. Unfortunately, the distribution of samples in class s is different according to c in many cases and then separated approach is inadequate.

Hierarchical approach looks more feasible. However, it lacks feedback from the guess of c to the guess of s. Then it has disadvantages which are similar to separated approach for s and to combined approach for c.

2. Proposed method

Suppose that we have a pair of classifiers $\boldsymbol{f}(\boldsymbol{x}, c) = (f_1(\boldsymbol{x}, c), \ldots, f_S(\boldsymbol{x}, c))$ and $\boldsymbol{g}(\boldsymbol{x}, s) = (g_1(\boldsymbol{x}, s), \ldots, g_C(\boldsymbol{x}, s))$, where $f_s(\boldsymbol{x}, c)$ and $g_c(\boldsymbol{x}, s)$ are estimations of conditional probabilities $p(s|\boldsymbol{x}, c)$ and $p(c|\boldsymbol{x}, s)$, respectively. The classifier $\boldsymbol{f}$ is trained for combined input $(\boldsymbol{x}(t), c(t))$ and simple output $s(t)$, while $\boldsymbol{g}(t)$ is for $(\boldsymbol{x}(t), s(t))$ and $c(t)$. These $\boldsymbol{f}$ and $\boldsymbol{g}$ are black-boxes throughout the proposed method: arbitrary classifiers can be used for $\boldsymbol{f}$ and $\boldsymbol{g}$ as far as they output conditional (posterior) probabilities of classes (Fig. 1).

For a new datum $\boldsymbol{x}$, we want to obtain $p(s|\boldsymbol{x})$ and $p(c|\boldsymbol{x})$ from the estimated $p(s|\boldsymbol{x}, c)$ and $p(c|\boldsymbol{x}, s)$. Note that $p(s|\boldsymbol{x})$ and $p(c|\boldsymbol{x})$ satisfy

$$p(s|\boldsymbol{x}) = \sum_{c=1}^{C} p(s|\boldsymbol{x}, c)p(c|\boldsymbol{x}), \quad (5)$$

$$p(c|\boldsymbol{x}) = \sum_{s=1}^{S} p(c|\boldsymbol{x}, s)p(s|\boldsymbol{x}). \quad (6)$$

These equations lead us to Markov chains naturally. Let

$$P = \begin{pmatrix} p(s=1|\boldsymbol{x}, c=1) & \cdots & p(s=1|\boldsymbol{x}, c=C) \\ \vdots & & \vdots \\ p(s=S|\boldsymbol{x}, c=1) & \cdots & p(s=S|\boldsymbol{x}, c=C) \end{pmatrix}, \quad (7)$$

$$Q = \begin{pmatrix} p(c=1|\boldsymbol{x}, s=1) & \cdots & p(c=1|\boldsymbol{x}, s=S) \\ \vdots & & \vdots \\ p(c=C|\boldsymbol{x}, s=1) & \cdots & p(c=C|\boldsymbol{x}, s=S) \end{pmatrix}. \quad (8)$$

Then, the goal values

$$\boldsymbol{p} = (p(s=1|\boldsymbol{x}), \cdots, p(s=S|\boldsymbol{x}))^T, \quad (9)$$

$$\boldsymbol{q} = (p(c=1|\boldsymbol{x}), \cdots, p(c=C|\boldsymbol{x}))^T \quad (10)$$

are invariant distributions of Markov chains whose transition matrices are PQ and QP, respectively, since $\boldsymbol{q} = Q\boldsymbol{p}$ and $\boldsymbol{p} = P\boldsymbol{q}$. These $\boldsymbol{p}$, $\boldsymbol{q}$ are directly obtained as the solution[1] of

$$(PQ - I)\boldsymbol{p} = O, \quad (11)$$

$$(QP - I)\boldsymbol{q} = O, \quad (12)$$

$$\mathbf{1}^T\boldsymbol{p} = \mathbf{1}^T\boldsymbol{q} = 1, \quad (13)$$

$$\mathbf{1} \equiv (1, \ldots, 1)^T, \quad (14)$$

without iterative procedures. Finally, the results of classification

$$\hat{s} = \arg\max_{s \in \mathcal{S}} p(s|\boldsymbol{x}), \quad (15)$$

$$\hat{c} = \arg\max_{c \in \mathcal{C}} p(c|\boldsymbol{x}) \quad (16)$$

are output.

[1]Existence of solution is guaranteed since PQ and QP are stochastic matrices, while uniqueness of solution is not guaranteed in general.

3. Experiment

3.1 Task

The proposed method is examined for basic artificial tasks (Fig. 2, 3: left). Parameters of the task are shown in Table 1, where I denotes the identity matrix.

Table 1. Parameters of experiments

number of classes	$(S, C) = (3, 3)$
number of samples	$n = 50 \times S \times C = 450$
dimension of data $\boldsymbol{x}$	$N = 2$
within-class distribution of data $\boldsymbol{x}$ (within-class variance)	Gaussian ($V = 0.3^2 I$)
classifiers f_s, g_c	Fisher linear discriminant
dimension of projected discriminant space	$L = 1$ or 2

3.2 Classifier

Classifier $\boldsymbol{f}$ for these experiments consists of three "experts" $\boldsymbol{f}(\cdot, 1)$, $\boldsymbol{f}(\cdot, 2)$, $\boldsymbol{f}(\cdot, 3)$ which correspond to $c = 1, 2, 3$, respectively. These experts are independently trained with only samples which have corresponding value of $c(t)$. Assuming that within-class distribution of $\boldsymbol{x}(t)$ in each (s, c) is Gaussian with a common unknown variance matrix V, we use Fisher linear discriminant[2] as each expert. Classifier $\boldsymbol{g}$ is also constructed similarly.

3.3 Result

Results of experiments are shown in Fig. 2, 3 (middle and right). In spite that the classifier consists of simple linear discriminant functions, feasible boundaries are obtained successfully.

4. Discussion

We can obtain smoother boundaries if we introduce relaxation

$$f'_s(\boldsymbol{x}, c) \propto f_s(\boldsymbol{x}, c)^\beta, \quad \sum_{s=1}^{S} f'_s(\boldsymbol{x}, c) = 1 \quad (17)$$

$$g'_c(\boldsymbol{x}, s) \propto g_c(\boldsymbol{x}, s)^\beta, \quad \sum_{c=1}^{C} g'_c(\boldsymbol{x}, s) = 1 \quad (18)$$

and use f'_s, g'_c as the components of P, Q instead of raw f_s, g_c, where β is a positive constant less than 1 (Fig. 2 lower).

In the above experiments, C "expert" classifiers $\boldsymbol{f}(\cdot, 1), \boldsymbol{f}(\cdot, 2), \ldots, \boldsymbol{f}(\cdot, C)$ are trained independently with only data which have corresponding value of $c(t)$. This makes number of available samples for each $\boldsymbol{f}(\cdot, c)$ smaller. In order to utilize informations in samples more efficiently, we have to adopt a classifier $\boldsymbol{f}$ which can deal with c more properly as a "hint". One natural idea is the use of single classifier with input vector

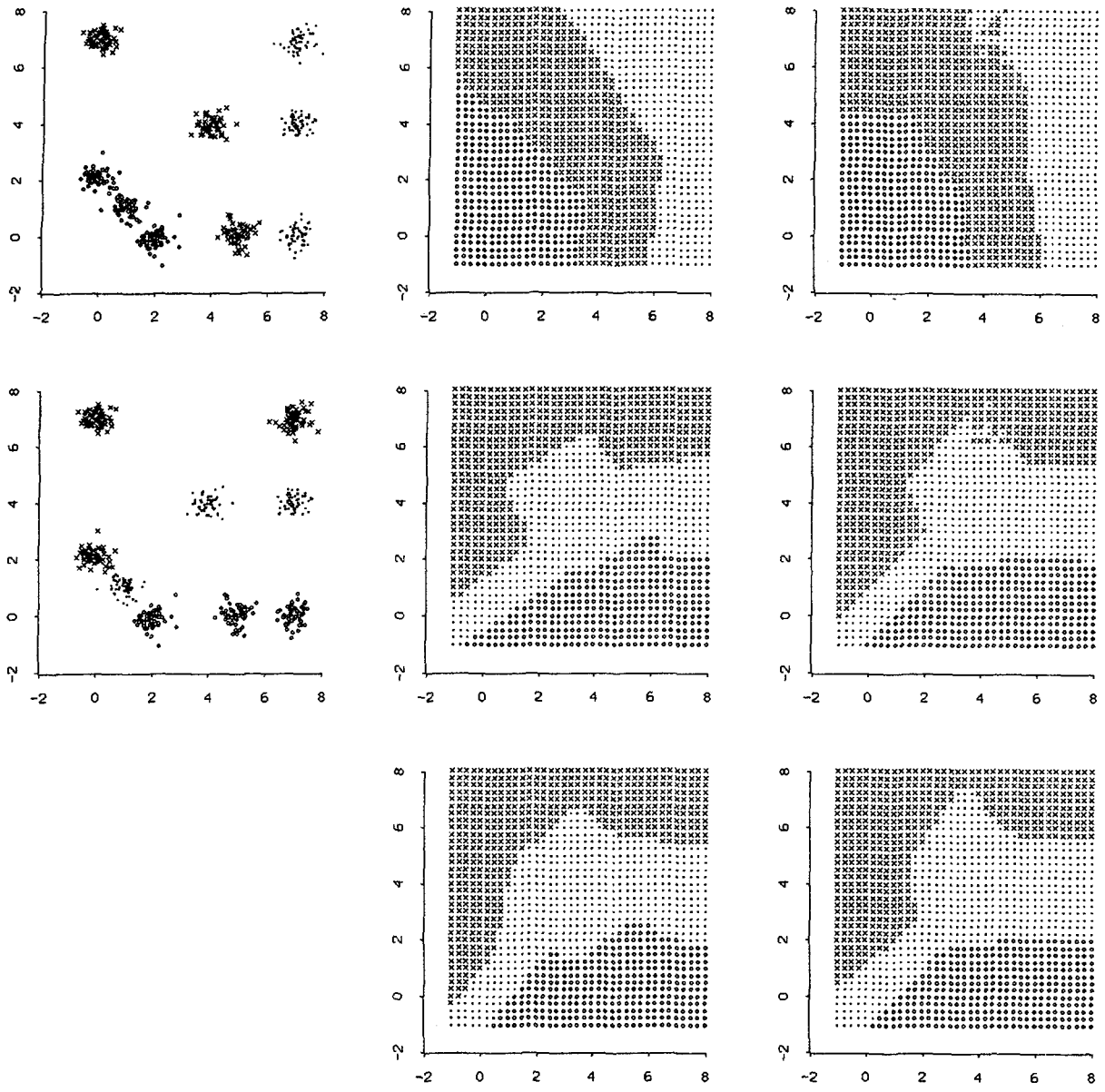

Figure 2. Experiment I (matrix-type structure)
Upper: attribute $s = 1(\bigcirc), 2(\cdot), 3(\times)$. Middle: attribute $c = 1(\bigcirc), 2(\cdot), 3(\times)$. Lower: s with relaxation $\beta = 0.1$ (see section 4).
Left: presented samples. Middle and Right: obtained boundaries of classification for $L = 1$ and $L = 2$, respectively, where L is the dimension of projected discriminant space.

$(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \delta_{1c}, \ldots, \delta_{Cc})^T$, where δ_{ij} is 1 if $i = j$ and 0 otherwise.

Another problem of the proposed method is inconsistency of $\boldsymbol{f}$ and $\boldsymbol{g}$. For arbitrarily given $\boldsymbol{f}$ and $\boldsymbol{g}$, in general, there is no distribution[2] $p(\boldsymbol{x}, s, c)$ whose conditional probabilities are $p(s|\boldsymbol{x}, c) = f_s(\boldsymbol{x}, c)$ and $p(c|\boldsymbol{x}, s) = g_c(\boldsymbol{x}, s)$. A mediation mechanism is thus desired. We are constructing a mediation algorithm from a point of view of information geometry [3]. We are expecting that the extension to three or more attributes may be also possible by use of this mediation.

[2]We can always find $p(s|\boldsymbol{x})$ and $p(c|\boldsymbol{y})$ which satisfy (5)(6). However, (5)(6) are not sufficient for guarantee of consistency. In order to guarantee the consistency, we need "detailed balance" condition $p(c|\boldsymbol{x}, s)p(s|\boldsymbol{x}) = p(s|\boldsymbol{x}, c)p(c|\boldsymbol{x})$ for all s, c.

5. Conclusion

A method for classification with double attributes has been proposed. The main idea is mutual suggestion of hints between a pair of classifiers. In spite that this idea sounds like an iterative algorithm, the solution can be obtained directly without iterative procedures. Its

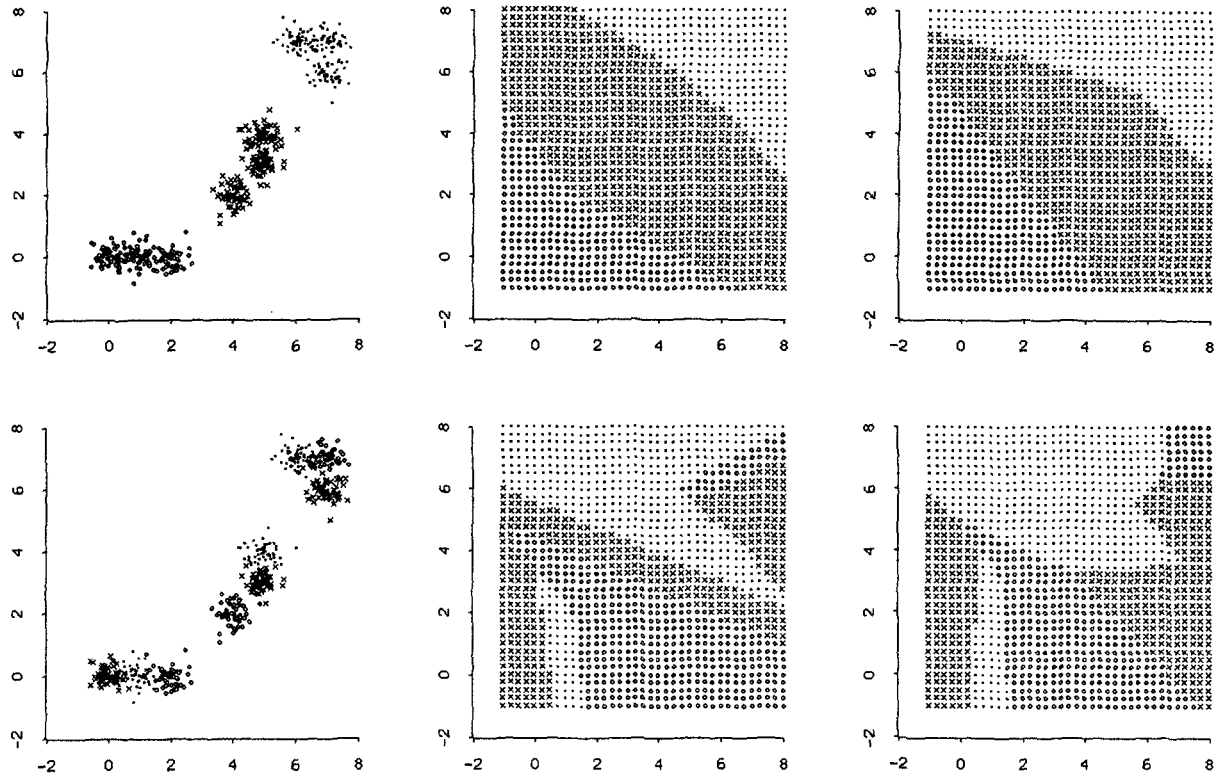

Figure 3. Experiment II (cluster-type structure)
Upper: attribute $s = 1(\bigcirc), 2(\cdot), 3(\times)$. Lower: attribute $c = 1(\bigcirc), 2(\cdot), 3(\times)$.
Left: presented samples. Middle and Right: obtained boundaries of classification for $L = 1$ and $L = 2$, respectively, where L is the dimension of projected discriminant space.

behavior has been examined for artificial tasks which have structures of "matrix" and "cluster".

Some issues on the proposed method has been discussed already in the previous section. It works well when samples have some structures, while it can fail when samples have no structure at all. At now, it is not clear which structures are suitable. In addition, experimental comparisons with [1] and naive methods for larger, real-world problems are also important future works.

This work has been partly supported by JSPS (Japan Society for the Promotion of Science), 14580405.

References

[1] Joshua B. Tenenbaum, "Separating style and content with bilinear models", Neural Computation, Vol. 12, pp. 1247–1283, 2000.

[2] Richard O. Duda, E. Hart, and David G. Stork, Pattern classification, 2nd ed., John Wiley & Sons, Inc., New York, 2001.

[3] Shun-ichi Amari, Differential-Geometrical Methods in Statistics, Lecture Notes in Statistics, 28, Springer-Verlag, 2nd printing, 1990.