ISS-P-319

SVM における超平面への距離を用いたクラス分類

濱﨑　邦秀　　三浦　孝夫
法政大学理工学部創生科学科

1. 目的

SVM は二値分類において、高い分類精度を期待できる分類器である。クラス数が 3 以上であるような分類、つまり多クラス分類に SVM を用いる場合は、本来の SVM とは異なった手法が必要となる。多クラス分類における SVM の手法として代表的なものに、one-versus-rest 法やペアワイズ法などが知られているが、どちらの手法にしても、所属クラスを一意的に決定することができない領域(以下、決定不能領域)が存在するなどの問題がある。
本研究では、超平面への距離を用いて、多クラス分類を行う。ベースラインをペアワイズ法として比較実験を行い、分類精度を評価する。

2. 二値分類器における SVM と多クラスへの拡張

SVM は、入力ベクトル空間上に、超平面と呼ばれる線形関数を構築し、超平面によって入力ベクトル空間を分離することで二値分類を行う。入力ベクトル空間が線形な超平面で分離不可能な場合、カーネル法と呼ばれる手法を用いる。各テスト事例において、識別関数を計算し、関数値の正負によって分類する。この識別関数の値を関数距離と呼び、超平面への距離を表す値である。関数距離はいうなれば、所属クラスへの確信度ともいえる。本研究では、この関数距離を利用して、多クラス分類を行う。

ペアワイズ法では、各クラス対ごとに、超平面を構築する。n 個のクラスがある場合、$n(n-1)/2$個の超平面を構築する。分類段階では、テスト事例を全ての各超平面で分類し、多数決を採り最も多くの票を得たクラスを所属クラスとする。このような多数決方式では、決定不能領域が生じる問題(図 2)がある。決定不能領域について、以下のような例を挙げる。クラス A,B,C に対して、図 2 に示すように各クラス対の超平面を構築したとする。ここで、あるテスト事例を分類する。3 つの超平面で囲まれた青い領域においては、3 つのクラスどれにも該当してしまう。

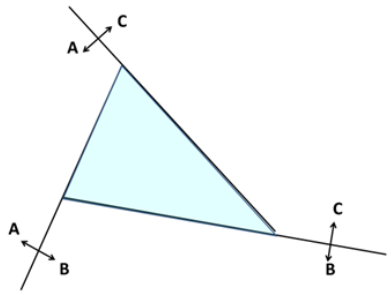

図 1:多クラス分類における決定不能領域

3. 本研究の手法と利点欠点

本研究の手法はペアワイズ法をベースに、多数決における投票値に関数距離を用いる点と関数距離のトップ k 個のみ多数決を行う点である。この手法では、トップ k 個内の関数距離の分散が、大きくなる場合がある。対策として、分散値の閾値σを設ける。さらに、閾値を超えたケースに対して、トップ k の平均値より下回った関数距離を多数決から除外する。

4. 実験環境と評価方法

実験に用いたコーパス読売新聞 2007 年版の新聞記事 1 年分である。各記事ごとに出現頻度名詞ベクトルを作成し、実験データとして用いる。各記事ごとに付与された記事分類をクラスラベルとして用いることにする。合計クラスラベル数 20 件、データ件数 9550 件とする。関数距離の分散値の閾値$\sigma = 2000$、関数距離の大きいトップ$k = 5$にて多数決を行う。カーネル法で用いるカーネル関数は RBF カーネルとする。評価方法は 100 分割の交差検定を行い、ベースラインにはペアワイズ法を用い、本研究の手法と比較実験を行う。

5. 実験結果と評価

以下のような実験結果が得られた。本研究の手法では、分類精度 73.33%、同様にペアワイズ法では、66.86%。本研究の手法は、実験データに対して、ベースラインより約 6.47%分類精度が高いことが分かる。テスト事例に使われたデータを解析した結果、誤分類されたテスト事例は、トップ 5 の関数距離の分散値が使用したテスト事例の分散値と比べて、比較的小さい分散値であるケースが多いことが分かる。(表 1)

誤分類されたテスト事例のヒストグラム		正分類されたテスト事例のヒストグラム	
分散値のデータ区間	頻度	分散値のデータ区間	頻度
0.0001	0	0.0001	0
0.0005	0	0.0005	0
0.001	0	0.001	0
0.005	0	0.005	0
0.01	0	0.01	0
0.05	4	0.05	0
0.1	1	0.1	0
0.5	5	0.5	0
1	1	1	0
5	8	5	0
10	0	10	1
50	5	50	13
100	2	100	13
500	0	500	24
1000	0	1000	12
5000	0	5000	7

表 1: 正誤分類されたテスト事例のヒストグラム

6. 結論

本研究では、ペアワイズ法をベースに、関数距離を多クラス分類に用いた。それによって、不決定領域の問題を解消し、SVM における多クラス分類の精度を約 6.47%向上させた。今度の改善点として、関数距離の分散値が小さい場合に、誤分類されるテスト事例を減らす問題が残された。

参考文献

[1] Nello Cristianini, Jahn Shawe-Taylor 著 大北剛　訳：「サポートベクターマシン入門」
[2] 高村大也　著：「言語処理のための機械学習入門」