Large Margin Nearest Neighbor の分類精度向上のための学習データ選択法

1X09C116-6　山崎 史博
指導教員　後藤 正幸

1　研究背景・目的

近年の情報技術の発展に伴い，大規模データからの知識発見手法が数多く提案されている．このうち，本研究ではベクトル空間モデルを用いた多次元データの分類問題に着目する．分類問題とは，カテゴリ情報が予め与えられた学習データから，カテゴリの与えられていない新たな入力データを正しいカテゴリに分類する問題である．ベクトル空間モデルを用いた分類では，データ間の関係性を表す距離尺度を用いて分類を行うため，適切な尺度の導入が極めて重要である．距離尺度には様々なものが存在するが，学習データの統計的特徴を学習し，分類に適した距離構造を得る手法も提案されており，メトリックラーニングと呼ばれ注目を集めている．

本研究では，メトリックラーニングの手法の中でも SVM 等で用いられるマージンの考え方を援用した Large Margin Nearest Neighbor(以下 LMNN) [1] に注目する．LMNN は，対象データとその近傍データの所属カテゴリの関係性を考慮し，対象データと同一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し，別カテゴリに所属するデータとのマハラノビス距離をマージンを考慮して大きくするような計量行列を学習する手法である．LMNN では，識別が比較的難しいカテゴリの境界付近のデータを重視し，カテゴリの識別に寄与する計量行列を学習する点がポイントとなっている．しかし，従来の LMNN では，全学習データが対象データとなるため，境界付近のデータが相対的に重視されずに計量行列が推定されてしまう場合がある．

そこで，本研究では，カテゴリの境界付近に存在するデータを選択し，それらを用いて LMNN の学習を行う手法を提案する．もし予めカテゴリの境界付近のデータを選択することができれば，これらのデータを学習対象とすることで，異なるカテゴリに所属するデータ同士の距離を大きくする距離構造を効率的に学習でき，かつ分類精度を向上させられる可能性がある．本提案手法の有効性を分類実験により示す．

2　準備

合計 N 個ある学習データ集合を $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$，離散カテゴリ集合を $\mathcal{C} = \{c_1, c_2, \cdots, c_G\}$ とする．$\boldsymbol{x}_i$ は d 次元の実数値ベクトルで，$y_i \in \mathcal{C}$ は $\boldsymbol{x}_i$ に付与されているカテゴリとする．全学習データ N 個のうちカテゴリ $c_g (g = 1, 2, \cdots, G)$ に所属する学習データ数を N_g とする．

メトリックラーニングは，学習データ間の統計的特徴を考慮した計量行列を学習することが目的である．計量行列 $\boldsymbol{M}$ を $d \times d$ の行列 $\boldsymbol{M} \in \mathbb{R}^{d \times d}$ とすると，データ $\boldsymbol{x}_i, \boldsymbol{x}_j$ 間のマハラノビス距離は式 (1) で定義される．

$$\mathrm{Dist}_{\boldsymbol{M}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sqrt{(\boldsymbol{x}_i - \boldsymbol{x}_j)^{\mathrm{T}} \boldsymbol{M} (\boldsymbol{x}_i - \boldsymbol{x}_j)} \tag{1}$$

ただし，T はベクトルの転置を表す．

3　Large Margin Nearest Neighbor

LMNN は SVM 等で用いられるマージンの考えを援用したメトリックラーニングの手法である．この手法は，対象データ $\boldsymbol{x}_i$ と同一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し，別カテゴリに所属するデータとのマハラノビス距離を大きくすることで，データを識別しやすくする．しかし，対象データ $\boldsymbol{x}_i$ と同一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と，計算コストが膨大となってしまう．そこで LMNN ではターゲットネイバという概念を導入する．ターゲットネイバは，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，これらと対象データとのマハラノビス距離を小さくする．上記に加え，データを識別するためにマージンという概念を導入する．マージンとは，異なるカテゴリに所属するデータ同士を離すための単位であり，識別しづらいカテゴリの境界付近のデータを分離し，識別を容易にする．

以上の下で，計量行列 $\boldsymbol{M}$ を次の最適化問題で学習する．

$$\min_{\mathbf{M}} \left\{ \sum_{ij} \eta_{ij} \mathrm{Dist}_{\boldsymbol{M}}(\boldsymbol{x}_i, \boldsymbol{x}_j) + h \sum_{ijl} \eta_{ij} (1 - \delta_{il}) \xi_{ijl} \right\} \tag{2}$$

subject to

$$\mathrm{Dist}_{\boldsymbol{M}}(\boldsymbol{x}_i, \boldsymbol{x}_l) - \mathrm{Dist}_{\boldsymbol{M}}(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq 1 - \xi_{ijl} \tag{3}$$

$$\xi_{ijl} \geq 0 \tag{4}$$

$$\boldsymbol{M} \succeq 0 \tag{5}$$

式 (2) の η_{ij} は，$\boldsymbol{x}_j$ 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である．同様に δ_{il} は，y_i と y_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．また，h は 0 から 1 の値をとる重みパラメータとし，式 (5) は $\boldsymbol{M}$ が半正定値行列である条件を示す．

式 (2) の第 1 項は対象データ $\boldsymbol{x}_i$ と，γ 個のターゲットネイバとのマハラノビス距離の和を表す．第 2 項は，同一カテゴリとの距離よりも，別カテゴリとの距離が小さい場合に正の値をとるため，ペナルティ項と解釈できる．

図 1 に計量行列 $\boldsymbol{M}$ の持つイメージを示す．ここで実線の円は，対象データ $\boldsymbol{x}_i$ を中心としたとき，最もユークリッド距離が大きいターゲットネイバとの距離半径を示し，点線の円は実線の円に対してマージンを加えた距離半径を示す．計量行列 $\boldsymbol{M}$ を用いると，対象データ $\boldsymbol{x}_i$ とターゲットネイバとのマハラノビス距離が小さくなり，対象データ $\boldsymbol{x}_i$ と別カテゴリデータとのマハラノビス距離は，点線の円の距離半径よりも大きくなる．これにより，カテゴリ間の識別を容易にする距離構造が得られる．

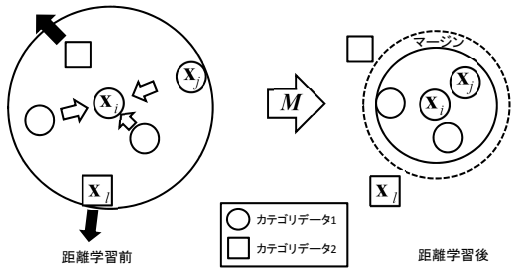

図 1. 計量行列 M による距離学習 ($\gamma = 3$ の例)

4　提案手法

4.1　背景

従来手法では，計量行列 $\boldsymbol{M}$ を学習するとき，全学習データを用いている．しかし，対象データ $\boldsymbol{x}_i$ の近傍に別カテゴリのデータが存在しない場合，式 (2) の第 2 項が 0 となる．このようなデータが多い場合，式 (2) の第 1 項が重視されすぎてしまい，LMNN の特徴が生かせない可能性がある．LMNN の基本的アイデアは，前述の通りマージンを設けるような計量行列を学習することで，境界付近の分類しづらいデータを分類しやすくする点にある．従って，別カテゴリ

データが近傍に多く存在すると考えられる境界付近のデータ (以下 Boundary Data) を選択し，それを学習に用いることができれば，分類精度が向上する可能性がある．

以上の議論から，本研究ではカテゴリの境界付近に存在すると考えられる Boundary Data を選択するアルゴリズムを構築する．そこで,「境界付近では，別カテゴリのデータが相対的に多く存在する」という仮説から，ある対象データの近傍データを探索し，その中で対象データとは別カテゴリに所属するデータが一定数以上存在するときに対象データを Boundary Data とする方法を考える．

4.2 学習データ選択アルゴリズムを用いた LMNN

Boundary Data を選択する際，各カテゴリに所属するデータ数は異なるため，カテゴリ毎のデータ数の差を考慮する必要がある．データ選択アルゴリズムでは，カテゴリ c_g に属するある対象データからユークリッド距離が小さい順にデータを p_g 件探索する．すなわち，p_g は対象とするカテゴリ c_g 毎に決められる値で，カテゴリに含まれるデータ数 N_g に比例して値を決定する．探索された p_g 個のデータのうち，対象データと別カテゴリに所属するデータの数が q_g 件以上となるような対象データを Boundary Data とする．

p_g, q_g はそれぞれ式 (6), (7) で与えられる．

$$p_g = \left\lceil \frac{N - \alpha N_g - 1}{\alpha(N - N_g) + 1} \right\rceil \tag{6}$$

$$q_g = \lceil \beta p_g \rceil \tag{7}$$

ただし，$\lceil \cdot \rceil$ は天井関数であり，ある実数に対してそれ以上の最小の整数を表す．式 (6), (7) における α, β は，Boundary Data か否かの判断基準を変化させるパラメータであり，[0,1] の任意の実数をとる．α は対象データの周辺のデータをどの程度探索するかを定めるパラメータであり，0 にすると $N-1$ 件の全データを探索し，1 にすると最近傍のデータのみを探索する．β は周辺から探索したデータ p_g 件のうち，別クラスの割合をどの程度許容するかのパラメータである．α, β 共に 0 とした場合にはデータ選択を行わず，従来の LMNN と同一となる．これらのパラメータを変動させることで，対象データの近傍データの探索数 p_g，Boundary Data とするための別カテゴリデータ数 q_g を決定する．提案アルゴリズムを以下に示す．

Step1) 全学習データのうちの 1 つを対象データとする．

Step2) 対象データからユークリッド距離が小さい順にデータを p_g 件探索する．

Step3) 探索した p_g 件のデータのうち，対象データと別カテゴリに所属するデータが q_g 件以上の場合，対象データを Boundary Data とみなす．

Step4) Step1〜Step3 を全学習データに対して行い，Boundary Data とみなされたデータ群を学習データとする．

Step5) Step4 で選択した学習データを入力データとして LMNN を実行する．

□

5 実験

5.1 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，提案手法の分類精度の評価を行った．実験の評価指標としては式 (8) を用いた．

$$\text{分類誤り率} = \frac{1 - \text{正しく分類されたテストデータ数}}{\text{テストデータの総数}} \tag{8}$$

実験では，公開データセット UCI 機械学習レポジトリのベンチマークデータセット 4 種類 (Iris, Bal, Digits, Isolet) を用いた．データセットの概要を表 1 に示す．

表 1. データセットの種類

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
Iris	4	3	112	38
Bal	4	3	465	160
Digits	50	10	7291	2007
Isolet	100	26	6238	1559

それぞれのパラメータは予備実験から，分類誤り率が良かった $\alpha = 0.3$, $\beta = 0.1$, $\gamma = 3$, $h = 0.5$ を用いた．比較手法としてデータ選択を行わない LMNN を用いた．

5.2 実験結果と考察

図 2 に実験の結果を示す．全ての実験において，提案手法は従来手法と比較し同程度以上の分類精度が得られている．また，表 2 に実験に要した計算時間を示す．全ての実験において，提案手法は従来手法より計算時間が短縮できたことがわかる．これは学習データ数が減少した効果が大きかったためと考えられる．Boundary Data 数の一例としては，Bal データセットのカテゴリのデータ数は 37, 210, 218 個に対し，Boundary Data 数は 37, 58, 61 個だった．

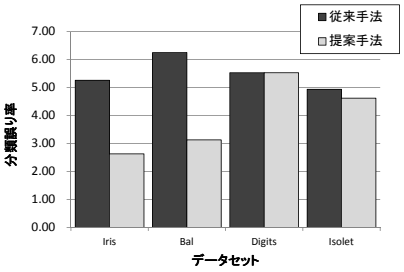

図 2. 分類精度の実験結果

表 2. 計算時間の実験結果 (sec)

	従来手法	提案手法
Iris	1.45	1.28
Bal	3.50	1.86
Digits	45.25	13.87
Isolet	85.20	70.82

本実験の結果から，カテゴリの境界付近のデータを用いることが分類精度の向上に寄与したことが分かる．近年，SVM や RVM のような分類器のアイデアにもあるように，カテゴリの境界を重視して学習する方法は，分類という目的には相性が良いことが本研究の成果からも分かる．また，元々学習データ数が少ないデータセットである Iris, Bal では，選択アルゴリズムで選択した学習データ数があまり変わらなかったため計算時間の短縮には結びつかなかった．一方，学習データが多い Digits, Isolet では，選択した学習データ数が減少したため，計算時間の低減に影響したと考えられる．

6 まとめと今後の課題

本研究では，LMNN の対象データと周辺データの関係性を考慮した計量行列を学習するという特性に着目し，Boundary Data を定義し，その選択方法を提案した．分類実験の結果，分類精度，計算時間の面で提案手法の有効性を示した．

今後の課題として，パラメータ α と β 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] K. Weinberger, J. Blitzer, and L. Saul, “Distance Metric Learning for Large Margin Nearest Neighbor Classification,” *Proc. NIPS, Eds. Cambridge, MA: MIT Press*, pp.1475–1482, 2006.