標本データとの非類似度の組を用いた最近隣法の高次元パターンに対する性質について

堀川　洋[†]（正員）

On Properties of Nearest-Neighbor Classifiers for High-Dimensional Patterns in Dissimilarity-Based Classification

Yo HORIKAWA [†], *Member*

[†] 香川大学工学部情報工学科，高松市
Faculty of Engineering, Kagawa University, Takamatsu-shi, 761–0396 Japan

あらまし　標本データとの非類似度の組を用いたパターン認識における最近隣法の性質について調べた．非類似度を要素とするベクトル空間における最近隣法は，もとのパターン空間において二次の識別境界を生成する．計算機実験によって，高次元パターンの識別において，対象とするクラスの分散・共分散行列が異なるとき，非類似度の組を用いた最近隣法がもとのパターン空間における最近隣法に比して良い識別性能をもつことを示した．

キーワード　パターン認識，非類似度，最近隣法，高次元，二次統計量

1. まえがき

近年，標本データとの非類似度（距離）の組（dissimilarity representation）を用いたパターン認識手法（dissimilarity-based classification）が提案され，応用が試みられている [1]．この手法は，標本データあるいはその一部を $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m$ とするとき，対象データ $\boldsymbol{x}$ を各標本データとの非類似度（距離）を要素とする m 次元ベクトル $\boldsymbol{d}(\boldsymbol{x}) = (d(\boldsymbol{x}, \boldsymbol{x}_1), d(\boldsymbol{x}, \boldsymbol{x}_2), \ldots, d(\boldsymbol{x}, \boldsymbol{x}_m))$ で表し，それを用いて識別を行うものである．このように，もとの特徴パターンではなくデータ間の非類似度のみを用いるため，featureless classification [2], [3]，relational discriminant analysis [4] などとも呼ばれている．以下，適切ではないかもしれないが，$\boldsymbol{d}(\boldsymbol{x})$ を「非類似度ベクトル」と呼ぶことにする．非類似度ベクトルを用いる手法は，特徴ベクトルとして表されていないタンパク質のアミノ酸配列などのパターンの識別に適用可能であり，また，非類似度ベクトルの次元は用いる標本データの個数（m）で定めることができるため小標本の高次元パターンにおける次元ののろい [5], [6] を回避することができるといった特徴を有している．更に，ガウスカーネルなどの距離に基づくカーネルを用いたサポートベクトルマシンなどのカーネル関数による認識手法 [6], [7] や複数の識別器の組合せ・統合手法 [8] などとも関連が深い．

文献 [1] では，このような標本データとの非類似度ベクトルを用いた様々な識別手法の比較実験が行われている．その中で，非類似度ベクトル空間における最近隣法の評価も行われているが，もとのパターン空間における最近隣法と比べたとき，その識別性能は対象とするパターンにより様々である．ここでは，特にデータのばらつきの大きさがクラスによって異なるとき，パターンの次元が大きくなるにつれ，非類似度ベクトル空間における最近隣法の識別率が高まることを，計算機実験によって示す．

2. 最近隣法による識別境界

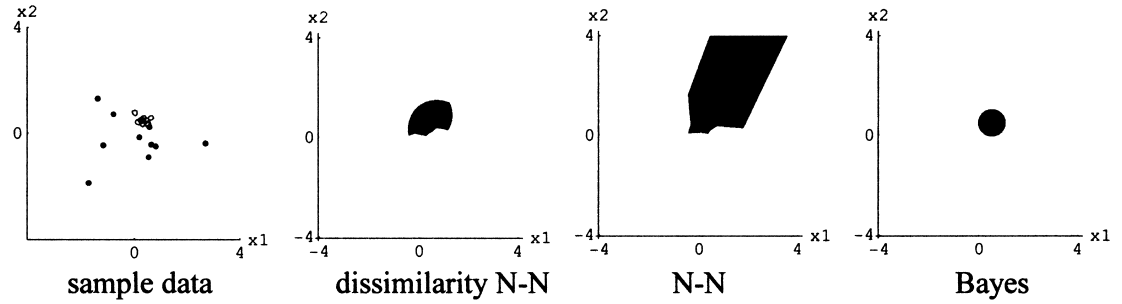
対象とするパターンが n 次元ベクトル $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$ で表されているものとする．そして，パターン同士の距離 $d(\boldsymbol{x}_i, \boldsymbol{x}_j)$，及び，非類似度ベクトル同士の距離 $d(\boldsymbol{d}(\boldsymbol{x}_i), \boldsymbol{d}(\boldsymbol{x}_j))$ は，ともにユークリッド距離を用いるものとする．このとき，非類似度ベクトル空間における最近隣法は，ユークリッド距離の計算における成分の 2 乗の項が消えないため，もとのパターン空間において二次形式の平方根からなる識別境界を構成する．

簡単な例として，以下のような二次元空間内の二つのクラス（C_1，C_2）に属する 4 個の標本データを考える．

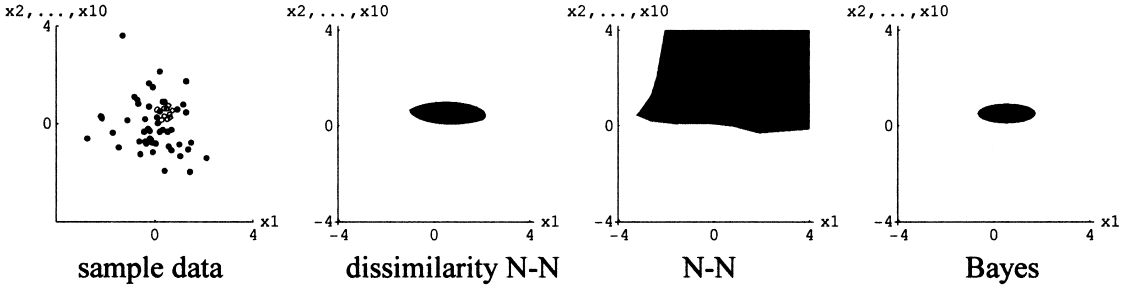
$$\boldsymbol{x}_1 = (-2, -2),\ \boldsymbol{x}_2 = (0, 0) \in C_1$$
$$\boldsymbol{x}_3 = (1, 0),\ \boldsymbol{x}_4 = (1.5, 0.5) \in C_2$$

このとき，対象データ $\boldsymbol{x}$ に対して，非類似度ベクトル $\boldsymbol{d}(\boldsymbol{x}) = (d(\boldsymbol{x}, \boldsymbol{x}_1), d(\boldsymbol{x}, \boldsymbol{x}_2), d(\boldsymbol{x}, \boldsymbol{x}_3), d(\boldsymbol{x}, \boldsymbol{x}_4))$ を用いた最近隣法による識別境界：$\min(d(\boldsymbol{d}(\boldsymbol{x}), \boldsymbol{d}(\boldsymbol{x}_1)), d(\boldsymbol{d}(\boldsymbol{x}), \boldsymbol{d}(\boldsymbol{x}_2))) - \min(d(\boldsymbol{d}(\boldsymbol{x}), \boldsymbol{d}(\boldsymbol{x}_3)), d(\boldsymbol{d}(\boldsymbol{x}), \boldsymbol{d}(\boldsymbol{x}_4))) = 0$ は，もとの二次元パターン空間において図 1 の実線で示した閉曲線になる [9]．なお，通常の最近隣法の識別境界は，当然ながら破線で示した直線である．ここで，識別境界は，クラス内の二つのデータ間の距離が小さいクラス（C_2）を囲う形になっていることが分かる．このことは，一般の場合には，ばらつきの小さい方のクラスのデータを囲む識別境界が生成されることを示唆している．そして，例えば，分散・共分散行列の異なる正規分布に従うデータにおいて，楕円型の二次形式で与えられる最適なベイズ識別関数に近い形状の識別境界が構成されることが期待される．

図 2 (a) は，分散の大きさの異なる正規分布（クラス C_1：$N({}^t(0,0), \boldsymbol{I})$ と C_2：$N({}^t(0.5, 0.5), (0.2)^2\boldsymbol{I})$，



(a) 2-dimensional patterns



(b) 10-dimensional patterns

図 2 二次元パターン空間における識別境界の例 (a) 二次元パターン（クラス C_1：$N({}^t(0,0), \boldsymbol{I})$ と C_2：$N({}^t(0.5,0.5), (0.2)^2\boldsymbol{I})$），標本数：20 個．(b) 十次元パターン（クラス C_1：$N({}^t(0,\ldots,0), \boldsymbol{I})$ と C_2：$N({}^t(0.5,\ldots,0.5), (0.2)^2\boldsymbol{I})$），標本数：100 個．左から右に，標本データ（●：クラス C_1，○：クラス C_2），非類似度の組を用いた最近隣法，通常の最近隣法，ベイズ（二次）識別関数による識別境界．

Fig. 2 Decision boundaries of 2-dimensional patterns of classes C_1: $N({}^t(0,0), \boldsymbol{I})$ and C_2: $N({}^t(0.5,0.5), (0.2)^2\boldsymbol{I})$ with 10 sample data (a) and 10-dimensional patterns of classes C_1: $N({}^t(0,\ldots,0), \boldsymbol{I})$ and C_2: $N({}^t(0.5,\ldots,0.5), (0.2)^2\boldsymbol{I})$ with 100 sample data (b). From left to right: sample data, the dissimilarity N-N, the original N-N and the optimal quadratic classifier.

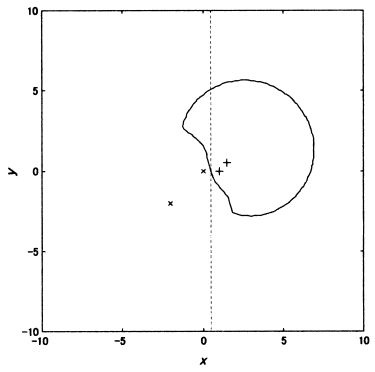


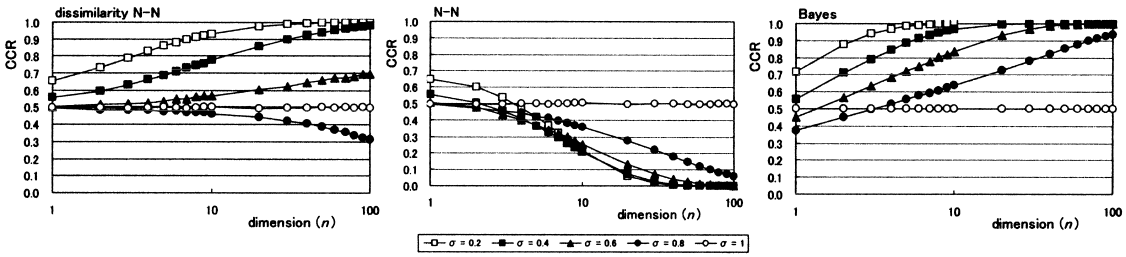
図 1 非類似度の組（実線）と通常（破線）の最近隣法による二次元パターン．$(-2,-2), (0,0) \in C_1$ (×)，$(1,0), (1.5,0.5) \in C_2$ (+) の識別境界

Fig. 1 Nearest-Neighbor decision boundaries of $(-2,-2)$, $(0,0)$ (×) in C_1 and $(1,0), (1.5,0.5)$ (+) in C_2 with dissimilarity representations (solid line) and with original patterns (dashed line).

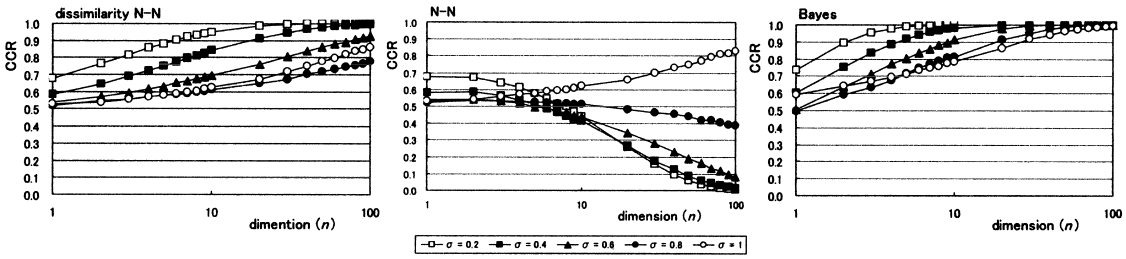
$\boldsymbol{I}$：単位行列）に従う二次元パターンにおける，標本データ数：$m = 20$（両クラスから 10 ずつ）の場合の識別境界の例である．標本数が少ないため通常の最近隣法による識別境界はクラス C_1 の標本がない右上方向に放射状に広がったものになるのに対して，非類似度ベクトルを用いた場合は円状のベイズ識別境界と似た閉じた形状になっている．また，図 2 (b) は，同様な十次元正規パターン（$m = 100$）の場合の，二次元断面（$x_1, x_2 = x_3 = \cdots = x_{10}$）上の識別境界の例である．ここで，標本データは二次元平面への射影像である．パターンの次元が大きくなると，標本数が増えても最近隣法の識別境界は閉じたものになっていないが，非類似度ベクトルを用いると楕円状の境界が得られている．

3. 計算機実験

上記で得られた知見に基づき，次のような n 次元正規分布に従うパターンに対して，非類似度ベクトル空



(a) $\mu = 0$



(b) $\mu = 0.5$

図 3　クラス $C_1 : N({}^t(0,0,\ldots,0), \boldsymbol{I})$ と $C_2 : N({}^t(\mu,\mu,\ldots,\mu), \sigma^2\boldsymbol{I})$ の識別における，非類似度ベクトル空間（左図），もとのパターン空間（中図）における最近隣法，及び，ベイズ（二次）識別関数（右図）を用いた，クラス C_1 のデータに対する正識別率．横軸はデータの次元（n）．(a) $\mu = 0$, (b) $\mu = 0.5$.

Fig. 3　Correct classification rates of nearest-neighbor classifiers for class C_1 data with dissimilarity representations (left) and original patterns (center) as well as the optimal quadratic classifier (right) in the classification of two classes C_1: $N({}^t(0,0,\ldots,0), \boldsymbol{I})$ and C_2: $N({}^t(\mu,\mu,\ldots,\mu), \sigma^2\boldsymbol{I})$ in the n-dimensional space. $\mu = 0$ (a), $\mu = 0.5$ (b).

間における最近隣法ともとのパターン空間における最近隣法の識別性能を，計算機実験によって調べた．

C_1: $N({}^t(0,0,\ldots,0), \boldsymbol{I})$,

C_2: $N({}^t(\mu,\mu,\ldots,\mu), \sigma^2\boldsymbol{I})$

（$\boldsymbol{I}$：n 次元単位行列）

クラス C_1 は n 次元標準正規分布であり，クラス C_2 は各成分の平均値が μ，標準偏差が σ である独立な n 次元正規分布である．標本データ数：$m = 10$（両クラスから 5 ずつ），テストデータ数：1000 として，二つの最近隣法による識別を，異なるランダムデータを用いて 1000 回行った．図 3 に，(a) $\mu = 0$ と (b) $\mu = 0.5$ の場合について，σ を 0.2 から 1.0 まで変えたときの，クラス C_1（分散の小さくない方）に属するパターンの正識別率を示す．横軸の次元 n は対数スケールをとっている．なお，最適なベイズ（二次）識別関数を用いた場合の正識別率を併せて示している．

図から，もとのパターン空間における最近隣法の正識別率は，パターンの次元が大きくなるにつれ低下することが分かる（ただし，平均値が異なり分散の差が小さい場合（$\mu = 0.5$, $\sigma \approx 1.0$ (b)）を除く）．それに対して，非類似度ベクトル空間における最近隣法の場合，次元とともに正識別率は増加する（ただし，平均値が同じで分散の差が小さい場合（$\mu = 0.0$, $\sigma = 0.8, 1.0$ (a)）を除く）．なお，ここでは示していないが，クラス C_2 のパターンに対する両者の正識別率はほとんど同じで，二つのクラスの統計量が同一（$\mu = 0$, $\sigma = 1$）のとき以外は次元とともに増加する．非類似度ベクトル空間における最近隣法による正識別率は，分散の差が大きいとき（$\sigma = 0.2$），最適な二次識別関数による正識別率と等分散を仮定した線形識別関数によるものとの間の値となっている．標本データ数を増やした場合（$m = 100, 1000$）にも，両者とも正識別率は高くなるが，同様な結果が得られる．また，超立方体内の一様分布のような非正規分布に従うパターン，例えば，$C_1 : U(-0.5, 0.5)^n$，$C_2 : U(-0.2, 0.2)^n$ のような場合に対しても同様である．

ただし，このようなもとのパターン空間における最近隣法と非類似度ベクトル空間における最近隣法との違いは，二つのクラスの平均値の差が分散に比して小

さいときにのみ見られるものである．例えば，今回と同じ分布の場合，もとのパターン空間における最近隣法の正識別率は，$\mu = 1.0$ では次元とともに増加するようになり，$\mu = 2.0$ になると両者の正識別率にはほとんど差が見られなくなる．

4. むすび

標本データとの非類似度を要素とするベクトル（非類似度ベクトル）空間における最近隣法，一般には区分線形識別器による識別関数は，もとのパターン空間においては二次の識別関数となる．そのため，クラスによってパターンの分布の二次統計量（分散・共分散行列）が異なる場合に，良い識別性能を有する場合があることを計算機実験によって示した．

クラス間でたとえ平均値に差があっても分散にも少し（1：0.9 程度でも）差があるような特徴が加わった場合，もとのパターン空間における最近隣法はかえって正識別率が低下してしまう．このことは，次元ののろいの一つの表れであり，直観的には，高次元空間における標本データのまばらさによるものと考えられる．例えば，原点を平均値としてばらつきの大きさが異なるクラスの識別には，原点の周りにばらつきの小さなクラスを囲むような識別境界が望ましい．ところが，n 次元空間において原点を $n-1$ 次元超平面で囲う場合，最小では単体のとき $n+1$ 枚の平面で済むが，一般には，$O(2^n)$ 枚の平面が必要となると考えられる．そのため，最近隣法において原点を囲むような識別境界を得るには $O(2^n)$ 個の標本データが必要となるが，それだけの標本データがない場合には，図 2 に示したように原点を囲めない場合が起こってしまうわけである．このことは，言い換えれば，標本データ数に対して次元が大きくなると，ばらつきの大きなクラスの標本データが存在しない方向（側）が増え，その方向の領域は原点近傍のばらつきの小さいクラスの標本データとの方が距離が小さくなってしまうことによる．高次元空間における最近隣法の性質については様々なものが知られているが [10]〜[14]，このような性質はこれまであまり指摘されていないように思われる．

それに対して，非類似度ベクトル空間における最近隣法では，クラス間で平均値が等しく分散のみが異なるような特徴からなるパターンの場合でも，パターンの次元が大きくなるにつれ，正識別率が高くなる．このことは，非類似度ベクトル空間においては，もとのパターンのクラス間の二次統計量の違いによって，次元の増大とともにクラスが線形に分離される傾向があることを示している．例えば，ばらつきの大きさの異なるクラスを考えた場合，標本データ数を一定としたとき，次元が大きくなると，ばらつきの小さなクラス（C_2）の標本データ間の非類似度に比べて，ばらつきの大きなクラス（C_1）の標本データ間の非類似度は平均として相対的に増大する．また，両クラスのデータ間の非類似度もやはり相対的に増大する．そのため，標本データとの非類似度は，クラス C_1 のデータでは，クラス C_1 の標本データとの類似度：大，クラス C_2 の標本データとの類似度：大となるのに対して，クラス C_2 のデータでは，クラス C_1 の標本データとの類似度：大，クラス C_2 の標本データとの類似度：小となる．このように，次元の増大とともにクラス間で非類似度ベクトルの差異が大きくなる傾向があるわけである．このような性質をもつ標本データとの非類似度の組の利用は，次元ののろいに対処する一つの有効なノンパラメトリックな方法であるといえよう．

文　献

[1] E. Pekalska and R.P.W. Duin, “Classifiers for dissimilarity-based pattern recognition,” Proc. 15th Int. Conf. Pattern Recognition (ICPR 2000), vol.2, pp.12–16, 2000.

[2] R.P.W. Duin, D. de Ridder, and D.M.J. Tax, “Experiments with a featureless approach to pattern recognition,” Pattern Recognit. Lett., vol.18, pp.1159–1166, 1997.

[3] V. Mottl, et al., “Featureless pattern recognition in an imaginary Hilbert space and its application to protein fold classification,” in Machine Learning and Data Mining in Pattern Recognition, ed. P. Permer, pp.322–336, Springer-Verlag, Berlin, 2001.

[4] R.P.W. Duin, E. Pekalska, and D. de Ridder, “Relational discriminant analysis,” Pattern Recognit. Lett., vol.20, no.11–13, pp.1175–1181, 1999.

[5] A.K. Jain and B. Chandrasekaran, “Dimensionality and sample size considerations in pattern recognition practice,” in Handbook of Statistics, vol.2, ed. P.R. Krishnaiah and L.N. Kanal, pp.835–855, North-Holland, 1982.

[6] R.P.W. Duin, “Classifiers in almost empty spaces,” Proc. 15th Int. Conf. Pattern Recognition (ICPR 2000), vol.2, pp.1–7, 2000.

[7] E. Pekalska, P. Paclík, and R.P.W. Duin, “A generalized kernel approach to dissimilarity-based classification,” J. Machine Learning Research, vol.2, pp.175–211, 2001.

[8] L.I. Kuncheva, J.C. Bezdek, and R.P.W. Duin, “Decision templates for multiple classifier fusion: An experimental comparison,” Pattern Recognit., vol.34, pp.299–314, 2001.

[9] Y. Horikawa, “Quadratic boundaries in N-N classifiers with dissimilarity-based representations,” Proc. 6th International Conference on Signal Processing (ICSP 2002), pp.1039–1042, 2002.
[10] K. Fukunaga, Introduction to Statistical Pattern Recognition (2nd ed.), Academic Press, San Diego, 1990.
[11] R. Weber, H.-J. Schek, and S. Blott, “A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces,” Proc. 24th Int. Conf. Very Large Database (VLDB’98), pp.194–205, 1998.
[12] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, “When is “nearest neighbor” meaningful?,” Proc. 7th Int. Conf. Database Theory (ICDT’99), pp.217–235, 1999.
[13] A. Hinneburg, C.C. Aggarwal, and D.A. Keim, “What is the nearest neighbor in high dimensional space,” Proc. 26th Int. Conf. Very Large Database (VLDB 2000), pp.505–515, 2000.
[14] 片山紀生，佐藤真一，“類似検索のための索引技術,” 情報処理，vol.42, pp.958–964, 2001.

（平成 16 年 9 月 3 日受付，11 月 12 日再受付）