サポートベクター回帰のモデル選択

Model Selection for Support Vector Regression

杉山 将

Masashi Sugiyama

東京工業大学 計算工学専攻

Department of Computer Science, Tokyo Institute of Technology

1 まえがき

Vapnik によって提案された Support Vector Machine (SVM) [5] は，高い汎化能力が得られる学習機械の一つとして近年非常に注目されている．しかし，SVM の能力を最大限に引き出すためには，正則化定数などのハイパーパラメータの値を適切に設定しなければならない．一般には，Cross-Validation (CV) によって予測誤差を推定し，その推定した予測誤差を最小にするようにハイパーパラメータの値を決定する．訓練データがたくさん入手できる場合は，CV によって予測誤差を正確に推定することができるため，適切にハイパーパラメータの値を決定することができた．しかし訓練データが少ない場合は，必ずしも正確に予測誤差を推定できるとは限らないため，更なる改良が必要とされていた．

Subspace Information Criterion (SIC) [4, 3] は，有限の訓練データに対して関数近似誤差の不偏推定量となることが保証される規準である．従って，SIC は訓練データが少ない状況でも優れた性能を発揮する．しかし，これまで SIC は線形な学習法にしか適用できなかったため，非線形な学習法を用いている SVM には適用できなかった．そこで本論文では，Bootstrap 法 [1] に基づいた SIC の拡張規準 Bootstrap Approximation SIC (BASIC) を提案する．そして，訓練データ数が少ない場合でも，BASIC によって SVM の正則化定数の値が適切に決定できることを計算機実験によって示す．

2 学習問題の定式化

教師付き学習は，関数近似の問題と捉えることができる．そこで，n 入力 1 出力の未知の実数値関数 $f(\boldsymbol{x})$ を推定する関数近似の問題を議論することにする．$f(\boldsymbol{x})$ の定義域を $\mathcal{D}$ $(\subset \mathbb{R}^n)$ で表す．学習に用いる訓練データを $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{\ell}$ で表す．ℓ は訓練データ数，$\boldsymbol{x}_i$ $(\in \mathcal{D})$ は標本点，y_i $(\in \mathbb{R})$ は標本値を表す．標本値 y_i には，平均 0，分散 σ^2 の分布に独立に従う未知の加法性雑音 ξ_i が加わっている場合を考える．即ち，$y_i = f(\boldsymbol{x}_i) + \xi_i$．また，学習したい未知の関数 $f(\boldsymbol{x})$ がある再生核ヒルベルト空間 $\mathcal{H}$ に含まれる場合を考える．$\mathcal{H}$ の再生核を $K(\boldsymbol{x}, \boldsymbol{x}')$ で表す．学習結果の関数を $\hat{f}(\boldsymbol{x})$ で表し，SVM によって求める [5]：

$$\hat{f}(\boldsymbol{x}) = \sum_{i=1}^{\ell} \alpha_i K(\boldsymbol{x}, \boldsymbol{x}_i) \tag{1}$$

$\{\alpha_i\}_{i=1}^{\ell}$ は訓練データから推定するパラメータであり，ある二次計画問題の解として得られる [5]．

関数近似問題の目的は，未知の関数 $f(\boldsymbol{x})$ にできるだけ近い学習結果の関数 $\hat{f}(\boldsymbol{x})$ を求めることである．本論文では，$\hat{f}(\boldsymbol{x})$ の「よさ」を次の関数近似誤差で測る．

$$\|\hat{f} - f\|_{\mathcal{H}}^2, \tag{2}$$

$\|\cdot\|_{\mathcal{H}}$ は $\mathcal{H}$ のノルムである．式 (2) は，関数近似の分野でよく用いられる誤差尺度である．

統計的学習理論では，雑音 $\{\xi_i\}_{i=1}^{\ell}$ だけでなく標本点 $\{\boldsymbol{x}_i\}_{i=1}^{\ell}$ も確率変数とみなす場合があるが，本論文では，$\{\boldsymbol{x}_i\}_{i=1}^{\ell}$ は任意に固定し，$\{\xi_i\}_{i=1}^{\ell}$ だけが確率変数である場合を考える．

3 Bootstrap Approximation SIC (BASIC)

$\{K(\boldsymbol{x}, \boldsymbol{x}_i)\}_{i=1}^{\ell}$ で張られる $\mathcal{H}$ の部分空間を $\mathcal{S}$ で表し，$f(\boldsymbol{x})$ の $\mathcal{S}$ への正射影を $f_K(\boldsymbol{x})$ で表す．$\mathcal{H}$ の内積を $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ で表せば，関数近似誤差 (2) は次のように変形できる ($\hat{f} \in \mathcal{S}$ に注意)．

$$\|\hat{f}-f\|_{\mathcal{H}}^2 = \|\hat{f}\|_{\mathcal{H}}^2 - 2\langle\hat{f}, f_K\rangle_{\mathcal{H}} + \|f_K\|_{\mathcal{H}}^2 + \|f-f_K\|_{\mathcal{H}}^2 \tag{3}$$

式 (3) 右辺の第 3 項と第 4 項は $\hat{f}$ に依らないので，無視することにする．式 (3) 右辺の第 1 項と第 2 項が関数近似誤差の本質的な部分であり，それらをまとめて J で表すことにする：

$$J = \|\hat{f}\|_{\mathcal{H}}^2 - 2\langle\hat{f}, f_K\rangle_{\mathcal{H}} \tag{4}$$

更に計算していくと，J は次式で表すことができる．

$$J = \boldsymbol{\alpha}^\top \boldsymbol{K} \boldsymbol{\alpha} - 2\boldsymbol{y}^\top \boldsymbol{\alpha} + 2\boldsymbol{\xi}^\top \boldsymbol{\alpha} \tag{5}$$

ここで，$^\top$ は転置，$\boldsymbol{K}$ は (i,j) 要素が $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ の ℓ 次元行列，$\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_\ell)^\top$，$\boldsymbol{y} = (y_1, y_2, \ldots, y_\ell)^\top$，$\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_\ell)^\top$ である．

式 (5) 右辺の第 1 項と第 2 項は与えられた訓練データから計算することができるが，第 3 項は雑音ベクトル $\boldsymbol{\xi}$ が未知のため直接計算できない．そこで，第 3 項を推定値で置き換えることにする．

まず，パラメータベクトル $\boldsymbol{\alpha}$ を線形推定する場合を考える．即ち，$\boldsymbol{\xi}$ と独立な ℓ 次元行列 $\boldsymbol{X}$ を用いて，

$$\boldsymbol{\alpha} = \boldsymbol{X}\boldsymbol{y} \tag{6}$$

で与えられる場合を考える．$\mathrm{E}_{\boldsymbol{\xi}}$ で雑音に関する平均を表し，$\mathrm{tr}\,(\cdot)$ で行列のトレースを表すことにすれば，

$\mathrm{E}_{\boldsymbol{\xi}}\boldsymbol{\xi}^\top\boldsymbol{\alpha} = \sigma^2\mathrm{tr}\,(\boldsymbol{X})$ が成り立つ．これより，次のような J の推定量が得られる．

$$\mathrm{SIC} = \boldsymbol{\alpha}^\top\boldsymbol{K}\boldsymbol{\alpha} - 2\boldsymbol{y}^\top\boldsymbol{\alpha} + 2\sigma^2\mathrm{tr}\,(\boldsymbol{X}) \qquad (7)$$

式 (7) は，Subspace Information Criterion (SIC) [4, 3] と本質的に等価である．SIC は $\mathrm{E}_{\boldsymbol{\xi}}\mathrm{SIC} = \mathrm{E}_{\boldsymbol{\xi}}J$ を満たす．元々の SIC の導出では，式 (6) が最初から仮定されていた．しかし上述の議論から，条件 (6) は式 (5) 右辺の第 3 項を近似するためだけに必要な条件であったことが分かる．従って，この第 3 項さえ推定する事ができれば，任意の非線形な学習法に SIC を適用することができる．

本論文では，式 (5) 右辺の第 3 項を Bootstrap 法 [1] によって近似する方法を提案する．その規準を Bootstrap Approximation SIC (BASIC) と呼ぶ：

$$\mathrm{BASIC} = \boldsymbol{\alpha}^\top\boldsymbol{K}\boldsymbol{\alpha} - 2\boldsymbol{y}^\top\boldsymbol{\alpha} + 2\mathrm{E}_{\boldsymbol{\xi}^b}{\boldsymbol{\xi}^b}^\top\boldsymbol{\alpha}^b \qquad (8)$$

第 3 項の Bootstrap 近似 $2\mathrm{E}_{\boldsymbol{\xi}^b}{\boldsymbol{\xi}^b}^\top\boldsymbol{\alpha}^b$ は，次の手順で計算する．

1. 訓練データ $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{\ell}$ を用いて通常通り SVM の学習を行ない，$\boldsymbol{\alpha}$ を求める．
2. 雑音の推定値 $\hat{\xi}_i = y_i - \sum_{j=1}^{\ell}\alpha_j K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ を求める $(i = 1, 2, \ldots, \ell)$.
3. $\{\hat{\xi}_i\}_{i=1}^{\ell}$ から復元抽出を行ない，Bootstrap 複製 $\{\xi_i^b\}_{i=1}^{\ell}$ を生成する．
4. $\{(\boldsymbol{x}_i, y_i^b) \mid y_i^b = \sum_{j=1}^{\ell}\alpha_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) + \xi_i^b\}_{i=1}^{\ell}$ を訓練データとして SVM の学習を行ない，$\boldsymbol{\alpha}^b$ を求める．
5. 3–4 を何度も繰り返し，${\boldsymbol{\xi}^b}^\top\boldsymbol{\alpha}^b$ の平均値を求める．

ここで注意すべき事は，標本点 $\{\boldsymbol{x}_i\}_{i=1}^{\ell}$ は固定したまま，雑音の推定値 $\{\hat{\xi}_i\}_{i=1}^{\ell}$ だけを複製することである．これは，本論文では固定した標本点 $\{\boldsymbol{x}_i\}_{i=1}^{\ell}$ を考えていることに起因している．

4 計算機実験

本節では，SVM の正則化定数 C の決定問題に BASIC を適用し，その性能を実験的に評価する．

入力次元 n は 1 次元とし，分散 1 のガウシアン RKHS を $\mathcal{H}$ として用いる．$f(x)$ は，$\mathrm{sinc}(x)$ とほとんど同じ形になるように係数を調整した 100 個のガウシアンの線形結合とする (図 1 左上参照)．標本点 $\{x_i\}_{i=1}^{\ell}$ は，$(-\pi, \pi)$ 上の一様分布から独立に採取する．雑音 $\{\xi_i\}_{i=1}^{\ell}$ は，平均 0，分散 0.01 の正規分布から独立に採取する．また，SVM の汎化性能を，$(-\pi, \pi)$ 上の一様分布から独立に採取した 1000 点のテスト入力に対する出力の平均二乗誤差で評価する．SVM の学習には，SVM^{light} を用いる [2]．SVM のパラメータ ϵ は 0.01 に設定する．SVM の正則化パラメータ C は $\{10^{-2}, 10^{-1.5}, 10^{-1}, \ldots, 10^{0.5}\}$ から選ぶ．BASIC の Bootstrap 近似の繰り返し回数は 100 回とする．以上の条件で，訓練データ数 $\ell = 10, 25, 50, 100, 200$ の 5 種類に対して，それぞれ 100 回実験を行なう．

図 1 に実験結果を示す．縦軸はテスト誤差を表し，100 回の分布を Box Plot 表記 (下から 5%，25%，50%，75%，95%) で示した．OPT は最適な正則化定数 C を選んだ場合を表し，CV は C の決定に Leave-One-Out Cross-Validation を用いた場合を表す．この結果から，BASIC の性能は CV と同等か少し上回っていることが分かる．特に，訓練データ数が少ない場合には非常に良い結果が得られている．

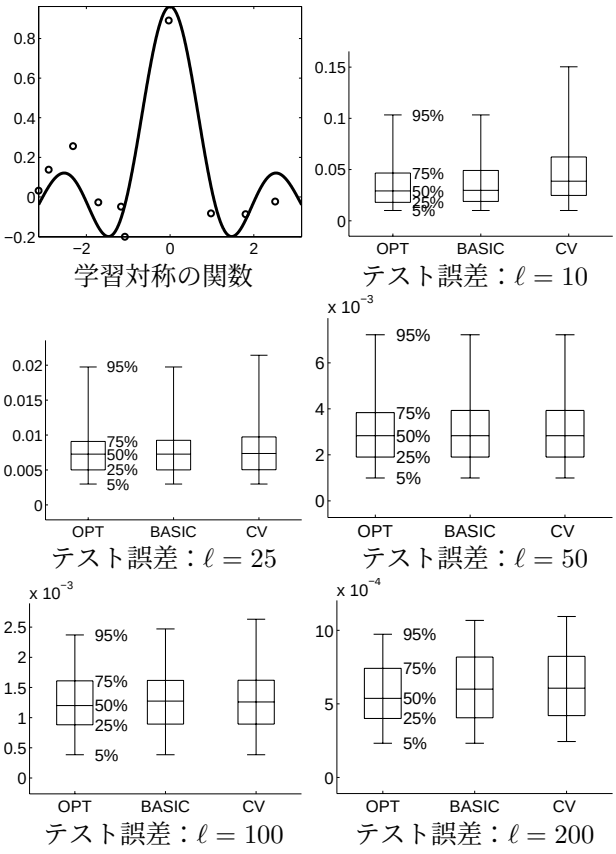

図 1　実験結果

今後は，BASIC の有効性を理論的に調べるとともに，様々な実データに BASIC を適用していく予定である．

謝辞：本研究は，科学研究費補助金 14780262，14380158 の援助により行なわれた．

参考文献

[1] B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

[2] T. Joachims. Making large-scale SVM learning practical. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods — Support Vector Learning*, pages 169–184. The MIT Press, Cambridge, MA, 1999.

[3] M. Sugiyama and K.-R. Müller. Selecting ridge parameters in infinite dimensional hypothesis spaces. In *Proceedings of International Conference on Artificial Neural Networks*, 2002. to appear.

[4] M. Sugiyama and H. Ogawa. Subspace information criterion for model selection. *Neural Computation*, 13(8):1863–1889, 2001.

[5] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, Berlin, 1995.