2L5-1in2

ES-SVM の解空間の解析

Analysis of the solution-space structure by the Exhaustive Search with Support Vector Machine

川端大貴[*1] 市川寛子[*2] 永田賢二[*1*3] 永福智志[*4] 田村了以[*4] 岡田真人[*1]
Daiki Kawabata Hiroko Ichikawa Kenji Nagata Satoshi Eifuku Ryoi Tamura Masato Okada

[*1]東京大学 大学院新領域創成科学研究科
Graduate School of Frontier Sciences, The University of Tokyo

[*2]東京理科大学 理工学部教養
Liberal Arts, Faculty of Science and Technology, Tokyo University of Science

[*3]JST さきがけ研究員
PRESTO Researcher, Japan Science and Technology Agency (JST)

[*4]富山大学 医学薬学研究部
Graduate School of Medicine and Pharmaceutical Sciences, University of Toyama

Feature selection in dealing with high-dimensional data is getting difficult, when features are more getting to be increase. To solve the feature selection problem, sparse estimation is often used such as Lasso. However, it have not been examined whether Lasso can exactly select the most optimal subsets of features among all the possible subsets of given features. We adopted Exhaustive Search with Support Vector Machine (ES-SVM) for two binary classification problems and calculated the generalization error by LOOCV for all the subsets of features. In the "solution-space" constructed with the CV errors and feature subsets, CV errors and the weight vectors of the decision boundaries found by ES-SVM are compared with those obtained by Lasso. We found that the most optimal feature subsets obtained only by ES-SVM consisted several clusters in the solution space. The feature subsets found by Lasso does not necessarily located on the cluster.

1. 序論

高次元データから情報表現に最適な特徴を選択するには，原理的には全ての特徴組み合わせを網羅的に探索する必要がある [Cover 77]．このとき計算量が変数の指数オーダーになってしまうという問題点を回避するため，L1 正則化ロジスティック回帰 (Lasso)[Tibshirani 96] などのスパース推定が広い分野で用いられている．しかしながら，スパース推定によって選択された特徴の組み合わせ（解）が，生じ得る全ての特徴組み合わせが分布する解空間において，どこに位置するかを評価する枠組みはこれまで検討されていない．

本研究では，高次元データを対象に変数の全組み合わせで判別性能を検討する ES-SVM (Exhaustive Search with Support Vector Machine) を行い，解空間のなかで Cross Validation(CV) の汎化誤差が最小となる解（最適解）の特性を解析した．さらに，スパース推定の一つである Lasso によって求められる解が，最適解の集合と比較して，解空間のなかでどこに位置するかを検討した．

2. スパース推定による特徴選択

スパース推定は，高次元かつ小サンプルのデータにおける特徴選択問題においては，用いるデータセットが少し異なっただけで選択される特徴が異なり，適切な特徴組み合わせを抽出することが難しいという問題点がある [五十嵐 15]．Nagata ら [Nagata 15] は，Lasso を用いて，23 次元の高次元かつ小サンプルのデータで二値判別問題における特徴選択を行った．この時，Lasso が特徴を選択したデータセットと，選択しないデータセットがあった．選択できた場合であっても，特徴組み合わせは CV ごとに多様に異なっていた．さらに Kitazono ら [Kitazono 13] は，Nagata ら [Nagata 15] と同じデータに対して ES-SVM を適用し，データセットによっては CV の汎化誤差が最小となる解（最適解）が 40 万通り以上存在することを示した．これらの知見から，Lasso が一意に選択する特徴が，多数存在する最適解の集合のどこに位置するかを解空間の中で評価する必要がある．

3. データの詳細

Eifuku ら [Eifuku 04] が行ったサルを対象とした神経生理学実験の結果を解析した．

実験では，サルが人物の識別に用いている脳神経細胞を同定するため，4 個体の顔それぞれについて，7 通りの角度で撮影した合計 $28(=4\times7)$ 枚の顔画像をマカクサルに観察させた．本研究では角度に依存しない個体識別を検討するため，各個体に対する角度の違いは無視し，1 個体あたり 7 サンプル得られたものとして扱った．このとき，大脳皮質の anterior inferior temporal cortex (AIT) と呼ばれる部位の神経細胞の活動を，23 箇所でシングルユニットレコーディングにより記録した．これら 23 個の神経細胞の発火率を入力データとし，どの神経細胞が角度によらない個体識別に寄与するかを検討するため，4 個体のうち 2 個体どうしを識別する際に必要な神経細胞，すなわち特徴の組み合わせを検討した．

本稿では，先行研究 [Kitazono 13][Nagata 15] でも取り扱われた，個体 3vs4，個体 1vs3 の識別について検討した結果について述べる．

連絡先: 岡田真人，okada@k.u-tokyo.ac.jp

4. ES-SVM

個体 3vs4，個体 1vs3 それぞれの識別問題において，23 個の特徴から生じうる全ての組み合わせに対して，Support Vector Machine(SVM) を適用した．

ES-SVM では，全ての特徴組み合わせ $2^{23}-1=8,388,607$ 通りを探索するために，組み合わせに用いる特徴を表わすベクトルとしてインディケーターを用いる．23 次元の入力データ $\boldsymbol{x}_n(=(x_1,\cdots,x_{23}))$ におけるインディケーター $\boldsymbol{C}_k(k=1,\cdots,8,388,607)$ は以下のように定義する．

$$\boldsymbol{C}_k=(C_{1,k},\cdots,C_{i,k},\cdots,C_{23,k})\in\{0,1\}^{23} \tag{1}$$

$C_{i,k}$ は，$C_{i,k}$ がその特徴組み合わせに含まれるとき $C_{i,k}=1$，含まれないとき，$C_{i,k}=0$ とする．SVM によって求める超平面の切片項を記述するため，入力データ $\boldsymbol{x}_n=(\boldsymbol{x}_n,1)$，インディケーター $\boldsymbol{C}_k=(\boldsymbol{C}_k,1)$ と置き換えた．インディケーター $\boldsymbol{C}_k$ において，$C_{i,k}$=1 である特徴の入力データ x_i のみを保持し，$C_{i,k}$=0 である特徴の入力データ x_i を 0 とするような特徴ベクトルを以下の式で定式化する．

$$\boldsymbol{C}_k\circ\boldsymbol{x}_n \tag{2}$$

$\circ$ はアダマール積であり，$\boldsymbol{C}_k\circ\boldsymbol{x}_n=(c_1x_1,c_2x_2,\cdots,c_{23}x_{23},1)$ となる．SVM は，ラベル y_n がついた学習データ $\boldsymbol{x}_n$

$$\{(\boldsymbol{x}_n,y_n)|\boldsymbol{x}_n=\{x_i\}\in R^D,y_n\in\{+1,-1\}_{n=1}^N\} \tag{3}$$

を用いて，y_n を $+1$ または -1 と判別する超平面 $y(\boldsymbol{x}_n)=\boldsymbol{x}_n\cdot\boldsymbol{w}$ の各係数 $\boldsymbol{w}=\{w_1,\cdots,w_{23},w_0\}^T$ を求める手法である．ただし $C_{i,k}=0$ である場合は $w_i=0$ とする．ここで，N はサンプル数，D は $C_{i,k}\neq0$ となる特徴数である．本研究では，超平面 $y(\boldsymbol{x}_n)$ の係数である $\boldsymbol{w}$ を重みベクトルと呼ぶ．この超平面 $y(\boldsymbol{x}_n)$ を用いた判別器 y_n を以下に記す．

$$y_n=\mathrm{sgn}((\boldsymbol{C}_k\circ\boldsymbol{x}_n)\cdot\boldsymbol{w}) \tag{4}$$

この判別器によってあるデータ $\boldsymbol{x}_n$ のラベル y_n を予測する．これを，全てのインディケーター $\boldsymbol{C}_k(k=1,\cdots,8,388,607)$ について行う．

5. Cross Validation Error

ES-SVM によって得られた 8,388,607 通りの解，すなわち重みベクトル $\boldsymbol{w}$ は，解が存在しうる 24 次元空間（解空間）に存在する．解空間上で重みベクトルがどのように分布するかを評価するため，その評価関数として未知データに対する汎化性能を表す Cross Varidation Error (CVE) を用いた．

Cross Varidation（交差検証）とは，得られたサンプルを訓練データとテストデータに分割し，訓練データを用いてモデルを学習した後，テストデータでそのモデルをテストした際の汎化性能を評価する手法である．本研究では，サンプル数 N のうち 1 つをテストデータとし，残りの $N-1$ を訓練データとする Leave-One-Out CV(LOOCV) を用いた．サンプルサイズが 14 であったため，14 試行の CV が可能であった．

重みベクトル $\boldsymbol{w}$ について，交差検証において誤判別したサンプルの割合の試行間平均を導出し，これを CVE とした．$\boldsymbol{X}$ をサンプルデータ，N をサンプル数，$\boldsymbol{x}_p$ をテストデータ，t_p を $\boldsymbol{x}_p$ の真のラベル，$y(\boldsymbol{X}\backslash\boldsymbol{x}_p)$ を訓練データより得られた超平面とすると，インディケーター $\boldsymbol{C}$ によって $CVE(\boldsymbol{C})$ は一意に表すことが出来る．$CVE(\boldsymbol{C})$ は以下の式で定式化を行う．ただし $\boldsymbol{X}\backslash\boldsymbol{x}_p$ は，$\boldsymbol{X}$ から $\boldsymbol{x}_p$ を引いた差集合であり，訓練データを表す．

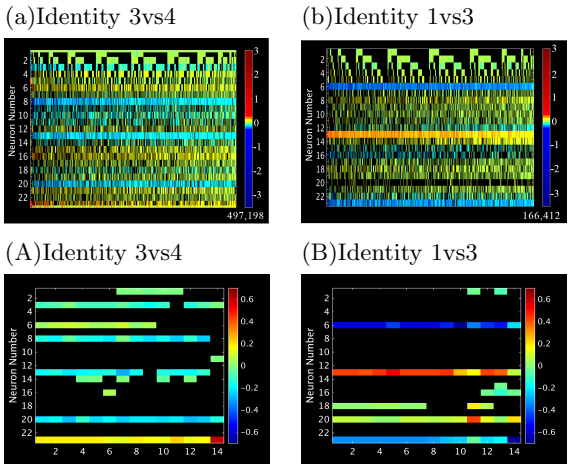

図 1: (a)(b) ES-SVM によって得られた最適解のカラーマップ．行は特徴の番号，列はそれぞれの最適解を表し，右下の数は列の数，すなわち最適解の数を表す．(A)(B) Lasso 解のカラーマップ．行は特徴の番号，列は交差検証の試行数を示す．いずれも，切片項を除いた 23 個の特徴への重みづけのみ示した．列ごとに，各特徴がその色で示される重みで選択されたことを表し，黒は選択されなかったことを示す．

$$CVE(\boldsymbol{C})=\frac{1}{N}\sum_{p=1}^{N}L(t_p,y_p(\boldsymbol{X}\backslash\boldsymbol{x}_p)), \tag{5}$$

$$L(t,y(\boldsymbol{x}))=\begin{cases}0 & (ty(\boldsymbol{x})>0)\\ 1 & (ty(\boldsymbol{x})<0)\end{cases} \tag{6}$$

CVE の最小値は 0 であった．CVE=0 となる解は複数存在し，個体 3 と 4 の識別では 497,198 通り，個体 1 と 3 の識別では 166,412 通りであった．このことから，個体 3 と 4 の識別データのほうが，個体 1 と 3 よりもデータのもつ識別能力が高いことがわかる．

本研究では，それぞれの識別において，CVE=0 となる重みベクトル $\boldsymbol{w}$ の集合を最適解と呼ぶ．図 1(a)(b) に，それぞれの識別における最適解を示す．

6. Lasso

個体 3 と 4，個体 1 と 3 それぞれの識別問題において，Lasso を用いた特徴選択を行った．Lasso によって選択された特徴の重みベクトル $\boldsymbol{w}$ を，以下，Lasso 解と呼ぶ．図 1(A)(B) に Lasso 解を示す．個体 3 と 4 の識別（図 1 (A)）では，14 試行の CV で 14 通りの異なる Lasso 解が得られ，そのうち 11 通りは選択された特徴の組み合わせが等しく重みベクトルのみが異なっていた．個体 1 と 3 の識別 (図 1 (B)) では，14 試行の CV で 14 通りの異なる Lasso 解が得られ，そのうち 6 通りは，選択された特徴の組み合わせが等しく重みベクトルのみが異なっていた．

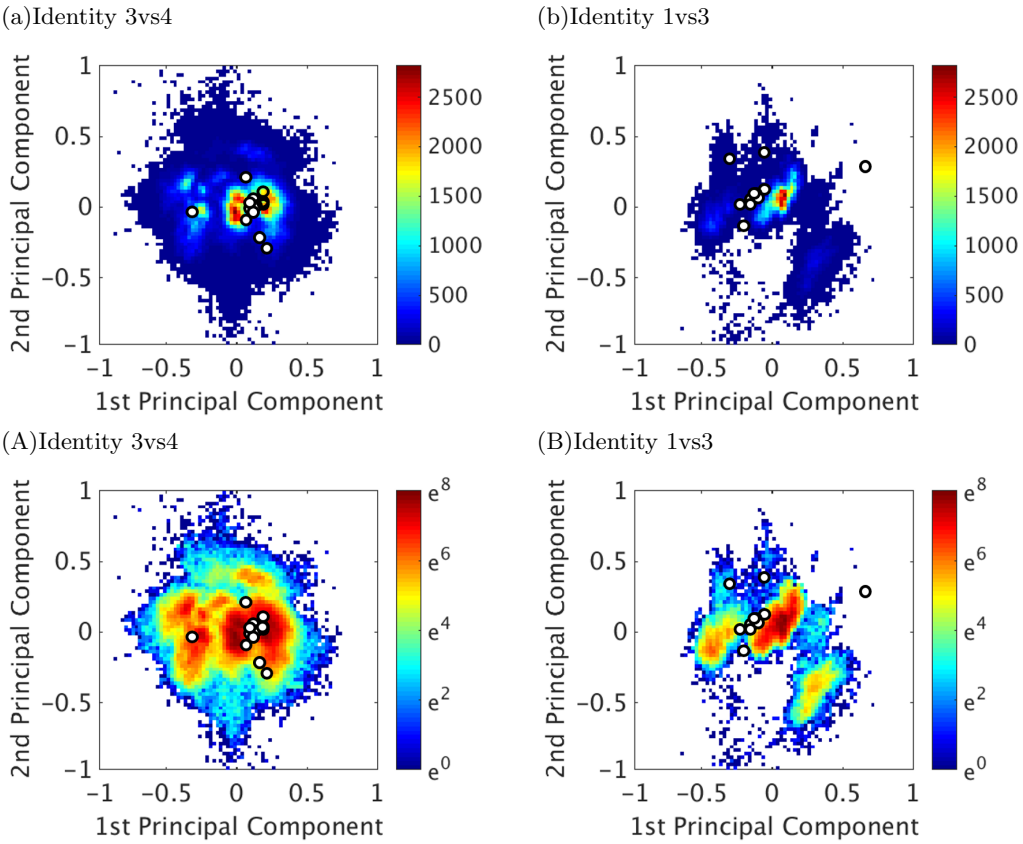

図 2: (a)(b) 最適解の頻度, (A)(B) 最適解の頻度の対数，それぞれを表すヒートマップ．24 次元の最適解に主成分分析を適用した際の第一主成分，第二主成分をそれぞれ縦軸，横軸とした．各ビンの白い部分は値がないことを表す．○は Lasso 解を主成分軸上に射影した時の位置である．

7. 最適解の解析

最適解（CVE=0 となる重みベクトル $\boldsymbol{w}$）が複数存在する場合，解空間でそれらがどのような構造をもつかを可視化するために，主成分分析を行った．

主成分分析とは，高次元データに対して，そのデータの分散を最大にする軸を求め，もとの次元より少ない軸上でデータを表現する目的で用いられる手法である．新しい軸は，寄与率の高い順からに第一主成分，第二主成分と呼ばれ，元データの特徴数と同じ数の主成分を用いた場合に累積寄与率は 100%となる．

ここでは，個体 3 と 4，個体 1 と 3 それぞれの個体識別における ES-SVM によって得られた最適解の構造を二次元上で可視化して記述するために，24 次元の最適解に主成分分析を適用し，得られた第一主成分と第二主成分を軸としてその頻度を 2 次元上にヒートマップで表した (図 2(a)(b))．さらに，それぞれの識別すべきデータに対して Lasso を行って得た，CV の試行ごとに異なる 14 個の Lasso 解についても，各空間上に表した．

図 2(a) を見ると，個体 3 と 4 の識別では，最適解が集中して存在するピークが 2 つあるが，Lasso 解はこれらの間に位置していた．これは，2 つのピークの間で Lasso 解が引かれ合い，最適解の頻度が低い谷間に落ちているように見える．次に図 2(b) を見ると，個体 1 と 3 の識別では，最適解のピークが 1 つ存在するが，Lasso 解はそれとずれた位置に固まっていた．これらの結果から，Lasso 解は最適解のピークを射抜いていないように見られる．ただし，それぞれの図では最適解が多数（個体 3 と 4 の識別では 49 万個，個体 1 と 3 の識別では 16 万個）存在したため，頻度の高い部分の記述にヒートマップの表現力が割かれることで，ほかの部分の構造が見えにくい可能性がある．そこで，図 2(a)(b) の各ビンにおける頻度の対数をとり，図 2 (A)(B) に示した．

図 2 (A) を見ると，個体 3 と 4 の識別においては，最適解の対数頻度のピークは複数あるが，頻度の低い領域を挟まずに 1 クラスターを形成しているように見える．Lasso 解は，最大のクラスターのピーク付近に 13 点が集中し，ほか 1 点は近接する別のピークの上に位置していた．すなわち，Lasso 解は解空間のなかで最適解のピーク付近に位置すると考えられる．一方で図 2(B) を見ると，個体 1 と 3 の識別においては，最適解の頻度のピークが 3 つあるように見える．ピーク同士の間には頻度の低い領域が見られることから，3 つのクラスター構造をもつ傾向が見られた．Lasso 解は，隣接するクラスターの谷間に落ちるように 11 点が存在し，3 点はいずれのクラスターピークから離れたところに孤立していた．このことは，個体 1 と 3 の識別では，Lasso 解は解空間のなかで最適解と近いながらも，中心から逸れた場所に位置すると考えられる．以上の結果から，Lasso は解空間のなかで最適解に近い解を導出しうること，しかしデータセットによっては，最適解が多数集まるピークの谷間に落ちてしまい，ピークを射抜くことができないことが示唆された．

8. 結論と考察

高次元かつ小サンプルのデータにおける特徴選択においては，ES-SVM において汎化誤差を最小にする解（最適解）が多数存在し，それらはクラスタ構造をもつ可能性が示された．一方で，解を一意に求める Lasso は，必ずしも最適解を求めず，最適解の特性とも異なる解を導出する場合があった．

Lasso が最適解のクラスタの近傍に位置するかは，用いるデータセットに依存して異なることが示唆された，先行研究と同様，高次元かつ小サンプルにおけるスパース推定による特徴選択は，データセットの違いによって大きく結果が異なるという知見と一致する．Lasso 解が最適解の近傍に多く位置したのは，本研究の個体 3vs4 の識別の場合であり，最適解の数が多数（生じうる全ての組み合わせの 5%程度（49 万／ 838 万通り））かつ，最適解の第一および第二主成分が 1 クラスタに近い構造をもっていた．

スパース推定のアルゴリズムを選択する上で，最適解の解空間における構造を評価することは必要不可欠であり，ES-SVM によって初めて可能になった．今後は，ES-SVM の唯一の問題点である計算量爆発を回避するアルゴリズムである，レプリカ交換モンテカルロ法による解空間の状態密度推定 [Nagata 15] を利用し，より効率的に解空間におけるスパース推定の性能評価を行う枠組みを検討する．

謝辞

本研究は，科学研究費補助金新学術領域研究（26120529，市川，25106506，26106504，永田；25120009，岡田）および基盤研究 (C)（25330283，永田）の助成を受けた．

参考文献

[Cover 77] Cover, T.M., and Van Canpenhout, J.M.: On the Possible Orderings in the Measurement Selection Problem, IEEE Trans. Systems, Man, and Cybernetics, Vol.7, No.9, 657-661(1977).

[Eifuku 04] Eifuku, S., De Souza, W. C., Tamura, R., Nisijo, H., and Ono, T.: Neuronal Correlates of Face Identification in the Monkey Anterior Temporal Cortical Areas, J Neurophysiol, Vol.91, 358-371(2004).

[五十嵐 15] 五十嵐康彦, 永田賢二, 岡田真人: マシンラーニング (第 3 回) ヒューマンインタフェースと特徴選択問題, Journal of Human Interface Society, Vol.17, No.4, 294-298(2015).

[Kitazono 13] Kitazono, J., Nagata, K., Nakajima, S., Manda, A., Eifuku, S., Tamura, R., and Okada, M.: Exhaustive Search of Feature Subsets for Support Vector Machine Classification, IPSJ SIG Technical Report, Vol. 2013-MPS-92, No. 8.(2013)

[Nagata 15] Nagata, K., Kitazono, J., Nakajima, S., Eifuku, S., Tamura, R., and Okada, M.: An exaustive search and stability of sparse estimation for feature selection problem, IPSJ Online Transactions. Vol.8, 25-32(2015).

[Tibshirani 96] Tibshirani, R.: Regression Shrinkage and Selection via tha lasso, J.Royal Stat. Soc. B, Vol.58, No.1, 267-288(1996).

[Vapnik 82] Vapnik, V. N.: Estimation of Dependences Based on Emprical Data, Springer(1982).

[Yamashita 08] Yamashita, O., Sato, M.A., Yoshioka, T., Tong, F., and Kamitani, Y.: Sparse estimation automatically selects voxels relevant for the decoding of fMRI activity patterns, Neuroimage, Vol.42, No.4, 1414-1429(2008).