1G2-3

交換モンテカルロ法による変数選択問題における解の効率的全数検索

An exhaustive search method for feature selection with the exchange Monte Carlo method

*1永田賢二 Kenji Nagata　*1北園淳 Jun Kitazono　*2中島伸一 Shin-ichi Nakajima　*3永福智史 Satoshi Eifuku　*4田村了以 Ryoi Tamura　*1岡田真人 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(株) ニコン光技術研究所 Optical Research Laboratory, Nikon Corporation
*3福島県立医科大学 Fukushima medical University
*3富山大学　医学薬学研究部 Graduate School of Medicine and Pharmaceutical Sciences, University of Toyama

Feature selection in machine learning is an important process for improving the generalization capability and interpretability of learned models through the selection of a relevant feature subset. In the last two decades, a number of feature selection methods, such as L1 regularization and automatic relevance determination have been intensively developed and used in a wide range of areas. We can select a relevant subset of features, by using these feature selection methods. In this study, we propose a new method of an exhaustive search method, instead of those method, by using the exchange Monte Carlo method.

1. まえがき

与えられた特徴量や変数の中から，意味のある特徴量の部分集合・組み合わせを選択する変数選択問題は，教師あり学習において，汎化性能を向上する上で重要な課題である．それを実現する手法として，L1 正則化などのスパース推定が近年，広く用いられている [Tibshirani 96]．L1 正則化は，線形計画問題に帰着させることができ，多項式オーダーの計算量で計算できる．その性質ゆえ，広い分野で応用されるようになっている．

変数選択の目的が，少ないデータからの再構成であり，未知データに対する汎化性能の向上の場合，得られた特徴は副産物であることが多い．一方で，得られた特徴から背後にある物理法則などを抽出することが目的である場合，高次元・少数サンプルの状況では，再構成を実現する変数の組み合わせは複数あることが一般的であり，解釈を誤ってしまう危険性がある．この問題を解決する為に，最適解として実現されうる変数の組み合わせ全てを網羅的に探索する手法の確立が重要である．

本研究では，交換モンテカルロ法 [Hukushima 96] を用いて，変数選択問題において最適解の分布を効率的に求める手法を提案する．提案手法では，変数の組み合わせに関する評価関数として，未知データに関する汎化性能を示す Cross Validation 誤差 (CV 誤差) を利用する．この CV 誤差をエネルギー関数として，モンテカルロ法を実装することで，解の候補の効率的な手法を提案する．さらに，マルチヒストグラム法 [Hukushima 02] に基づき，モンテカルロ法のサンプリング結果から解の個数を推定する方法を提案する．これらの手法を顔識別に寄与する神経細胞の選択問題に応用することで，有効性を検証する．

2. 提案手法

本研究では，交換モンテカルロ法を用いて，変数選択問題における最適解の分布を効率的に求める手法を提案する．

2.1 ボルツマン分布の導入

本提案手法では，変数の組み合わせに関する評価関数として，未知データについての汎化性能を表す Cross Varidation 誤差を利用する．Cross Validation では，データを分割し，データの一部を用いてモデルを学習させ，学習に用いなかった残りのデータによって，学習したモデルのテストを行い未知のデータに対する予測性能を評価する．交差検定は，学習に用いることが出来るデータ数が限られている場合に有効である．

本手法では，評価関数である CV 誤差を最小にする変数の組み合わせのうち一つを求めるだけではなく，そうした変数の組み合わせ全てを網羅的に探索する手法を目指す．そのために，CV 誤差をエネルギー関数としたボルツマン分布を導入する．

D 個の特徴量からなるデータを考える．特徴量の部分集合を $\boldsymbol{S} = (S_1, \ldots, S_D) \in \{+1, -1\}^D$ で表すとする．すなわち，S_i がその部分集合に含まれるとき $S_i = 1$，含まれないとき，$S_i = -1$ とする．また，この部分集合を用いた際の CV 誤差をエネルギー $E(\boldsymbol{S})$ とし，次のボルツマン分布を定義する．

$$p(\boldsymbol{S}; \beta) = \frac{1}{Z_\beta} \exp(-\beta E(\boldsymbol{S})). \tag{1}$$

β は逆温度と呼ばれる変数，Z_β は規格化因子である．この分布は，CV 誤差が低い値となる部分集合について大きな確率をとる．提案手法では，この分布からサンプリングを行うことによって，CV 誤差が低い値となる部分集合を重点的に探索する．

2.2 交換モンテカルロ法

交換モンテカルロ法は，古典的なマルコフ連鎖モンテカルロ法の拡張である．交換モンテカルロ法を用いることによって，MCMC 法で問題となる，遅い緩和の問題を避け，効率的なサンプリングを行うことが可能となる．

交換 MC 法では，異なる逆温度を持つレプリカを用意する．すなわち異なる逆温度 $0 = \beta_1 < \beta_2 < \cdots < \beta_M$ を定義し，次の結合確率を考える．

$$p(\boldsymbol{S}_1, \ldots, \boldsymbol{S}_M) = \prod_{m=1}^{M} p(\boldsymbol{S_m}; \beta_m). \tag{2}$$

右辺の $p(\boldsymbol{S}; \beta)$ は，式 (1) で定義される確率分布である．交換 MC 法は，大きく分けて次の二つのステップからなる．

各温度ごとの通常のサンプリング

各逆温度ごとに通常のメトロポリス法のサンプリングを

連絡先: 岡田真人:okada@k.u-tokyo.ac.jp

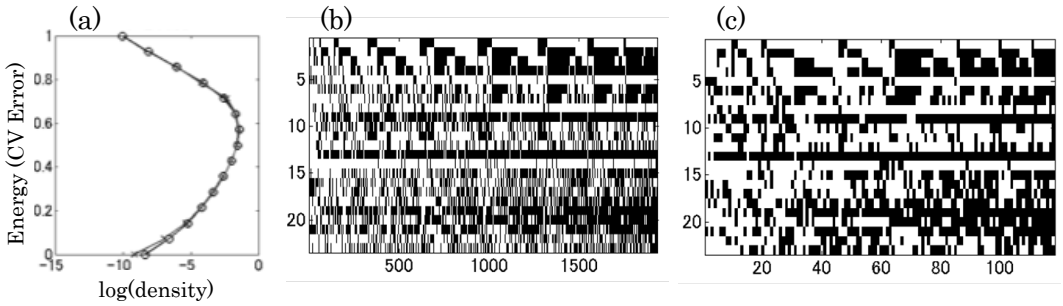

図 1: (a)CV 誤差のヒストグラムの推定結果．○が全数検索の結果，×が交換 MC 法の推定結果である．(b)(c)CV 誤差最小を示す

行う．すなわち，$p(\boldsymbol{S_m};\beta_m)$ 分布からメトロポリス法によって，$\boldsymbol{S}_m$ をサンプリングする．

隣り合う温度間での状態交換

隣り合う温度間ごとでの状態を交換し，$\{\boldsymbol{S}_m, \boldsymbol{S}_{m+1}\} \rightarrow \{\boldsymbol{S}_l, \boldsymbol{S}_{l+1}\}$ とする．交換する確率は，以下で定める．

$$p(\boldsymbol{S}_m \leftrightarrow \boldsymbol{S}_{m+1}) = \min(1, v) \tag{3}$$

$$v = \frac{p(\boldsymbol{S}_{m+1};\beta_m)p(\boldsymbol{S}_m;\beta_{m+1})}{p(\boldsymbol{S}_m;\beta_m)p(\boldsymbol{S}_{m+1};\beta_{m+1})} \tag{4}$$

$$= \exp\left((\beta_{m+1}-\beta_m)(E(\boldsymbol{S}_{m+1})-E(\boldsymbol{S}_m))\right) \tag{5}$$

2.3 マルチヒストグラム法

交換 MC 法のもう一つの利点として，メトロポリス法などの通常のモンテカルロ法では推定が困難である，状態密度 $g(E)$ と規格化定数 Z_β の計算が可能なことが挙げられる．状態密度 $g(E)$ は，エネルギーの値が E となる状態数のことをいう．すなわち，エネルギーのヒストグラムである．これらの二つの量は，交換 MC 法のサンプリング結果から，以下に説明するマルチヒストグラム法を用いて，推定することができる．

交換 MC 法によるサンプリングの結果，各逆温度ごとにエネルギー値 E をとる状態が $N_m(E)$ 個得られたとする．このとき，エネルギー値 E をとる状態の状態数 $g(E)$ は，

$$g(E) = \frac{\sum_{m=1}^{M} N_m(E)}{\sum_{m=1}^{M} n_m \mathrm{e}^{f_m-\beta_m E}} \tag{6}$$

$$\mathrm{e}^{-f_m} = \sum_E g(E)\mathrm{e}^{-\beta_m E} \tag{7}$$

によって与えられる．ここで n_m は逆温度 β_m での総サンプル数であり，f_m は，自由エネルギーと呼ばれる量であり，規格化定数 Z_β の対数である（$f_m = \log(Z_{\beta_m})$）．式 (6) と (7) を交互に繰り返し解くことによって，ヒストグラム $g(E)$ と自由エネルギー f_m を求めることができる．

3. シミュレーションによる検証

提案手法の有効性を検証するために，いくつかのシミュレーションを行った．ここでは，その結果についてまとめる．

用いるデータは，マカクサルに顔画像を提示した際の神経細胞の活動を記録したデータである．提示した顔画像は，4 つの identity それぞれに対して，7 つの異なる角度から見たものであり，合計 $28 = 4 \times 7$ 枚である．この 28 枚の刺激をそれぞれ提示した際の，anterior inferior temporal cortex (AIT) と呼ばれる部位の神経細胞の活動を，電極を用いて記録を行った．活動を記録した神経細胞数は，23 個である．

上記データに対し，23 個のニューロンを特徴量とし，どのニューロンの計測データを用いれば，顔の向きによらず Identity のペアの識別を行えるか解析した．そのような変数選択問題に対し，提案手法を適用した．

図 1(a) は，Identity 1 と 4 の識別に関する CV 誤差のヒストグラムの推定結果である．図中の○は，組み合わせ総数 8388607 通り全てを調べて作成したヒストグラムで，×は，交換 MC 法で 82800 回サンプリングして得られたヒストグラムである．そのため，計算量はおよそ 100 分の 1 である．この結果から，交換 MC 法とマルチヒストグラム法により，CV 誤差のヒストグラムが精度よく推定できていることが伺える．

図 1(b),(c) は，それぞれ全数探索および提案手法により求められた，CV 誤差最小となる変数の組み合わせを示したものである．横軸は組み合わせの番号，縦軸は神経細胞の番号を表す．黒は該当する部分集合に神経細胞が選ばれていることを表し，白は選ばれないことを表す．全数探索により求められた CV 誤差最小の組み合わせは 1938 通りであり，サンプリングされた組み合わせは 118 通り (118/1938=6.09%) であった．図 1(a) で示したように，サンプリングの途中でヒストグラムを推定できるため，サンプリングされた割合も同時に推定可能である利点がある．

4. まとめ

本研究では，交換モンテカルロ法を用いた，変数選択問題における解の効率的な全数探索手法を提案した．また，マルチヒストグラム法による解の個数の推定法も提案し，顔識別データにおいて提案手法の有効性を検証した．

参考文献

[Eifuku 11] S.Eifuku, W. C. De Souza, R. Nakata, T. Ono, and R. Tamura, “Neural Representations of Personally Familiar and Unfamiliar Faces in the Anterior Inferior Temporal Cortex of Monkeys”, PLoS One, 6, e18913 (2011)

[Hukushima 96] K. Hukushima and K. Nemoto, “Exchange Monte Carlo Method and Application to Spin Glass Simulations”, J. Phys. Soc. Jpn., 65, 1604-1608 (1996)

[Hukushima 02] K. Hukushima, “Extended ensemble Monte Carlo approach to hardly relaxing problems”, Comput. Phys. Commun, 147, 77-82 (2002)

[Tibshirani 96] R. Tibshirani, “Regression shrinkage and selection via the lasso”, J. Royal Stat. Soc. B, 58, 267-288 (1996)