複数の局所距離計量の学習法とその統合による分類法の提案

情報数理応用研究　　5214C016-1　齋藤洋
指導教員　後藤正幸

A Proposal of Local Distance Metric Learning and Classification Method Based on the Combination of Local Distances

SAITO Hiroshi

1　研究背景・目的

近年，情報技術の発展により大量データの取得と保存が可能になり，データの自動分類の重要性が増している．データの自動分類は，カテゴリ情報が予め付与された学習データから分類規則を構成し，カテゴリ情報の与えられていない新規入力データのカテゴリを推定する問題として定義される．このような自動分類の基本的なアルゴリズムとして，新規入力データのカテゴリを入力データと最も距離の近い k 個の学習データの多数決により決定する k 近傍法がある [1]．この手法は用いる距離尺度に性能が大きく依存することが知られており，適切な距離尺度の導入が重要な課題である．このような点に対し，マハラノビス距離における計量行列を対象問題に適した距離構造となるように学習する，メトリックラーニングと呼ばれる手法が近年数多く提案されている．

本研究では，メトリックラーニングの代表的な手法である Large Margin Nearest Neighbor (以下 LMNN) [2] に着目する．LMNN は任意の学習データとその近傍にある同カテゴリデータとのマハラノビス距離を小さくしながら，異なるカテゴリのデータとのマハラノビス距離が大きくなるよう計量行列を学習する手法である．k 近傍法では，新規入力データのカテゴリを近傍の学習データの多数決により決定するため，カテゴリ境界付近に複数のカテゴリのデータが混在している時に分類が困難になる．この問題に対し，LMNN においてはカテゴリ境界付近のデータを重視した計量行列の学習を行うことにより，分類に有効な距離を得ることを可能にしている．しかし，複数のカテゴリを持つ学習データにおいてはカテゴリ間に統計的特徴の差異が存在すると考えるのが自然であり，学習データ全体に対し大域的な距離計量を学習する LMNN ではその差異によって適切な距離計量を学習ができない可能性がある．

この問題の解決のため，本研究では LMNN がカテゴリ境界付近のデータを重視した計量行列を学習することに着目し，カテゴリごとの統計的特徴をモデル化する複数の局所距離計量を用いた学習法を提案する．さらに未知カテゴリデータの正確な分類を可能とするため，複数の局所距離計量の統合による分類法を提案する．提案手法の有効性をベンチマークデータと人工データを用いた分類実験により実証する．

2　準備

いま，全データ数が N 個である学習データの集合を $\mathcal{X} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$ とする．ここで，$\boldsymbol{x}_i \in \mathbb{R}^d$ は d 次元特徴ベクトルで，$y_i \in \mathcal{C}$ は $\boldsymbol{x}_i$ が属するカテゴリとする．ただし，$\mathcal{C} = \{c_1, c_2, \cdots, c_g\}$ はカテゴリの集合である．このとき，新規の入力データ $\boldsymbol{x}$ を g 個のカテゴリのいずれかに自動分類する問題を考える．いま，計量行列を $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ とすると，任意の 2 点 $\boldsymbol{x}, \tilde{\boldsymbol{x}} \in \mathbb{R}^d$ 間のマハラノビス距離 $d_{\boldsymbol{A}}(\boldsymbol{x}, \tilde{\boldsymbol{x}})$ は式 (1) で定義される．

$$d_{\boldsymbol{A}}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \sqrt{(\boldsymbol{x} - \tilde{\boldsymbol{x}})^{\mathrm{T}} \boldsymbol{A} (\boldsymbol{x} - \tilde{\boldsymbol{x}})}. \tag{1}$$

ただし，T は行列，ベクトルの転置を表す．また，$d_{\boldsymbol{A}}$ が距離尺度となることを保証するため，計量行列 $\boldsymbol{A}$ は半正定値対称行列である必要がある．この計量行列 $\boldsymbol{A}$ を学習データから学習することで，問題に適した距離構造が得られる．メトリックラーニングの代表的な手法としては Information-Theoretic Metric Learning [3] 等があるが，中でも特に分類を目的とした有効な手法として，LMNN が知られている．

3　Large Margin Nearest Neighbor

3.1　概要

LMNN は，異なるカテゴリに所属するデータ間のマハラノビス距離に関する制約のもとで，対象データ $\boldsymbol{x}_i$ と同じカテゴリに所属するデータとのマハラノビス距離を小さくすることにより距離構造を学習するための手法である．しかし，対象データ $\boldsymbol{x}_i$ と同じカテゴリに所属する全てのデータとのマハラノビス距離を計算することは莫大な計算コストを要する．さらに，k 近傍法が近傍データとの関係性のみを考慮することを考えると，全てのデータ間のマハラノビス距離の情報が分類精度向上に有用であるとは考えられない．そのため，LMNN ではターゲットネイバという概念を導入している．ターゲットネイバとは対象データ $\boldsymbol{x}_i$ と同一カテゴリに所属する近傍データのうち上位 γ 個のデータであり，LMNN ではこのターゲットネイバと対象データとのマハラノビス距離のみを小さくすることで実務的に有効な手法を構成している．

さらに LMNN では Support Vector Machine [4] 等の学習に用いられているマージンの概念を用いる．マージンとは異なるカテゴリ同士の距離の最小値であり，LMNN では学習にマージン最大化の制約を加えることで分類精度の向上を図る．具体的にはマージン最大化制約により，複数のカテゴリのデータが混在している識別が難しいカテゴリ境界を重視した学習が行われ，精度が向上される．以上の LMNN による距離学習のイメージを図 1 に示す．

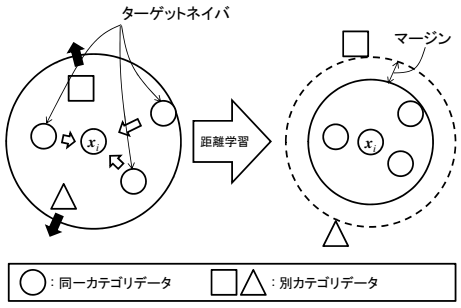

図 1. LMNN による距離学習のイメージ

3.2　最適化問題

前述の特徴より，LMNN の最適化問題は，計量行列の半正定値条件とマージンに関する制約のもとで目的関数を最小化する半正定値計画問題として，以下のように定式化される．

$$\underset{\boldsymbol{A}}{\text{minimize}} \sum_{i,j} \eta_{ij} d_{\boldsymbol{A}}(\boldsymbol{x}_i, \boldsymbol{x}_j) + h \sum_{i,j,l} \eta_{ij}(1-\delta_{il})\xi_{ijl}, \quad (2)$$

$$\text{subject to } \forall i,j,l \in \mathbb{R}[\eta_{ij}=1, \delta_{il}=0]$$

$$d_{\boldsymbol{A}}(\boldsymbol{x}_i, \boldsymbol{x}_l) - d_{\boldsymbol{A}}(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq 1 - \xi_{ijl}, \quad (3)$$

$$\xi_{ijl} \geq 0, \quad (4)$$

$$\boldsymbol{A} \succeq 0. \quad (5)$$

ここで，h は 0 から 1 の範囲の値をとる重みパラメータであり，通常は交差検定によって決定する．また，一般に線形分離不可能なデータに対しては，異なるカテゴリに所属するデータ間の距離全てがマージン制約を満たすような解は求まらない．よって，マージン境界の外側に存在するデータに対し非負のスラック変数 ξ_{ijl} を導入し，マージンの制約を ξ_{ijl} だけ反することを許容することで解の導出を可能にする．式 (2) には，任意のデータ $\boldsymbol{x}_j$ が対象データ $\boldsymbol{x}_i$ のターゲットネイバである場合は 1 を，そうでなければ 0 をとるインジケータ関数 η_{ij} が用いられている．同様に δ_{il} は，任意の 2 データ $\boldsymbol{x}_i$ と $\boldsymbol{x}_l$ に関してカテゴリが一致している場合は 1 を，一致しなければ 0 をとるインジケータ関数である．

このとき，目的関数の第一項は対象データとそのターゲットネイバ間のマハラノビス距離の総和を最小化するための項となる．また第二項はスラック変数の総和を最小化するための項である．式 (3) は異なるカテゴリ同士の距離に関する制約，つまりマージンの制約である．式 (5) は計量行列の半正定値条件であり，これにより最適化問題は半正定値計画問題となる．

3.3 アルゴリズム

LMNN の計量行列 $\boldsymbol{A}$ は，式 (2)–(5) で定義される最適化問題を勾配法をベースとしたアルゴリズムで解くことにより学習される．いま，$\boldsymbol{C}_{ij}$ を $\boldsymbol{C}_{ij} = (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^{\mathrm{T}}$，$\mathrm{tr}(\cdot)$ を行列のトレースとしたとき，任意の 2 点 $\boldsymbol{x}_i, \boldsymbol{x}_j$ 間のマハラノビス距離 (式 (1)) は以下のように変形できる．

$$d_{\boldsymbol{A}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \mathrm{tr}(\boldsymbol{A}\boldsymbol{C}_{ij}). \quad (6)$$

いま，対象データ $\boldsymbol{x}_i$ に対し $\boldsymbol{x}_j$ がターゲットネイバとなる場合の添え字の組み合わせを $i \to j$ と表記する．また，$(i,j,l) \in \mathcal{N}^t$ を勾配法の t 回目の更新において式 (3) の制約条件が等号となるデータ $\boldsymbol{x}_i, \boldsymbol{x}_j, \boldsymbol{x}_l$ の添え字の組み合わせ，すなわち，任意の対象データ $\boldsymbol{x}_i$ に対しそのターゲットネイバ $\boldsymbol{x}_j$ より近くに異なるカテゴリのデータ $\boldsymbol{x}_l$ が存在する時の $\boldsymbol{x}_i, \boldsymbol{x}_j, \boldsymbol{x}_l$ の添え字の組み合わせとする．このとき，式 (3) の制約のもとで式 (2) を最小化する最適化問題の損失関数は以下の式 (7) のように定式化できる．

$$\varepsilon(\boldsymbol{A}) = \sum_{i \to j} \mathrm{tr}(\boldsymbol{A}\boldsymbol{C}_{ij}) + h \sum_{(i,j,l) \in \mathcal{N}^t} (1 + \mathrm{tr}(\boldsymbol{A}\boldsymbol{C}_{ij}) - \mathrm{tr}(\boldsymbol{A}\boldsymbol{C}_{il})). \quad (7)$$

以上より，t 回目の更新における計量行列を $\boldsymbol{A}^t$，勾配法の更新幅を α としたとき，勾配および更新式は以下の式 (8), (9) で求められる．

$$\nabla_{\boldsymbol{A}} \varepsilon(\boldsymbol{A}) = \sum_{i \to j} \boldsymbol{C}_{ij} + h \sum_{(i,j,l) \in \mathcal{N}^t} (\boldsymbol{C}_{ij} - \boldsymbol{C}_{il}), \quad (8)$$

$$\boldsymbol{A}^{t+1} = \boldsymbol{A}^t - \alpha \nabla_{\boldsymbol{A}} \varepsilon(\boldsymbol{A})\big|_{\boldsymbol{A}=\boldsymbol{A}^t}. \quad (9)$$

$d_{\boldsymbol{A}}$ が距離尺度となることを保証するため，計量行列 $\boldsymbol{A}$ は半正定値行列である必要がある．いま， 計量行列の固有値分解を $\boldsymbol{A}^t = \boldsymbol{Q}^t \boldsymbol{\Lambda}^t \boldsymbol{Q}^{t\mathrm{T}}$ とし，各固有値をその対角成分に持つ行列 $\boldsymbol{\Lambda}^t$ に対し負の固有値を 0 に置き換えた行列を $\boldsymbol{\Lambda}^{+t}$ とすると，LMNN においては以下の式 (10) を用いて計量行列の半正定値性を保証する．

$$\boldsymbol{A}^t = \boldsymbol{Q}^t \boldsymbol{\Lambda}^{+t} \boldsymbol{Q}^{t\mathrm{T}}. \quad (10)$$

以上の LMNN の最適化アルゴリズムをまとめると以下のようになる．

[最適化アルゴリズム]

Step1) 更新数 $t=0$ とし，行列 $\boldsymbol{A}^0$ に初期値を与える．

Step2) 式 (8) より勾配を求め，式 (9) により計量行列を更新する．

Step3) $\boldsymbol{A}^t$ に固有値分解を行い，式 (10) により計量行列の半正定値性を保証する．

Step4) 式 (9) における更新の変化量が閾値以下となるまで Step2 と Step3 を繰り返す．

Step5) 計量行列 $\hat{\boldsymbol{A}}$ を出力して終了する． □

3.4 k 近傍法への適用

k 近傍法では，新規入力データのカテゴリをその近傍データ k 件のカテゴリの多数決により決定する．通常の k 近傍法では近傍点の探索にユークリッド距離を用いるのに対し，LMNN において学習された計量行列 $\hat{\boldsymbol{A}}$ により定義されるマハラノビス距離を近傍点の探索に用いることで k 近傍法に LMNN を適用し，分類精度の向上を図る．

4 提案手法

4.1 概要

LMNN では学習データ全体に対し大域的な距離計量を学習するため，カテゴリ間の統計的特徴に差異がある学習データに対しては，それらが相殺された計量行列を学習してしまう問題がある．この問題の解決のため，LMNN の学習におけるマージン最大化の制約は，複数のカテゴリのデータが混在するカテゴリ境界付近のデータを重視した学習を狙いとしている点に着目する．具体的には，学習データに存在する複数のカテゴリ境界ごとに異なる距離計量を学習すれば，カテゴリ間の統計的特徴の差異を考慮した距離計量の学習が可能になると考えられる．

ゆえに本研究では上記の LMNN の問題点を解決するため，任意のカテゴリ境界に着目した複数の局所距離計量の学習を行う手法を提案する．さらにマージンを考慮した学習法である LMNN の特性を用い，局所距離計量の統合を図る．これを k 近傍法の近傍探索に適応することで複数の局所距離計量を用いた分類が可能となる．

4.2 複数の局所距離計量の学習

提案手法では，カテゴリごとの統計的特徴の差異を考慮した距離計量の学習を行うため，まず学習データの持つ全てのカテゴリを任意の 2 つのカテゴリ集合に分割する．このとき，個々の学習データは 2 つのカテゴリ集合のどちらかに属し，g 個のカテゴリを 2 カテゴリ集合に分割する組み合わせは $2^{g-1}-1$ 通り考えられる．カテゴリの分割後，全ての 2 カテゴリ集合の組み合わせに対して独立に LMNN を実行し複数の局所計量行列 $\boldsymbol{A}_n (n = 1, 2, ..., 2^{g-1}-1)$ を学習することにより，$2^{g-1}-1$ 個の局所距離計量 $d_{\boldsymbol{A}_n}$ が定義される．これにより，各学習で全ての学習データを用いながら，異なるカテゴリ間の境界に着目することが可能となる．この任意のカテゴリ集合間の統計的特徴の差異を考慮した学習は，大域的な距離計量を学習する従来手法では表現できない，局所的に適切な分離のための距離尺度を表現可能にする．ゆえに，提案手法で学習される複数の局所距離計量を分類に用いることで，分類精度の向上が期待される．

4.3 局所距離計量の統合による分類

本研究では，学習される複数の局所距離計量を分類に用いるため，複数の局所距離計量の k 近傍法への適用を考える．いま，提案手法における k 近傍法の近傍探索において，新規入力データ $\boldsymbol{x}_p$ と学習データ $\boldsymbol{x}_q$ の距離を測ることを考えると，局所計量行列が $2^{g-1}-1$ 個学習されるため，$\boldsymbol{x}_p$ と $\boldsymbol{x}_q$ に対し $2^{g-1}-1$ 通りのマハラノビス距離 $d_{\boldsymbol{A}_n}(\boldsymbol{x}_p, \boldsymbol{x}_q)$ が定義される．

局所計量行列 $\boldsymbol{A}_n$ はある 2 つのカテゴリ集合の境界を明確にするよう学習されるため，マハラノビス距離 $d_{\boldsymbol{A}_n}(\boldsymbol{x}_p, \boldsymbol{x}_q)$ はその 2 つのカテゴリ集合を分類するのに適した局所的な距離尺度となっている．新規入力データ $\boldsymbol{x}_p$ と学習データ $\boldsymbol{x}_q$ の距離を測る際は，それぞれのデータが存在する領域に適した局所的尺度を用いたいが，最適な局所的尺度をただ一つ決定するのは困難である．そこで，各マハラノビス距離 $d_{\boldsymbol{A}_n}(\boldsymbol{x}_p, \boldsymbol{x}_q)$ に対しデータ $\boldsymbol{x}_p, \boldsymbol{x}_q$ が存在する領域によって変化する重み w_{pq}^n を導入し，各マハラノビス距離を式 (11) のように統合することを考える．

$$D(\boldsymbol{x}_p, \boldsymbol{x}_q) = \sum_{n=1}^{2^{g-1}-1} w_{pq}^n \, d_{\boldsymbol{A}_n}(\boldsymbol{x}_p, \boldsymbol{x}_q). \tag{11}$$

この式により局所距離計量の統合を行うことで任意のデータ間の距離が一意に定義され，k 近傍法への適用が可能になる．重み w_{pq}^n の導入では，任意のデータ間の距離を測る際，全てのマハラノビス距離の学習時に重視したカテゴリ境界からデータが遠いほど小さくなる重みを定義したい．この重みの定義においては，任意のデータとカテゴリ境界との距離を一意に定義する必要がある．

ゆえに本研究では，各局所計量行列 $\boldsymbol{A}_n$ に対しその学習の際に重視したカテゴリ境界の代表点 $\boldsymbol{b}_n$ を定義することを考える．具体的には，カテゴリ境界の代表点 $\boldsymbol{b}_n$ を，計量行列学習の際の LMNN に使われるスラック変数を用いて定義する．式 (2), (3) より，スラック変数の値はその値が大きいほどデータがカテゴリ付近に存在し計量行列学習に強い影響を与えたことを意味する．ゆえに，$\boldsymbol{b}_n$ を以下の式 (12) のようにデータのスラック変数を重みとした重み付き重心として定義することにより，複数のカテゴリのデータが混在するカテゴリ境界の代表点として定義することができる．

$$\boldsymbol{b}_n = \frac{\sum_{(i,j,l)\in\mathcal{N}} \xi_{ijl} \boldsymbol{x}_l}{\sum_{(i,j,l)\in\mathcal{N}} \xi_{ijl}}. \tag{12}$$

式 (12) のようにカテゴリ境界の代表点を定義することにより，任意のデータのカテゴリ境界との距離が一意に定義される．これを用いることで重み w_{pq}^n を以下のように定義する．

$$w_{pq}^n = \frac{1}{d_{\boldsymbol{A}_n}(\boldsymbol{x}_p, \boldsymbol{b}_n) + d_{\boldsymbol{A}_n}(\boldsymbol{x}_q, \boldsymbol{b}_n)}. \tag{13}$$

データ $\boldsymbol{x}_p, \boldsymbol{x}_q$ と代表点 $\boldsymbol{b}_n$ の距離の和の逆数をマハラノビス距離 $d_{\boldsymbol{A}_n}(\boldsymbol{x}_p, \boldsymbol{x}_q)$ の重み w_{pq}^n とすることにより，データがカテゴリ境界から遠いほど小さくなる重みが設定される．これにより，$\boldsymbol{x}_p, \boldsymbol{x}_q$ に近いカテゴリ境界を識別するための計量行列の重みが相対的に大きくなり，$\boldsymbol{x}_p, \boldsymbol{x}_q$ に近い領域の局所的構造を学習した計量行列を重視した分類が可能になる．

4.4 提案アルゴリズム

提案手法のアルゴリズムを以下に示す．

[提案アルゴリズム]

Step1) 学習データのカテゴリを任意に 2 つのカテゴリ集合に分割する．

Step2) 全ての 2 カテゴリ集合の組み合わせに対し LMNN を実行し，局所計量行列 $\boldsymbol{A}_n (n = 1, ..., 2^{g-1}-1)$ を学習する．

Step3) 各局所計量行列に対し，その学習時に重視したカテゴリ境界の代表点 $\boldsymbol{b}_n$ を式 (12) により計算する．

Step4) 全てのテストデータと学習データの組み合わせに対し，式 (13) で計算される重みを用いて式 (11) により近傍探索を行い，k 近傍法による分類を行う．

□

5 実験

提案手法の分類精度向上効果の有効性を検証するため，ベンチマークデータセットを用いた実験を行うとともに，提案手法がカテゴリごとの統計的特徴の差異を反映できていることを検証するため，人工データセットを用いた実験を行った．

5.1 ベンチマークデータセットを用いた実験

5.1.1 実験概要

本節では，ベンチマークデータセットに対して分類実験を行い，提案手法の分類精度の評価を行った結果を示す．提案手法による分類は複数の局所距離計量を用いた k 近傍法であるため，通常のユークリッド距離を用いた k 近傍法と従来の LMNN を用いた k 近傍法を比較対象として分類実験を行った．分類精度を評価する指標として，全テストデータのうち誤分類されたデータの割合である分類誤り率を用いた．

実験には 2010 年の毎日新聞の記事データを用い，8 カテゴリ各 300 件のデータを用意した．各カテゴリ 200 件のデータを学習データとし，残り各 100 件のテストデータを分類したときの分類誤り率を測る実験を 3 回行った．特徴量として単語頻度を用い，頻度 10 以上の単語（名詞，動詞）によって特徴空間を構成する．また，提案手法はカテゴリごとの統計的特徴の差異を考慮した手法であるため，カテゴリの数が少ない場合の従来手法との差を検証する必要がある．よって，8 カテゴリ全てを用いた実験の他に，一部のカテゴリのデータを除いたデータセットを用意した．各データセットの情報を以下の表 1 に示す．

表 1. データセット情報

データセット名	次元数	カテゴリ数	データ数
毎日新聞 4 カテゴリ	1998	4	1200
毎日新聞 5 カテゴリ	2476	5	1500
毎日新聞 6 カテゴリ	2849	6	1800
毎日新聞 8 カテゴリ	2000	8	2400

8 カテゴリ全てを用いたデータセットに含まれる頻度 10 以上の単語数は 3846 と高次元になってしまうため，潜在的意味インデキシングにより次元数を 2000 に削減した．

5.1.2 実験結果と考察

図 2 に分類誤り率の実験結果を示す．

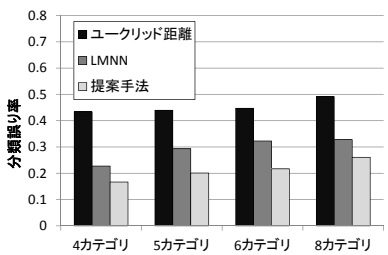

図 2. 分類誤り率の実験結果

提案手法は全てのデータセットにおいて最も低い分類誤り率を示した．この結果より，提案手法のカテゴリごとの統計的特徴の差異を考慮した分類が，従来の LMNN による分類に対して分類精度の面で優れていることが示された．

5.2 人工データセットを用いた実験

5.2.1 実験概要

ベンチマークデータセットを用いた検証実験により，提案手法が分類誤り率の面で優れていることを示すことができた．これに加え以下では，提案手法においてカテゴリごとの差異を反映した局所距離計量の学習が行われていることを検証する．その検証のため，次元数が 2，カテゴリ数が 5 の人工データを作成し，実際に従来手法と提案手法により学習される距離計量により定義される特徴空間にデータを射影しその分布の図示を行う．

人工データの生成では，カテゴリごとに異なる統計的特徴を持つように，5 つのカテゴリに対応した 5 つの異なる平均と分散共分散行列を持つ 2 次元正規分布から各カテゴリ 10500 個のデータを発生させる．図 3 にデータを生成させる 2 次元正規分布の確率密度関数の等高線を示す．

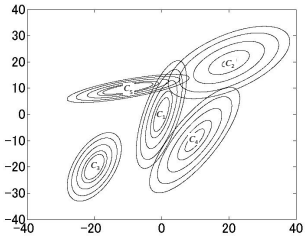

図 3. 確率密度関数の等高線

本節の人工データを用いた実験では，以上のようなカテゴリごとに統計的特徴の差異がある人工データに対し従来の LMNN と提案手法による距離計量の学習を行い，k 近傍法に適用することで提案手法の性質を明らかにする．

5.2.2 実験結果と考察

各カテゴリの人工データ 10500 件のうち，500 件を学習データとし，10000 件をテストデータとした時の各手法の分類誤り率の結果は以下の通りとなった．

表 2. 分類実験の結果

ユークリッド距離	LMNN	提案手法
0.00356	0.00362	0.00286

従来の LMNN により学習した距離計量を用いた k 近傍法の分類誤り率は通常のユークリッド距離を用いた k 近傍法とほぼ同じ結果となった．一方，提案手法の複数の局所距離計量を用いた分類手法の分類誤り率は，他の 2 手法より低い結果となった．

この結果を考察するため，本研究においては従来手法と提案手法により学習される距離計量により定義される特徴空間に各カテゴリの人工データを射影し，その分布を見る．ここで，提案手法では複数の局所距離計量により複数の特徴空間が定義されるため，例として学習データの 5 カテゴリを $\{1,2,3,5\}$ と $\{4\}$ の 2 つのカテゴリ集合に分割し得られた特徴空間にデータを射影する．従来手法と提案手法により得られた分布は以下の通りとなった．

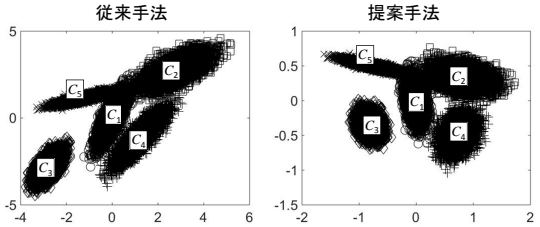

図 4. 射影したデータの分布

図 3 と図 4 より，従来手法の LMNN により学習された距離計量により定義される特徴空間への射影では人工データの分布がほぼ変わらないことが分かった．よって，LMNN の単一の距離構造では学習データの持つカテゴリごとの統計的特徴の差異が反映できず，分類に有効な距離計量が学習できていないことが示された．

一方提案手法においては，局所距離計量の学習によりカテゴリ 4 と他のカテゴリの境界が明確化していることが分かる．よって，提案手法の学習は特定のカテゴリ間の統計的特徴の差異を考慮した学習を可能とすることが示された．加えて，提案手法が全てのカテゴリのテストデータに対して有効性を持つことを検証するため，従来の LMNN と提案手法を用いたそれぞれの分類において，誤分類されたテストデータの件数をカテゴリごとに算出した．この結果を表 3 に示す．

表 3. 各カテゴリの誤分類データ数

	LMNN	提案手法
カテゴリ 1	55	48
カテゴリ 2	10	8
カテゴリ 3	0	0
カテゴリ 4	7	6
カテゴリ 5	109	81
合計	181	143

表 3 より，提案手法は全てのカテゴリのデータに対して従来の LMNN より誤分類数が減少するという結果が得られた．以上より，提案手法がカテゴリごとの統計的特徴の差異を考慮し，分類に有効な距離構造の学習を可能とすることが示された．

6 まとめと今後の課題

本研究では，LMNN の持つカテゴリごとの統計的特徴の差異を考慮できないという問題を解決するため，LMNN のカテゴリ境界付近のデータを重視した計量行列の学習に着目し，任意のカテゴリ集合間の統計的特徴の差異を反映した複数の局所距離計量を学習する手法を提案した．学習した局所距離計量を用いて未知カテゴリデータの正確な分類を行うため，全ての局所距離計量を統合し k 近傍法に適用することによる分類手法を提案した．提案手法の有効性を検証するため，ベンチマークデータセットを用いた分類実験を行いその有効性を示した．加えて，提案手法がカテゴリごとの統計的特徴の差異を反映できていることを検証するため，人工データセットを用いた実験を行いその有効性を示した．

今後の課題として，局所距離計量の統合にカテゴリ境界との距離を用いていることを考慮し，距離計量の学習自体にもよりカテゴリ境界のデータの情報を反映させる方法の検討が挙げられる．さらに，カテゴリ境界が複雑化したときに提案手法の有効性がどのように変化するかを検証することも今後の課題である．

参考文献

[1] T. Cover and P. Hart,“Nearest Neighbor Pattern Classification,” *IEEE Trans. Inform. Theo.*, Vol. IT-13, pp. 21–27, 1967.

[2] K. Weinberger, J. Blitzer and L. Saul,“Distance Metric Learning for Large Margin Nearest Neighbor Classification,” *Proc. NIPS, Eds.*, pp. 1475–1482, 2006.

[3] J. Davis, B. Kulis, P. Jain, S. Sra and I. Dhillon,“Information-Theoretic Metric Learning,” *Proc. 24th International Conference on Machine Learning*, pp. 209–216, 2007.

[4] V. N. Vapnik, *Statistical learning theory*, John Wiley and Sons, New York, 1998.