カーネル密度推定法を用いた非線形判別手法の提案

山本けい子 函館工業高等専門学校　寒河江雅彦 金沢大学経済学類

1．はじめに

パターン認識における非線形判別問題は，サポートベクターマシン（SVM）[1]の出現により，大きな進展をとげている．我々は，カーネル密度推定法（以降，KDE と略す）を用いた非線形判別手法を提案する．カーネル密度推定法は，データの分布を仮定しないノンパラメトリックな手法であり，複雑な現象を確率的かつ柔軟にとらえて表現できることから様々な応用が期待される．本稿では，数字判別問題に対する 2 つの状況下での SVM との比較を通して KDE の特性を検証する．

2．カーネル密度推定法を用いた非線形判別器

多変量カーネル密度推定法(1)は，データ点にカーネル関数と呼ばれる基底関数を配置し，それらを領域内で足し合わせることによって滑らかな推定量を得るものである．

$$\hat{f}(\mathbf{x};\mathbf{H}) = \frac{1}{n}\sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{x}-\mathbf{X}_i) \qquad (1)$$

ただし，$\mathbf{X}_i \quad (i=1,\cdots,n)$ はデータベクトル，K_H はバンド幅行列 H をもつカーネル関数である．

多変量データに対する KDE は，構築の困難さや推定精度の問題から直接的な適用は難しい．そこで，1 変量 KDE の積で表現するプロダクト(積型)カーネル推定法によって近似する．たとえば，2 変量プロダクト KDE は(2)式で定義される：

$$\hat{f}(\mathbf{z};h_x,h_y) = \frac{1}{n}\sum_{i,j} K_{h_x}(x-X_i)K_{h_y}(y-Y_j) \qquad (2)$$

以下に，(2)に示した KDE を利用した非線形判別の流れを示す．

1) クラス別正解確率分布の作成
学習用データに対し，クラスごとに KDE を適用し，クラス別確率分布の推定を行う．

2) 評価確率分布の作成
判別対象の評価用データに対し，KDE を適用し，評価確率分布の推定を行う．

3) 正解分布と評価分布間での類似量の算出
判別対象データに基づく評価分布と各クラスの正解分布間の類似量を平均積分二乗誤差によって算出する．

4) クラスの判別
評価分布と最も類似する正解分布のクラスへ判別する．

クラス別正解分布を作成しておくことで，実際の判別時には，判別対象データの分布推定と類似量の算出のみでクラス判別を行うことができるため，効率的かつ汎用的な手法といえる．

3．手書き数字判別問題

手書き数字判別問題に対して，カーネル密度推定法を用いた非線形判別器（数字判別器）を構築し，その判別性能を SVM と比較し，評価する．実装には，オープンソースの統計解析システム R[3]を用いた．

3.1 実験用データ

U.S. Postal Service (USPS) ZIP code datasets の手書き数字を対象とした．USPS データは 1 つの数字を 16×16 ピクセルのグレースケール値(-1 から 1 の範囲の値）によって表す正規化されたデータである．データ数を表 1 に記載する．

表 1　USPS データセット

データ	0	1	2	3	4
学習用	1194	1005	731	658	652
評価用	359	264	198	166	200
データ	5	6	7	8	9
学習用	556	664	645	542	644
評価用	160	170	147	166	177

グレースケール値は，0 から 1 の間の値に変換し確率値として使用した．

3.2 ビン化カーネル密度推定法

通常のカーネル密度推定法は，データ点に対して基底となるカーネル関数を用いる．カウント(頻度)データの場合は，カウント値に比例する重みつきのカーネル関数を用いたビン化カーネル推定法が使用される．USPS データの性質から，16×16 ピクセルにグレースケール値の高さを持つデータとみなし，ビン化カーネル密度推定法を適用した．数字判別におけるビン化カーネル推定法は(3)式で定義される．

$$\hat{f}(\mathbf{z};h_x,h_y) = \sum_{i,j} q_{ij}K_{h_x}(x-i\delta)K_{h_y}(y-j\delta) \qquad (3)$$

ただし，$q_{ij} \equiv g_{ij}/\sum_{i,j} g_{ij}$ であり，$g_{ij}, i\delta, j\delta$ は，それぞれ (i,j) ビンにおけるグレースケール値とビン中点である．

3.3 正解確率分布の作成

表 1 に示した学習用データを数字ごとに各ビンで平均し，ビン化カーネル密度推定を行う．推定

点は各ビンの中点を加えた 33×33 点，カーネル関数は，2 変量正規プロダクトカーネル，バンド幅は，各次元とも 1 に設定した．グレースケール値が標本数ではないため，理論的な議論は省く．推定した各数字の正解確率分布を図 1 に示す．

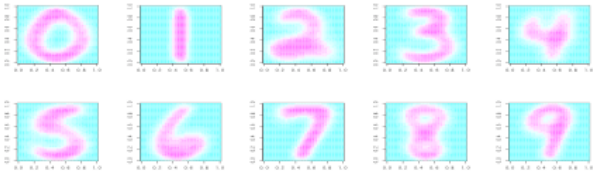

図 1　推定した正解確率分布

3.4 評価用データを用いた数字判別

3.3 で作成した"0"から"9"までの正解確率分布と評価用データで作成した評価確率分布との類似度を(4)式で与えられる平均積分二乗誤差(MISE)を用いて算出する．

$$\mathrm{MISE}[\hat{f}(\cdot;h_x,h_y)] = \int \mathrm{E}\left[\left\{\hat{f}(\mathbf{z};h_x,h_y)-f(\mathbf{z})\right\}^2\right] d\mathbf{z} \qquad (4)$$

MISE の最も小さかった(最も類似した)クラスを評価データのクラスとして分類する．

4．結果

4.1 判別性能の検証

手書き数字判別問題に対する各数字の判別性能を表 2 に示す．

表 2 に示すように，カーネル推定法を用いた数字判別器は，判別する評価用データの数字によって，判別率にばらつきがあるものの，平均して 0.82 であった．一方，SVM はすべての数字において，判別率 0.9 以上，平均して 0.94 という高性能な結果であった．

表 2 評価用データの判別率

判別率	カーネル	SVM
0	0.87	0.98
1	0.96	0.96
2	0.75	0.91
3	0.81	0.91
4	0.77	0.93
5	0.77	0.92
6	0.82	0.95
7	0.82	0.93
8	0.76	0.9
9	0.8	0.97
計	0.82	0.94

4.2 頑健性の検証

カーネル密度推定法による非線形判別器の頑健性を調べるため，ノイズを付加した手書き数字判別に関する実験を行った．

手書き数字データにノイズを加え，判別性能を評価する．16×16 ピクセルのうち，ノイズの大きさを正規乱数(平均 0，標準偏差 0.5)で与え，ノイズを付加するピクセルの割合を増やしたときの判別性能のグラフを図 2 に示す．

図 2 より，SVM の判別性能はノイズの量とともに低下するが，カーネル推定法による数字判別器の性能の低下は少ないことがわかる．

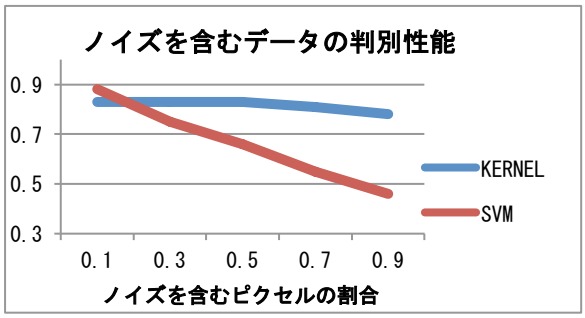

図 2　ノイズを含むデータの判別性能

4.3 クラスタリングによる判別性能の改善

学習用データに応じて数字毎に複数個の正解確率分布を作成し，判別性能の改善を試みた．学習用データを数字ごとに k-means 法によってクラスタリングし，クラスターごとに正解密度分布を推定する．評価用データでの判別の際には，各クラスターが所属していた数字へ判別する．クラスター数を変化させたときの判別性能を図 3 に示す．

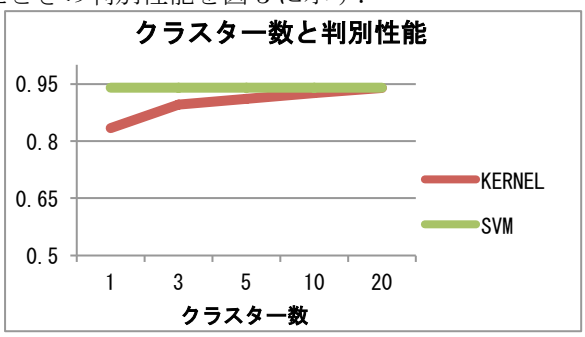

図 3　クラスター数と判別性能

図 3 より，クラスター数を増やすと，SVM に近い判別性能が得られていることがわかる．

5．考察

カーネル密度推定法を用いたパターン認識手法について，手書き数字判別の例を取り上げて検討した．カーネル密度推定法による数字判別器は，SVM の判別性能よりも劣っていたが，ノイズのあるデータに対しては，SVM よりも性能低下の影響は受けにくく，頑健性がみられた．判別する数字によって性能にばらつきが見られるため，クラスタリングによって複数個の正解密度分布を準備することで，SVM とかわらない判別性能が得られた．実用化に向け，さらなる改善が期待できる．

参考文献

[1] V.Vapnik, "The Nature of Statistical Learning Theory", Springer, (1995).
[2] M. P. Wand, M. C. Jones "Kernel Smoothing", Monographs on Statistics and Applied Probability, Chapman & Hall, (1995)
[3]R Development Core Team "R: A language and environment for statistical computing", R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.