社団法人　情報処理学会　研究報告
IPSJ SIG Technical Report

2006－MPS－58（17）
2006／3／17

# 1 対 1 識別スコアの総和による多種識別 ～スコアの制約と種類変動時の再計算

田中 秀俊 [1]

**概要**　多種識別法のうち、1 対 1 の識別を総合して行う方法を採用する場合、1 対 1 識別の結果のペアワイズ関数とその総合関数にそれぞれ選択肢がある。総合関数としてペアワイズ関数の総和による比較を採用すると、識別すべき種類が識別後に変動した場合に再計算にかかる計算量が少ないこと、ただし総合結果との整合性を保つには、ペアワイズ関数の総合スコアを識別結果の否定の総和とすべきことを示す。

# Multiclassification by Summation of Pairwise Classification Scores

Hidetoshi Tanaka[1]

**Abstract**　Multiclassification problems are often binarized into pairwise classifications in order to utilize basic classification methods. The binarization contains two functions to determine: the score function of the pairwise classification and their aggregation function. Summation aggregation costs small computation in class reconfiguration, and requires negative score functions to be consistent with the pairwise classifications.

## 1　多種識別の方法

多種識別の問題で、その部分問題である 1 種類対 1 種類での分離性の良さが、線形判別や SVM などの 1 対 1 を識別する方法を用いて既に分かっている場合、多種の得意な決定木や 1 種類対他種類の他の識別法を改めて試すのではなく、その性能の良い 1 対 1 識別を総合して多種を識別したい。対象が 2 次元であればボロノイ図を描く問題に帰着すべきだが、高次元の場合には、1 対 1 識別の結果をスコア化し、リーグ戦方式やランダムなトーナメント戦方式で勝者を決める方が、実装や結果確認の容易さなどメリットが大きい。トーナメント構造をランダムに決める方法 [1] は、最も有望な結果を決めるには最速の方法だが、2 位以下を決めるためには再度トーナメントの必要があり、また、3 すくみの領域の結果が、そこが 3 すくみだったという情報ぬきにランダムに 1 つ出てしまうという欠点がある。3 すくみの領域の 2 つの例を図 1に示す。図 1には種類 A,B,C に属するサンプルがそれぞれ a,b,c で示されている。最も基本的な線形カーネルの SVM で 1 対 1 識別平面を作成すると、例えば図のように a と b、b と c、c と a の垂直二等分超平面が境界超平面となるイメージで境界面が構成される。このとき A 対 B では A、B 対 C では B、C 対 A では C という識別結果になる図中の (*) のような箇所 (3 すくみ領域) が発生し、種類決定不能となる。

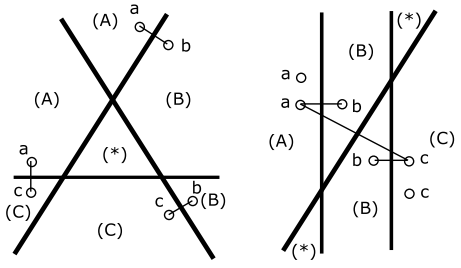


図 1: 3 すくみの領域 (*)

[1] 三菱電機（株）情報技術総合研究所
Mitsubishi Electric Corp., Info. Tech. R&D Center

一方、リーグ戦方式には、ペアワイズ関数と総合関数という 2 つの決定すべきスコア関数がある。ここでペアワイズ関数とは 1 対 1 識別の結果を表現するもので、勝者を 1 点、敗者を −1 点、引分けを 0 点とする単純な関数を端的な例として、勝ち方や負け方を評価する適当なメンバーシップ関数を採用することができる。総合関数は、ペアワイズ関数の値を総合して最高スコアを判定結果とするための関数で、単純に値を加算する方法やファジィ論理積を用いる方法などがある。これらスコア関数の決め方の違いは、上述の 3 すくみ領域において表面化しやすい。著者らは、ペアワイズ関数に二重否定値を、総合関数に総和を採用した二重否定値加算法 [2] を提案し、検討を進めている。

## 2 二重否定スコアと種類削減

大量サンプルの識別作業の途中に、いくつかの種類についての存在が否定され、あるいは識別の必要がなくなって、それまでに識別済のサンプルの識別結果をすべて再計算しなければならない場合がある。そのようなとき、総合関数に値の総和を用いておくと、再計算の手間が少なくてすむ。削減される種類とのペアワイズスコアを改めて算出するか、あるいは記憶しておけば、総合スコアから引き算するだけで新たな総合スコアを得られるからである。例えば、サンプル数が $n$ 個で、種類数が $N$ から 1 つ減った場合、ペアワイズスコア総再計算回数は $\frac{1}{2}n(N-1)(N-2)$ に対し、削減種類とのペアワイズスコア再計算回数は $n(N-1)$ ですむ。

本章では、この削減種類のペアワイズ再計算による結果修正の妥当性について、簡単な検証結果を示す。実験は、UCI machine learning repository から cardiac arrhythmia データベース (13 クラス) を対象とした [3]。識別には SVM-light[4] を用いた。スコア関数には、SVM の判別関数をペアワイズ関数のベースとする肯定スコア、二重否定スコア、ファジイソフトマージンスコアの 3 種類を用いた。これらのスコアについて簡単に述べる。

種類 A 対種類 B の判別超平面を支える点として、種類 A に所属する点 $\boldsymbol{a}$ と種類 B に所属する点 $\boldsymbol{b}$ を考える。SVM で得られた判別関数は、両点を適切にとることによって $\mathrm{SVM}(\boldsymbol{x},\boldsymbol{a},\boldsymbol{b})$ の形で式 (1) のように表せる。

$$\mathrm{SVM}(\boldsymbol{x},\boldsymbol{a},\boldsymbol{b}) = \frac{2}{||\boldsymbol{a}-\boldsymbol{b}||^2}(\boldsymbol{a}-\boldsymbol{b})(\boldsymbol{x}-\frac{\boldsymbol{a}+\boldsymbol{b}}{2}) \quad (1)$$

式 (1) は $\mathrm{SVM}(\boldsymbol{a},\boldsymbol{a},\boldsymbol{b}) = 1$、$\mathrm{SVM}(\boldsymbol{a},\boldsymbol{b},\boldsymbol{a}) = -1$ となっている。点 $x$ が種類 A 対種類 B の判別で種類 A に所属するメンバーシップ値 $a_b(\boldsymbol{x})$ は、例えば式 (1) を用いて式 (2) のように定義できる。

$$a_b(\boldsymbol{x}) = \begin{cases} 0 & (\mathrm{SVM}(\boldsymbol{x},\boldsymbol{a},\boldsymbol{b}) \le 0) \\ \mathrm{SVM}(\boldsymbol{x},\boldsymbol{a},\boldsymbol{b}) & \\ 1 & (\mathrm{SVM}(\boldsymbol{x},\boldsymbol{a},\boldsymbol{b}) \ge 1) \end{cases} \quad (2)$$

各種類についてこのペアワイズ関数の総和をとる総合スコアを、本稿では肯定スコアと呼ぶ。各種類の可能性を否定するペアワイズ関数以外の総和をとる総合スコアを二重否定スコアと呼ぶ。また、式 (3) の $a'_b(\boldsymbol{x})$ のように下限を設けないペアワイズ関数を用い、総合関数にそのファジイ論理積を用いる総合スコアを本稿ではファジイソフトマージンスコアと呼ぶ [5]。

$$a'_b(\boldsymbol{x}) = \begin{cases} \mathrm{SVM}(\boldsymbol{x},\boldsymbol{a},\boldsymbol{b}) & (\mathrm{SVM}(\boldsymbol{x},\boldsymbol{a},\boldsymbol{b}) < 1) \\ 1 & (\mathrm{SVM}(\boldsymbol{x},\boldsymbol{a},\boldsymbol{b}) \ge 1) \end{cases} \quad (3)$$

学習フェーズにおいて、SVM-light を用い、13 種類の 1 対 1 識別器、計 78 個を $C = 0.001$ で生成して各関数の値を得、多種識別を実施した。その結果を表 1に示す。PO は肯定スコア、DN は二重否定スコア、FS はファジイソフトマージンスコアの結果である。テストフェーズでは学習に用いた全データをそのまま用いた（自己認識評価）。X/Y は正解 (X) 対不正解 (Y) を示す。DN は FS と同じ結果となった。一方、PO は自己認識評価の場合、このように若干良い成績となる。

ここで、種類 C16 が削除された場合を想定する。PO と DN では、C16 に関する 12 個の 1 対 1 識別スコア (C01–C16 ～ C15–C16) が再計算され、各サンプルについて引き算が行われる。FS ではスコアは改めて全部再計算される。結果を表 2に示す。$X$–$Y$ は $X$ 個のサンプルが C16 削除前には識別されていたが、削除後には $Y$ 個になったことを表す。スコアを再計算した FS に対し、DN は引き算で同じ結果を出している。

表 1: 自己認識評価

| 種 | 計 | PO | DN | FS |
|---|---|---|---|---|
| C01 | 245 | 243/2 | 240/5 | 240/5 |
| C02 | 44 | 34/10 | 29/15 | 29/15 |
| C03 | 15 | 15/0 | 15/0 | 15/0 |
| C04 | 15 | 13/2 | 13/2 | 13/2 |
| C05 | 13 | 7/6 | 7/6 | 7/6 |
| C06 | 25 | 9/16 | 7/18 | 7/18 |
| C07 | 3 | 2/1 | 2/1 | 2/1 |
| C08 | 2 | 2/0 | 2/0 | 2/0 |
| C09 | 9 | 9/0 | 9/0 | 9/0 |
| C10 | 50 | 44/6 | 44/6 | 44/6 |
| C14 | 4 | 4/0 | 4/0 | 4/0 |
| C15 | 5 | 5/0 | 5/0 | 5/0 |
| C16 | 22 | 5/17 | 4/18 | 4/18 |

表 2: C16 の削除

| Class | PO | DN | FS |
|---|---|---|---|
| C01 | 12–15 | 14–16 | 14–16 |
| C02 | 1–2 | 1–1 | 1–1 |
| C03 | 0–1 | 0–1 | 0–1 |
| C04 | 0–0 | 0–1 | 0–1 |
| C06 | 1–1 | 0–0 | 0–0 |
| C10 | 2–2 | 2–2 | 2–2 |
| C15 | 1–1 | 1–1 | 1–1 |
| C16 | 5–0 | 4–0 | 4–0 |

## 3 総和で総合する場合のスコア関数

総合関数にペアワイズスコアの総和を用いる場合、個々の 1 対 1 識別結果と整合しない総合結果が出てしまうケースが簡単に想定できる。例えば A,B,C の 3 種類の識別において、ペアワイズスコアとして 0 〜 1 の実数値、総合関数としてその加算を採用すると、A 対 B では A に 0.5、A 対 C でも A に 0.4 だったとしても、B 対 C で B に 1.0 が加算されると総合結果は 0.1 の差で B になり、A 対 B の結果と不整合になる。本章では、ペアワイズスコアと総合スコアとの整合性をとるための方法について検討する。

種類 A 対種類 B の 1 対 1 識別に基づいて、ある対象 $\boldsymbol{x}$ の種類 A に対するメンバーシップ値を $\mathcal{S}(\boldsymbol{x}, A, B)$、種類 B に対するメンバーシップ値を $\mathcal{S}(\boldsymbol{x}, B, A)$ とする。関係する種類が A と B だけの場合、$\mathcal{S}(\boldsymbol{x}, A, B) > \mathcal{S}(\boldsymbol{x}, B, A)$ であれば、この 1 対 1 識別結果は A になる。

関係する種類全体を $U$ とする。ごく普通の肯定スコアによる総合では、各種類 $u \in U$ の総合スコアを式 (4) のように定め、最大の $T(\boldsymbol{x}, u)$ となる $u$ を $\boldsymbol{x}$ の総合識別結果とする。

$$\mathcal{T}(\boldsymbol{x}, u) = \sum_{w \in U - \{u\}} \mathcal{S}(\boldsymbol{x}, u, w) \tag{4}$$

しかしこのような肯定スコアでは、1 対 1 識別結果と総合結果の整合性を常に得ることはできない。それを以下に示す。1 対 1 識別結果と総合結果の整合性がとれている状態とは、$\forall u, v \in U$ において式 (5) となることである。

$$\mathcal{T}(\boldsymbol{x}, u) > \mathcal{T}(\boldsymbol{x}, v) \Leftrightarrow \mathcal{S}(\boldsymbol{x}, u, v) > \mathcal{S}(\boldsymbol{x}, v, u) \tag{5}$$

$u$ と $v$ の総合スコアの差から式 (6) が導かれる。

$$\begin{aligned} &\mathcal{T}(\boldsymbol{x}, u) - \mathcal{T}(\boldsymbol{x}, v) \\ &= \textstyle\sum_{w \in U - \{u,v\}} [\mathcal{S}(\boldsymbol{x}, u, w) - \mathcal{S}(\boldsymbol{x}, v, w)] \\ &+ \mathcal{S}(\boldsymbol{x}, u, v) - \mathcal{S}(\boldsymbol{x}, v, u) \end{aligned} \tag{6}$$

ここで $q \leq \mathcal{S}(\boldsymbol{x}, v, w) \leq p$ のように仮に上限 $p$ と下限 $q$ を設ける。種類 A は他の種類 $w$ とのペアワイズスコアについて、その差が微小値 $\delta > 0$、種類 B は種類 A 以外の種類 $v$ とのペアワイズスコアについて、その差が最大値 $p - q$ とする。これは種類 B に対し種類 A が「最も不利」な状況と言える。これを式 (6) に代入すると、$N$ 種類の多種識別において正解種類 A の総合関数とそうでない種類 B の総合関数との差は式 (7) となる。

$$\mathcal{T}(\boldsymbol{x}, A) - \mathcal{T}(\boldsymbol{x}, B) = (N-2)(q-p) + \delta \tag{7}$$

総合スコアとペアワイズスコアの整合性をとるためには、これが正となる必要があり、$\delta$ に式 (8) に示す下限が生じる。定義により $\delta \leq p - q$ なので、これは $N > 2$ では実現できない。

$$\delta > (N-2)(p-q) \tag{8}$$

例えば、種類 A,B,C において、A 対 C で A に下限 $q$、B 対 C で B に上限 $p$ というスコアの場合、A 対

B での A のスコアと B のスコアとの差は、$p-q$ より大きくならないと総和による総合識別結果が A にならないが、そのようなスコアはつけられない。

一方、否定スコアでは整合性を確保することができる。総合スコアを式 (9) のように定め、否定スコア $T'(\boldsymbol{x},u)$ が最小となる $u$ を $\boldsymbol{x}$ の総合識別結果とする。

$$\mathcal{T}'(\boldsymbol{x},u)=\sum_{w\in U-\{u\}}\mathcal{S}(\boldsymbol{x},w,u) \tag{9}$$

ここで 1 対 1 識別結果と総合結果の整合性がとれている状態は、$\forall u,v\in U$ において式 (10) となることである。

$$\mathcal{T}'(\boldsymbol{x},u)<\mathcal{T}'(\boldsymbol{x},v)\Leftrightarrow\mathcal{S}(\boldsymbol{x},u,v)<\mathcal{S}(\boldsymbol{x},v,u) \tag{10}$$

$u$ と $v$ の総合スコアの差から式 (11) が導かれる。

$$\begin{aligned}&\mathcal{T}'(\boldsymbol{x},u)-\mathcal{T}'(\boldsymbol{x},v)\\&=\textstyle\sum_{w\in U-\{u,v\}}[\mathcal{S}(\boldsymbol{x},w,u)-\mathcal{S}(\boldsymbol{x},w,v)]\\&+\mathcal{S}(\boldsymbol{x},v,u)-\mathcal{S}(\boldsymbol{x},u,v)\end{aligned} \tag{11}$$

否定スコアでの $\delta$ の下限は、式 (12) のように、他種類とのすべてのペアワイズスコアで等しく下限 $q$ をとり、対象となる 2 種類間にのみ微差がある状況で、この条件は $\delta>0$ と同値である。

$$\mathcal{T}'(\boldsymbol{x},A)-\mathcal{T}'(\boldsymbol{x},B)=-\delta<0 \tag{12}$$

以上から、総和による総合スコアにおいて、ペアワイズスコアとの整合性をとるためには、否定スコアを採用するべきであることが分かる。

## 4 まとめと課題

多種識別法のうち、1 対 1 の識別を総合して行う方法を採用する場合、1 対 1 識別の結果のペアワイズ関数とその総合関数にそれぞれ選択肢がある。総合関数としてペアワイズ関数の加算による比較を採用すると、識別すべき種類が識別後に変動した場合に再計算にかかる計算量が少ない。その際、総合結果との整合性を保つには、ペアワイズ関数の総合方法を否定的な総和にしておくべきである。

検討中の二重否定値加算法は、否定的スコアの総和を採用しているので、種類削減に対する再計算量が抑えられ、ペアワイズ関数との整合性もとれている方法と言える。今後はファジイソフトマージン法との性能比較を、複数の対象について数値実験する予定である。

## 参考文献


[1] Kreissel,U.H.-G., “Pairwise Classification and Support Vector Machine,” Scholkopf, Burges and Smola (eds.), Advances in Kernel Methods, pp.255–268. (1999).

[2] 田中, “1 対 1 識別の非帰属度を総合する多種識別,” 情報処理学会研究報告, 2004-MPS-51, pp.33–36, (2004).

[3] Guvenir,H.A., Acar,B., Demiroz,G., and Cekin,A., “A Supervised Machine Learning Algorithm for Arrhythmia Analysis,” Proc. the Computers in Cardiology Conference, (1997).

[4] T. Joachims. Making large-scale SVM learning practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods*, pp.169–184. MIT Press, 1999.

[5] Inoue,T. and Abe,S., “Fuzzy Support Vector Machines for Pattern Classification,” Proc. Int. Joint Conf. on Neural Networks, pp.1449–1454, (2001).