G-015

中間点情報を用いたサポートベクターマシンに関する研究
A Study on Support Vector Machines using Midpoint Data

山下 真吾† 田村 宏樹† 淡野 公一†
Shingo Yamashita Hiroki Tamura Koichi Tanno

1. はじめに

サポートベクターマシン（以降 SVM と呼ぶ）は，高次元特徴空間でのマージン最大化をすることで高い識別率を得ることができる識別器の 1 つである．しかし，入力空間における識別境界線の情報は考慮されていない．そのため，入力空間では歪な境界線になる場合がある．それを改善する方法として，ソフトマージン SVM などがあるが，新たなパラメータの設定が問題となってしまう．

入力空間での識別境界線に“偏り”が生じるように SVM が識別境界線を決定した場合の改善手法として，著者らは中間点検証法を提案している[1]．中間点検証法の利点は，SVM 構築後に適用可能であること，新たなパラメータの設定が必要ないこと，計算量が少ないなどの利点があり，ベンチマーク問題において従来のハードマージン SVM より識別率を改善することに成功している．

しかし，文献[1]では大きく 2 つ検証されていない問題があった。まず 1 つ目として、クラス間に重なりがある部分での中間点とそうでない中間点を考慮していない問題があった．2 つ目として，適用したのがハードマージンの SVM だけであり，ソフトマージン SVM への適用結果についての検証がなされていなかった．

そこで本発表では，クラス間に重なりがある部分での中間点，そうでない中間点の 2 タイプに中間点の情報を分け，それぞれの場合での識別結果について検証を行う．また，ソフトマージン SVM へ適用したときの結果についても検証を行う．

2. 提案手法

中間点検証法では，まず中間点データを異なるクラスの既知の訓練データから作成する．次に中間点を生成するフローを示す．また，図 1 に中間点生成の概念図を示す．

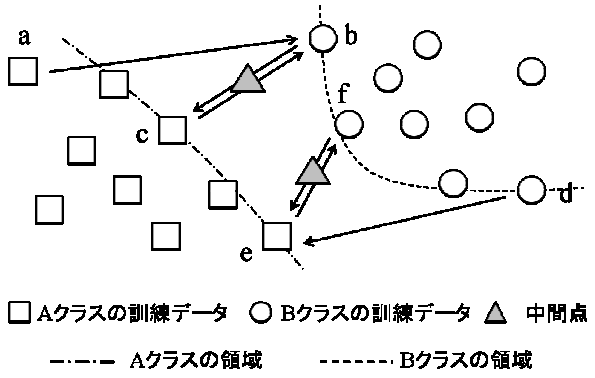

図 1　中間点生成の概念図

STEP1	: A クラスのデータから任意の点 a を選択.
STEP2	: 点 a から最も近い点を B クラスのデータから選択(図 1 中の点 b).
STEP3	: STEP2 で得られたデータから最も近い点を A クラスのデータから選択(図 1 中の c).
STEP4	: 点 a と STEP3 で得られたデータが一致したとき STEP5 へ．不一致の場合 STEP2 へ.
STEP5	: STEP2 と STEP3 で得られたデータの中間点を生成する.

以上の処理をすべての訓練データに適用し，中間点データを生成する．

次に，生成された中間点データを，クラス間に重なりがある部分での中間点と，そうでない中間点を簡易的に分類する方法について述べる．図 2 にその手順のイメージ図を示す．

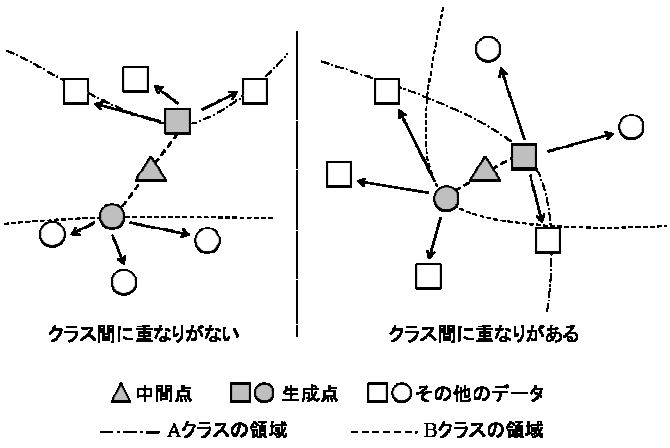

図 2　中間点分類のイメージ図

分類には，中間点を生成する際に選択された各クラスの代表点を用いて行う．以下，これを生成点と呼ぶ．

まず，各クラスの生成点について，生成点との距離が短い点を上位から各々3 点選択する．そして，双方の生成点について，選択された 3 点の内 2 点が生成点と同クラスであった場合，それらの生成点によって生成された中間点はクラス間に重なりがない部分に位置していると見なす．逆に，それ以外の場合には，クラス間に重なりがある部分に位置していると見なした．

最後に，中間点データを用いた中間点検証法を SVM へ適用する手順について述べる．中間点データは理想的な場合では識別境界線に近くなると仮定し，中間点データに対する SVM による理想の出力値は，0 に近くなるようにする．その方法として SVM の出力状態である式(1)，(2)

† 宮崎大学工学部　Factory of Engineering, University of Miyazaki

に式(3)のように式(5)で算出する B を加算することを行う．式(5)で算出する B は中間点データに対する SVM の出力の平均値に符号を反転させたものである．つまり B を加算することで，中間点データに対する SVM の出力の平均を 0 にする．これは SVM の出力を中間点データに対する偏りがないように線形的に調整したことになる．

$$g(x)=\sum_{i=1}^{N} w_i K(x_i,x)+b \tag{1}$$

$$K(x_i,x)=\exp\left(\frac{-\|x_i-x\|^2}{2\delta^2}\right) \tag{2}$$

$$h(x)=g(x)+B \tag{3}$$

$$O=sign(h(x)) \tag{4}$$

$$B=-\frac{1}{M}\sum_{m=1}^{M} g(x_m) \tag{5}$$

式(5)の M は中間点 x_m のデータ数である．今回の検証では中間点を分類し，それぞれの種類のみの中間点を使用する．よってその場合，M は選択した種類の中間点の個数となる．また，N は識別関数の決定に関与するデータ(サポートベクトル)の数である．

ここで，提案手法の処理を以下に示す．

STEP1	：SVM を既知の訓練データにより構成
STEP2	：中間点を生成
STEP3	：中間点の分類
STEP4	：中間点データに対する SVM の出力値を算出
STEP5	：式(5)に従って B の値を計算する
STEP6	：式(3)，式(4)に従い提案手法の出力を計算

3. 計算機実験

本発表では，Ionosphere data，Pima-indians-diabetes data，Wisconsin breast cancer data，Sonar data，Liver-disorders data の計 5 つのベンチマーク問題[2]にて，提案手法の有効性を計算機実験によって検証した．

表 1 はハードマージン SVM に提案手法を適用した結果，表 2 はソフトマージン SVM の 1 種である C-SVC[3]に提案手法を適用した結果である．ここで，SVM および C-SVC はそれぞれ提案手法を適用していない通常の状態での結果である．そして，その他は提案手法を適用したものであり，それぞれ，+ALL は式(5)においてすべての中間点の出力を使用したときの結果，+Aonly はクラス間に重なりがないと分類された中間点のみを使用したときの結果，+Bonly はクラス間に重なりがあると分類された中間点のみを使用したときの結果である．

これらの結果から，まずハードマージン SVM では，Iono を除く問題について，通常の手法に比べ提案手法の方が良い識別率を示した．また，全ての中間点を使用したものに対して，中間点を分類して適用したものを比較すると，+Aonly は同等，+Bonly は若干改悪した結果となっている．

表 1　ハードマージン SVM への適用結果(%)

	SVM	+ALL	+Aonly	+Bonly
Iono	84.1	68.9	52.3	71.5
Pima	78.6	80.7	78.6	79.2
Wis	82.4	98.5	98.5	92.4
Sonar	88.5	**91.3**	**91.3**	89.4
Liver	63.5	76.5	**77.4**	63.5

表 2　C-SVC への適用結果(%)

	C-SVC	+ALL	+Aonly	+Bonly
Iono	**98.0**	72.8	43.7	87.4
Pima	**82.3**	80.7	81.8	81.3
Wis	**98.8**	98.5	98.5	97.9
Sonar	89.4	89.4	90.4	89.4
Liver	74.8	71.3	74.8	67.8

C-SVC では，提案手法の結果が同等あるいは改悪された結果となった．ただし Sonar については，+Aonly にて改善が見られる．また，提案手法の中では Iono を除く結果では，+Aonly が最も良い結果となっている．

4. おわりに

本発表では，SVM の中間点検証法を中間点を分類して適用した場合の検証，およびソフトマージン SVM への適用についての検証を計算機実験で行った．

計算機実験では，ハードマージン SVM に提案手法を適用した場合，クラス間に重なりがないと分類された中間点データのみを使用した結果がすべての中間点を使用した結果とほぼ同等の識別率となった．また，ソフトマージン SVM に適用した場合，提案手法は識別率の改善を行うことができなかったが，提案手法の中での比較では，クラス間に重なりがない部分の中間点のみの使用が最も良い結果となった．この結果から，クラス間に重なりがないと分類される中間点のみを使用することの有効性が確認できた．

実験結果より，提案手法である中間点検証法は，識別境界線が入力空間において偏りが生じるような問題に対しては有効であると考えられる。しかし，クラス間が重なっていることで識別境界線が上手く決定できないような問題に対しては，提案手法よりソフトマージン SVM の方が適している場合が多いと考えられる．

今後の課題としては，1)中間点を分類する手法をより正確にすること，2)中間点を用いる有効性を再検討すること，3)中間点を用いてより適切に SVM の改善ができる手法の提案が挙げられる．

参考文献

[1]Hiroki TAMURA, Koichi TANNO, “Midpoint-Validation Method for Support Vector Machine Classification”, IEICE Trans. on Information and Systems, Vol.E91-D No.7 pp.2095-2098,Jul. 2008.

[2]C.L. Blake, C.J Merz, UCI Repository of machine learning databases,http://www.ics.uci.edu/~mlearn/MLRepository.html, University of California, Department of Information and Computer Science, Irvine, CA, 1998.

[3]LIBSVM: http://www.csie.ntu.edu.tw/~cjlin/libsvm/ Chang and Lin, 2001.