2I2-4

K-スパース全状態探索法による識別問題における変数選択

Feature selection by K-sparse Exhaustive-Search (ES-K-SVM) in linear classification

市川寛子 *1 Hiroko Ichikawa
川端大貴 *2 Daiki Kawabata
五十嵐康彦 *2 Yasuhiko Igarashi
永田賢二 *3*4 Kenji Nagata
永福智志 *5 Satoshi Eifuku
田村了以 *5 Ryoi Tamura
岡田真人 *2 Masato Okada

*1東京理科大学 理工学部
Faculty of Science and Technology, Tokyo University of Science
*2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
*3JST さきがけ研究員
PRESTO Researcher, Japan Science and Technology Agency (JST)
*4国立研究開発法人産業技術総合研究所人工知能研究センター
The Artificial Intelligence Research Center, The National Institute of Advanced Industrial Science and Technology (AIST)
*5富山大学 医学薬学研究部
Graduate School of Medicine and Pharmaceutical Sciences, University of Toyama

Exhaustive search (ES) is the only way to find the optimal feature subset in linear classification, however, it requires huge computational complexity when adopted to high-dimensional data. Recently Igarashi et al. modified ES and proposed the ES-K method, in which the optimal feature subsets is assumed to be K-sparse and the K-sparse subsets are exhaustively searched for sparse feature selection in linear regression. In this study, we adopted the ES-K method in linear classification with SVM (ES-K-SVM) and investigated the optimal feature subsets found in each K. Adopted to high-dimensional small-sample data, the ES-K-SVM founds multiple optimal subsets. Most of optimal subsets commonly consisted of the limited number of feature. When we assume that any data can be described using K-sparse features, such limited number of features should be select as key features. To confirm our insights experimentally, we would conduct virtual measurement and analysis (VMA).

1. 序論

高次元データから説明変数を選択する際，最適な変数を厳密に求めるためには全ての変数組み合わせを網羅的に探索する必要がある [Cover 77]．全状態探索法の最大の問題点は，計算量が変数の指数オーダーになることであるが，五十嵐ら [五十嵐 16] によって提案された K-スパース全状態探索法によって部分的に克服できる．K-スパース全状態探索法では，最適な説明変数はたかだか K 個であると仮定し，変数の個数 K の値を 1 から順に増やしながら，K 個の変数組み合わせごとに最適な変数組み合わせを網羅的に探索していく手法である．

本研究では，K-スパース全状態探索法を SVM[Vapnik 82] を識別器とする識別問題に適用し（ES-K-SVM），選択すべき変数について考察を行った．高次元データにおける識別問題では，特にサンプルが少ない場合に最適な変数組み合わせが複数得られることがあり [五十嵐 15]，いずれの組み合わせを選ぶべきかを判断する基準がない．川端らは，K-スパース全状態探索によって得られる最適な変数組み合わせを K を追って観察し，K が変わっても選ばれ続ける変数があることを示しており [川端 16]，本研究では変数の特徴についてさらに考察した．

2. データ

23 変数，14 サンプルからなる実データ解析を行った．データは，Eifuku ら [Eifuku 04] が行った神経生理学実験において計測された，サルの脳神経活動の発火率であった．実験では，サルに顔写真を観察させ，4 人の人物を同定させる遅延見本合わせ課題を行わせた．このとき，人物同定に関連すると考えられる大脳皮質 anterior inferior temporal cortex (AIT) 領域の 23 個の神経細胞においてシングルユニットレコーディングを行った．

識別問題の目的は，23 個の神経細胞の発火率を入力データとし，2 人の人物に対して差別的に応答する神経細胞を抽出することであった．本稿では，先行研究 [Kitazono 13][Nagata 15][川端 16] でも取り扱われた，個体 1vs3 の識別問題に適用した結果のみ記載する．

連絡先: 岡田真人，okada@k.u-tokyo.ac.jp

3. 定式化

3.1 インディケーターの導入

入力変数の数が D 個である時，それら全ての変数の組み合わせは 2^D-1 通りである．各組み合わせに使用する変数を表わすベクトルとしてインディケーターを導入する．D 次元の入力データ $\boldsymbol{x}_n(=(x_{n,i},\cdots,x_{n,D}))$ におけるインディケーター $\boldsymbol{C}_k(k=1,\cdots,2^D-1)$ を以下のように定義する．

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

$c_{k,i}$ は，k 個目のインディケーターが i 番目の変数を含んでいる場合は $c_{k,i}=1$，含んでいない場合は，$c_{k,i}=0$ とする．このインディケーター $\boldsymbol{C}_k$ と入力データ $\boldsymbol{x}_n$ において，$c_{k,i}$=1 の時，$x_{n,i}$ を保持し，$c_{k,i}$=0 の時，$x_{n,i}$ を 0 とするような変数

ベクトルを以下の式で定式化する．

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

$\circ$ はアダマール積であり，$\boldsymbol{C}_k \circ \boldsymbol{x}_n = (c_{k,1}x_{n,1}, \cdots, c_{k,D}x_{n,D})$ となる．この入力データを基に SVM によって $y(\boldsymbol{x}) = (\boldsymbol{C}_k \circ \boldsymbol{x}) \cdot \boldsymbol{w} + w_0 = 0$ という識別平面を求める．ただし $c_{k,i} = 0$ である場合は $w_i = 0$ とする．本研究では，各インディケーターから得られた識別平面 $y(\boldsymbol{x})$ を評価するため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 validation error (CVE) を用いた．

3.2 ES-K-SVM

ES-K-SVM では，インディケーター $\boldsymbol{C}_k$ の要素 $c_{k,i}$ の非ゼロである個数を K 個と決め，各 K について，変数の全ての組み合わせを探索する手法である．この ES-K-SVM は，データの全ての変数の数が D 個であるとすると，${}_DC_K$ 通りのインディケーターに対して計算を行う．本研究では，選択される変数がスパースであると仮定し，$K = 1, 2, 3, \cdots$ と少数の変数から ES-K-SVM を行った．$K = 1, \cdots, D$ までの全ての K で計算を行えば，SVM 全状態探索（ES-SVM）と同様の計算量になるが，K が D に達する前に計算を中断できれば，ES-SVM と比べて計算量を抑えられる．

4. 結果

まず K=1 で ES-K-SVM を行った結果，CVE が 0 となる変数組み合わせはなかった．次に，K=2 で ES-K-SVM を行った結果，CVE が 0 となる変数組み合わせが 2 通りあった．それぞれ，6 番目と 13 番目の神経細胞の組み合わせ（以下（6,13）のように表記する），および (16,20) であった．K=3 では，CVE が 0 となる変数組み合わせが 16 通りであった．このうち，14 通りは K=2 で得られた組み合わせに新たな変数を 1 つ加えた組み合わせであった．このとき，K=2 で得られた組み合わせを「親」と呼び、K=3 で得られた、「親」に任意の変数 1 つを足した上位集合を「子」と呼ぶこととする．親子構造の概念図を図 1 に示す．

K=4 では，CVE が 0 となる変数組み合わせが 91 通りであった．このうち，81 通りは K=2 および K=3 で得られた「親」に新たな変数を 1 つ加えた組み合わせ，すなわち「子」であった．

5. 考察

高次元少サンプルの識別問題で得られる複数の最適な変数組み合わせには，限られた数の変数が繰り返し用いられていることが示された．最適な変数組み合わせが複数あったとしても，そのほとんどが「親」の上位集合として表現できる場合には，よりスパースな K で得られる「親」を最適な変数組み合わせとして採用することが考えられる．これは，変数の数 K ごとに分断して，最適な変数組み合わせを抽出する K-スパース全状態探索によって初めて達成されたことである．

今後は，今回の実データ解析で得られた結果が，どのようなデータ構造によって得られたかを議論するために，実データの生成モデルを考え，ここから人工的に生成した仮想データに同様の手法を適用して解析する．この枠組みは VMA(Virtual Measurement and Analysis, 仮想計測解析) [五十嵐 16] と呼ばれ，これによって実データ解析によって得られた結果，およびこれにもとづく知見がどれだけ真実に近いかを議論することができる．

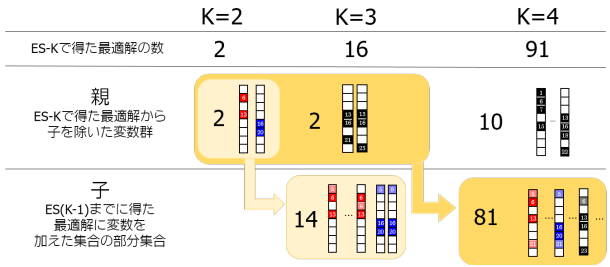

図 1: 各 K において得られた最適な変数組み合わせに見られる，変数の親子構造．

謝辞

本研究は，科学研究費補助金新学術領域研究（26120529,16H01555，市川；25120009，岡田）の助成を受けた．

参考文献

[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).

[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 16] 五十嵐康彦，竹中光，中西（大野）義典，植村誠，池田思朗，岡田真人: 全状態探索による線形回帰のスパース変数選択, 信学技報, vol. 116, no. 300, 313-320(2016).

[川端 16] 川端大貴，市川寛子，五十嵐康彦，永田賢二，永福智志，田村了以，岡田真人: SVM 全状態探索法 (ES-SVM) によるスパース変数選択, 信学技報, vol. 116, no. 300, 361-368(2016).

[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).

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