低次元計量行列の学習とその結合による計量行列学習の計算量削減法

1X10C048-6　齋藤 洋
指導教員　後藤 正幸

1　研究背景・目的

近年，情報技術の発展によりデータの自動分類について多くの研究が行われている．データの自動分類とは，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学習し，カテゴリ情報の与えられていないデータのカテゴリを推定する問題である．分類規則の最も基本的かつ有効な学習法としてデータ間の距離を用いる手法があるが，そのパフォーマンスは用いる距離尺度に大きく依存する．そのため，適切な距離尺度の導入法として，マハラノビス距離を対象に問題に適した距離構造を学習するメトリックラーニングとよばれる手法が提案されている．本研究ではこのうち，半正定値計画問題を勾配法で解くことによりマハラノビス距離における計量行列を学習する Mahalanobis Metric for Clustering (以下 MMC)[1] に着目する．MMC はデータの自動分類に有効な距離尺度の学習を可能にするが，学習時に全学習データ間の距離計算および計量行列の固有値分解を繰り返し行うため，学習データ数の増加や特徴空間の高次元化により計算量が著しく増加してしまうという問題がある．

一方，集合学習の分野では Random Forest と呼ばれる手法が注目されている．これは，少ない変数を選択して学習した複数のモデルを混合することで予測精度を高める手法である．MMC の計算量が学習データ数と特徴空間の次元数に依存する点に着目すれば，Random Forest の考え方を MMC に導入することでその分類精度を劣化させずに計算量削減が可能と考えられる．そこで本研究では，学習データの少数抽出，特徴空間の低次元化を行ったのち MMC の計量行列を学習する操作を繰り返し，得られた複数の計量行列を結合する手法を提案する．提案手法の有効性をベンチマークデータセットを用いた分類実験により検証する．

2　Mahalanobis Metric for Clustering

データを g 個のカテゴリに分類する問題を考える．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，$\boldsymbol{x}_i$ は d 次元特徴ベクトルで，$y_i \in \{1, 2, \cdots, g\}$ は $\boldsymbol{x}_i$ が属するカテゴリとする．このとき，同一カテゴリに所属するデータペア $\{\boldsymbol{x}_i, \boldsymbol{x}_j\}$ の集合を $\mathcal{S}$，別カテゴリに所属するデータペア $\{\boldsymbol{x}_i, \boldsymbol{x}_j\}$ の集合を $\mathcal{D}$ とする．

MMC は，集合 $\mathcal{D}$ に含まれるデータペアのマハラノビス距離の総和を大きくし，集合 $\mathcal{S}$ に含まれるデータペアのマハラノビス距離の総和を小さくする計量行列を学習する．いま，計量行列を $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ 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 $d_{\boldsymbol{A}}(\boldsymbol{x}_i, \boldsymbol{x}_j)$ は式 (1) で定義される．

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

このとき，計量行列 $\boldsymbol{A}$ の学習は以下の最適化問題を解くことにより行われる．

$$\max_{\boldsymbol{A}} \quad g(\boldsymbol{A}) = \sum_{\{\boldsymbol{x}_i, \boldsymbol{x}_j\} \in \mathcal{D}} d_{\boldsymbol{A}}(\boldsymbol{x}_i, \boldsymbol{x}_j), \tag{2}$$

$$\text{s.t.} \quad f(\boldsymbol{A}) = \sum_{\{\boldsymbol{x}_i, \boldsymbol{x}_j\} \in \mathcal{S}} d_{\boldsymbol{A}}(\boldsymbol{x}_i, \boldsymbol{x}_j)^2 \leq 1, \tag{3}$$

$$\boldsymbol{A} \succeq 0. \tag{4}$$

式 (2) は異なるカテゴリに属するデータペアのマハラノビス距離の総和に対する最大化であり，式 (3) は同一カテゴリに属するデータペアのマハラノビス距離の総和を抑えるための制約である．式 (4) は行列 $\boldsymbol{A}$ の半正定値条件を示す．

いま，式 (3), (4) を満たす領域をそれぞれ $C_1 = \{\boldsymbol{A} : \sum_{\{\boldsymbol{x}_i, \boldsymbol{x}_j\} \in \mathcal{S}} d_{\boldsymbol{A}}(\boldsymbol{x}_i, \boldsymbol{x}_j)^2 \leq 1\}$, $C_2 = \{\boldsymbol{A} : \boldsymbol{A} \succeq 0\}$ とし，行列 $\boldsymbol{A}$ の更新幅を決めるパラメータを α としたとき，以下のアルゴリズムで最適解を求める．なお，$\|\cdot\|_F$ は行列のフロベニウスノルム，∇ は関数の勾配を表わす演算子，$\perp$ はベクトルの垂直方向への写像を表す演算子である．

Step1) 行列 $\boldsymbol{A}$ に初期値を与える．

Step2) $\boldsymbol{A}$ が収束するまで式 (5), (6) により $\boldsymbol{A}$ を更新する．

$$\boldsymbol{A} := \arg\min_{\boldsymbol{A}'} \{\|\boldsymbol{A}' - \boldsymbol{A}\|_F^2 : \boldsymbol{A}' \in C_1\}, \tag{5}$$

$$\boldsymbol{A} := \arg\min_{\boldsymbol{A}'} \{\|\boldsymbol{A}' - \boldsymbol{A}\|_F^2 : \boldsymbol{A}' \in C_2\}. \tag{6}$$

Step3) 式 (7) により $\boldsymbol{A}$ を更新する．

$$\boldsymbol{A} := \boldsymbol{A} + \alpha(\nabla_{\boldsymbol{A}} g(\boldsymbol{A}))_{\perp \nabla_{\boldsymbol{A}} f}. \tag{7}$$

$\boldsymbol{A}$ が収束条件を満たしていなければ Step2 に戻る．

□

式 (5) は同一カテゴリに属する学習データ間の距離に関する制約の下での 2 次の最適化問題であり，各学習データ間の距離を計算するため，その計算量は学習データ数の 2 乗に比例する．式 (6) は計量行列の半正定値条件の下での 2 次の最適化問題であり，計量行列の固有値分解を行うためその計算量は特徴空間の次元数の 2 乗から 3 乗に比例する．これらを収束まで繰り返すため，MMC の計算量は学習データ数，特徴空間の次元数の増加に伴い大きく増加してしまう．

3　提案手法

3.1　概要

前節で示したように，MMC の計算量は学習データ数と特徴空間の次元数の増加と共に急速に増大してしまう．そこで本研究では，MMC の分類精度を維持したまま計算量を削減することを目的とし，Random Forest の考え方を援用する．具体的には，学習データの少数抽出，特徴空間の低次元化により作成したサブデータセットから計量行列を学習する操作を繰り返し，得られた複数の低次元計量行列を結合する手法を提案する．提案手法のイメージを図 1 に示す．

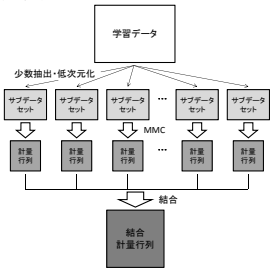

図 1. 提案手法イメージ

3.2　サブデータセットの作成

いま，作成するサブデータセット数を b とし，サブデータセット内のデータ数を $N'(\leq N)$，低次元化後の特徴空間の次元数を $d'(\leq d)$ とする．まず，N 個の学習データの集合 $\mathcal{X}$ から，N' 回ランダムに復元抽出する操作を b 回繰り返し，重複を許した N' 個のデータを持つような b 個のサブデータセット $\mathcal{B}_k (k = 1, 2, ..., b)$ を作成する．次に，$\mathcal{B}_k$ に含まれる N' 個の d 次元データに対し，重複を許さずランダムに d' 個の変数番号 $r_l^{(k)} (l = 1, 2, ..., d', 1 \leq r_l^{(k)} \leq d)$ を選び，その次元のみを取り出すことによって特徴空間の低次元化を行い，得られた N' 個の d' 次元データの集合を ${\mathcal{B}'}_k$ とする．

3.3　低次元計量行列の学習と結合

作成されたそれぞれのサブデータセット ${\mathcal{B}'}_k$ に対して MMC を行い低次元計量行列を学習し，これを $\hat{\boldsymbol{A}}_k$ とする．

ただし，変数 $r_l^{(k)}(l = 1, 2, ..., d')$ を持つベクトルにより構成されるサブデータセット ${\mathcal{B}'}_k$ から学習された低次元計量行列 $\hat{\boldsymbol{A}}_k = [\hat{a}^k_{p,q}] \in \mathbb{R}^{d' \times d'}$ の要素 $\hat{a}^k_{p,q}$ は，変数 $r_p^{(k)}$ と $r_q^{(k)}$ の関係性を表す要素である．

いま，b 個の $\hat{\boldsymbol{A}}_k$ を 1 つに結合することで得られる結合計量行列を $\hat{\boldsymbol{A}} = [\hat{a}_{i,j}] \in \mathbb{R}^{d \times d}$ とすると，この要素 $\hat{a}_{i,j}$ が変数 i と j の関係性を表す要素である必要がある．そこで式 (8) により各低次元計量行列 $\hat{\boldsymbol{A}}_k$ で同じ変数間の関係性を持つ要素を統合し，結合計量行列 $\hat{\boldsymbol{A}}$ の要素とすることでこの性質を満たすようにする．

$$\hat{a}_{i,j} = \frac{1}{b} \sum_{k=1}^{b} \sum_{\{r_p^{(k)}=i, r_q^{(k)}=j\}} \hat{a}^k_{p,q}. \tag{8}$$

ここで低次元計量行列 $\hat{\boldsymbol{A}}_k$ は特徴空間の次元削減時に選択された変数間の要素しか持たないため，擬似的に削減された変数間の要素を全て 0 とすることにより低次元計量行列を元の d 次元に拡張した行列を ${\hat{\boldsymbol{A}}'}_k = [{\hat{a}'}^k_{s,t}] \in \mathbb{R}^{d \times d}$ とする．式 (8) で与えられる結合計量行列 $\hat{\boldsymbol{A}}$ は，この半正定値行列 ${\hat{\boldsymbol{A}}'}_k$ の線形和となる．任意のベクトルを $\boldsymbol{x}$ としたとき，

$$\boldsymbol{x}^{\mathrm{T}} \hat{\boldsymbol{A}} \boldsymbol{x} = \boldsymbol{x}^{\mathrm{T}} \left(\frac{1}{b} \sum_{k=1}^{b} {\hat{\boldsymbol{A}}'}_k \right) \boldsymbol{x} = \frac{1}{b} \sum_{k=1}^{b} \boldsymbol{x}^{\mathrm{T}} {\hat{\boldsymbol{A}}'}_k \boldsymbol{x} \geq 0, \tag{9}$$

が成り立つため，結合計量行列 $\hat{\boldsymbol{A}}$ は半正定値行列であることが保証される．

3.4 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．

Step1) 学習データの集合 $\mathcal{X}$ に対して N' 回ランダムに復元抽出する操作を b 回行い，データ数 N' のサブデータセット $\mathcal{B}_k(k = 1, 2, ..., b)$ を作成する．

Step2) $\mathcal{B}_k$ に含まれる N' 個の d 次元データをランダムな変数選択によって d' 次元に低次元化し，サブデータセット ${\mathcal{B}'}_k$ にする．

Step3) ${\mathcal{B}'}_k$ に対して MMC を行い，低次元計量行列 $\hat{\boldsymbol{A}}_k$ を学習する．

Step4) 得られた b 個の低次元計量行列 $\hat{\boldsymbol{A}}_k$ を，式 (8) により結合計量行列 $\hat{\boldsymbol{A}}$ に結合する．

□

4 実験

4.1 実験条件

提案手法の有効性を示すため，ベンチマークデータセットに対して分類実験を行い，提案手法の計算時間及び分類誤り率の評価を行った．分類誤り率は式 (10) により計算する．

$$\text{分類誤り率} = 1 - \frac{\text{正しく分類されたテストデータ数}}{\text{テストデータ数}}. \tag{10}$$

実験では，UCI 機械学習レポジトリのベンチマークデータセット 3 種類 (Bal, Ionosphere, Musk) を用いた．また予備実験の結果よりサブデータセットのデータ数 N'，次元数 d'，作成数 b を表 1 の通り設定した．

表 1. データセットの概要とパラメータ

データセット名	次元数	カテゴリ数	学習データ数	テストデータ数	N'	d'	b
Bal	4	3	465	160	50	2	100
Ionosphere	34	2	264	87	50	6	200
Musk	166	2	4948	1650	100	10	500

比較手法として全ての学習データから直接計量行列を学習する MMC を用いた．なお，提案手法の計算時間については全処理の合計計算時間の加え，並列処理した場合を想定し，一回の計量行列の学習に要した最遅時間を評価する．

4.2 実験結果と考察

表 2 に実験に要した計算時間を示し，図 2 に実験の分類誤り率を示す．

表 2. 計算時間の実験結果 (sec)

	従来手法	提案手法 (全処理)	提案手法 (並列処理)
Bal	8.99	18.00	0.33
Ionosphere	50.30	103.62	5.47
Musk	167096.07	1382.02	14.67

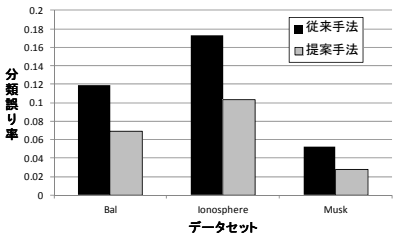

図 2. 分類誤り率の実験結果

提案手法の計算時間は従来手法に比べ，並列処理した場合は全てのデータセットにおいて低下し，並列処理しなかった場合は Musk においてのみ低下した．MMC の計算量は，学習データ数に対して 2 乗オーダ，特徴空間の次元数に対して 2 乗から 3 乗オーダである．本実験に用いたデータセットは学習データ数 N に比べ特徴空間の次元数 d が大きくないため，提案手法を用いることによる計算量削減効果は主として学習データ数の削減による影響である．データ数 N があまり大きくない Bal と Ionosphere においては，データ数の削減による計算量削減効果は大きくなく，この計算を b 回繰り返すことによる計算量増加分が削減分を上回り，従来手法よりも提案手法の全体計算量が上がってしまっている．しかしその場合においても，並列計算が可能であれば大幅に計算量の削減が可能である．次元数とデータ数が共に大きい Musk においては，直列計算であっても計算時間が 1/100 以上に低減されており，並列計算に至っては 3/10000 程度にまで計算時間が低減できる．本提案手法は，データ数 N と次元数 d が非常に大きい問題において著しい効果が期待できる．

一方，提案手法の分類誤り率は全てのデータセットにおいて従来手法より低下した．MMC は学習するパラメータの数が特徴空間の次元数の 2 乗に比例するため，自由度が高く過学習を起こしやすい．しかし，提案手法においてはデータの少数抽出と特徴空間の低次元化によりパラメータ数が減らされるため過学習が抑えられる．また，提案手法は抽出した一部の学習データからの計量行列学習を繰り返すため，学習データに少数の外れ値が含まれていた場合も，学習データの抽出時に外れ値が抽出される回数は少なくなる．そのため，提案手法により学習される結合計量行列は外れ値の影響を受けにくい．実験においてはこれらの効果により，全てのデータセットにおいて分類誤り率が低下したと考えられる．これらの結果から，計算時間と分類誤り率の両面での提案手法の有効性が示された．

5 まとめと今後の課題

本研究では MMC の計算量を分類精度を維持したまま削減することを目的として，学習データを少数抽出しかつ特徴空間を低次元化したのち計量行列を学習する操作を繰り返し，得られた複数の低次元計量行列を結合する方法を提案した．ベンチマークデータセットに対する分類実験の結果から，計算時間と分類誤り率の両面での提案手法の有効性を示した．今後の課題として，サブデータセットのデータ数と次元数の自動決定アルゴリズムの検討が挙げられる．

参考文献

[1] E. P. Xing, A. Y. Ng, M. I. Jordan and S. Russell, "Distance Metric Learning with application to clustering with side-information," *Advances in Neural Information Processing Systems 15*, pp.505-512, 2003.