Vol.2016-MPS-109 No.10
2016/7/25

温度交換MCMC法を用いたびまん性肺疾患画像の特徴量選択

小岩井 誠[1,a)]　飯田 のどか[1]　庄野 逸[1]　木戸 尚治[2]

概要：びまん性肺疾患の疾病部の X 線 CT 画像には多様な陰影パターンが含まれており，診断において判別が困難な場合がある．本研究では，このような肺疾患パターン識別に有効な特徴量の選択問題を機械学習的なアプローチで取り扱うことを考える．このような画像から得られる特徴量の内のどれを選択するかという問題は，最適解を求めようとすると全ての組合せを探索する組合せ最適化問題となるため，特徴数の指数オーダーに比例した計算困難を伴う．近年，この特徴量選択問題に対して温度交換マルコフ連鎖モンテカルロ法を用いた手法が Nagata らによって提案されており，現実的な時間のうちに解集合が得られる可能性があることが示唆されている．そこで本研究では，びまん性肺疾患の陰影パターンから得られる 39 種類の特徴量から，各クラス識別に有効な特徴量の組合せを Nagata らの手法によって選択することを試みた．その結果，我々は有効な特徴量の組み合わせ候補に関する示唆を得ることが出来た．

キーワード：特徴量選択，医療画像，温度交換マルコフ連鎖モンテカルロ法

Feature Selection for Diffuse Lung Disease using Exchange Markov Chain Monte-Carlo Method

Makoto KOIWAI[1,a)]　Nodoka IIDA[1]　Hayaru SHOUNO[1]　Shoji KIDO[2]

Abstract: Diffuse lung disease (DLD) in high resolution computed tomography images show a lot of variations even in the same class, and this variations make difficulty in diagnosis. In this study, we treat a effective feature selection problem for this DLD pattern classification using machine learning approach. In order to obtain the best feature selection for classification, we should search whole combination of features, which requires exponential order calculation cost. Recently, Nagata et al. proposed an application of Exchange Markov Chain Monte Carlo (ExMCMC) method for this problem, and suggested that they reveals hidden feature structures for classification. Thus, we tried their method to select the effective feature combination for each DLD classification from 39 types of features, which are obtained from typical texture analysis method in the image processing. As the result, we obtained the effective feature combination candidates for each DLD classification problem.

Keywords: Feature Selection, Medical Image, replica exchange MCMC

1. はじめに

びまん性肺疾患とは肺の広範囲にわたり疾病部が拡がる疾患の総称であり，代表的な疾患である間質性肺炎では肺胞の隔壁や血管などに炎症が発生する．患者は肺の酸素吸収力が低下し，病状が進行すると疾患は難治化し死亡率が高まるため，疾病部が拡大していない早期段階での発見が望まれる．びまん性肺疾患の診断には高解像度 CT 画像が使用されているが，画像上での陰影パターンは多様かつ複雑なものとなり，診断は専門医においてもその経験に左右される．このような背景のもとで，医師の補助を目的とし

[1] 電気通信大学
UEC, Chofu, Tokyo 182-8585 Japan
[2] 山口大学
YU, Ube, Tokyo 755-8611 Japan
[a)] makoto.koiwai@uec.ac.jp

た計算機診断支援（Computer Aided Diagnosis: CAD）システムの構築が望まれており，多くのパターン識別システムを用いた研究が為されてきた [2][5].

機械学習分野におけるパターン認識システムは，特徴量抽出器と識別器に分けて議論され，システム内部では入力データを特徴量を介して表現し，この表現に対する学習モデルを構築することで識別器が構成される [1]．菅田らによる先行研究では，びまん性肺疾患の CT 画像に対する “テクスチャ特徴” と呼ばれる特徴量を抽出し，識別及び解析結果の比較が行われた [6]．このような研究ではパターン認識に使用する特徴量を実験的に選択しているが，本来，特徴量は識別器の性能を左右する重要な要素である．特に特徴量の数に関しては，その数が少なければ表現能力が過少になり，その数が多くても “次元の呪い” といった問題を起きるため，パフォーマンス低下を引き起こす可能性がある．特徴量選択とは，複数の特徴量の中からパターン認識において有効な特徴量の組合せを選ぶ処理であり，学習モデルの識別精度を向上させるための重要な要素である．本研究では，特徴量選択問題において，学習モデルの識別率に汎化誤差を規範に用いて特徴量の組合せを探索することを考える．

組合せ最適化問題における最も単純な探索法は，全ての組合せに対してモデルのパフォーマンスを算出する方法であり，これは全状態探索法と呼ばれる．この手法を用いることで最適解を探索することができるが，探索対象の次元数が大きくなるにつれて指数オーダーの計算コストを必要とする．これは探索対象の組合せ数が次元数の指数的に増加するためであり，全状態探索法を用いて高次元の組合せ最適化問題を扱うことは困難であると考えられてきた．特に特徴量選択においては，探索対象が高次元になることに加え，対象の種類が利用するデータセットに依存して変化してしまう．

この問題に関して，近年 Nagata らによる温度交換マルコフ連鎖モンテカルロ（Exchange Markov Chain Monte Carlo: ExMCMC）法を用いた特徴量選択が提案されており，神経科学やスペクトル分析の分野において成果を挙げている [4]．本研究では，ExMCMC 法を用いた Nagata らの手法をびまん性肺疾患画像の特徴量選択へ適用し，学習モデルの汎化誤差の分布に関する効率的なサンプリング手法を検討する．

2. 特徴量選択手法

本研究で用いた特徴選択手法の模式図を図 1 に示す．この手法では，従来の識別システム同様にデータセットからテクスチャ特徴量を抽出し，これら一つ一つに対する使用／不使用の選択を行った上で識別器を構築する．構築される識別器の性能は利用する特徴量の組合せによって変化し，それぞれに対して性能のスコアが付けられる．これら

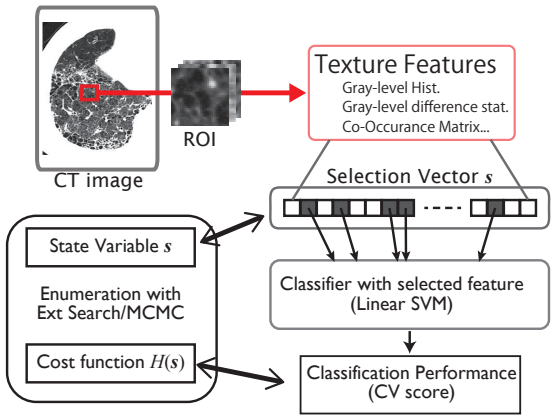

図 1　本研究で構築する特徴量選択手法の模式図

全ての組合せのうち，スコアが高いものが良い特徴量の組合せであると判断できる．特徴量が少ない場合は全状態探索による全スコアの算出が可能であるが，特徴量が多い場合では膨大な組合せ数のために探索が困難になる．そこで本研究では，Nagata らの手法に従って ExMCMC 法を用いた組合せ探索について検討する．以降では，本実験で用いる各手法についての説明を行う．

2.1 識別器

機械学習における識別器の学習とは，“訓練データ” と呼ばれる入力データを基に，入力空間を分割する面を構成するアルゴリズムの総称である．訓練データとは分類クラスを示すラベルが付与されたデータのことであり，識別器の性能は訓練データに依存する．線形 Support Vector Machine (SVM) のような 2 クラス分類器を用いて他クラス分類を行う場合，一対他識別器などを構築し，この識別結果を組合せてデータの分類を行うことが一般的である．一対他識別器とは任意の 1 クラスと他のクラスを分類する識別器のことであり，この組合せは対象クラスの数だけ考えることができる．

2.2 交差検定（CV）法

本研究では，識別器の能力を測る指標として交差検定 (Cross Validation: CV) 法によって得られるスコアを採用した．以下ではこのスコアのことを CV スコアと呼ぶ．CV スコアは識別器の汎化能力の指標を示しており，これがより低い値を示すことは識別精度が高いことを意味する．CV 法では初めに，標本データを k 等分し分割データを作成する．1 つの分割データを評価用として除き，それ以外の分割データを学習用として識別器の構築に使用する．その後，得られた識別器を用いて評価用データのラベルを予測し，予測ラベルと正解ラベルとの平均二乗誤差を距離として測定する．評価用データと学習用データの全ての組合せにおける識別器の精度を評価し，得られた識別精度の平均値を CV スコアとする．

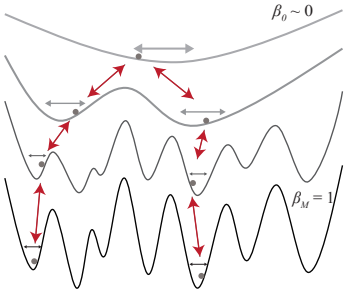

図 2　温度交換 MCMC 法のコンセプト図．横軸は状態空間を，縦軸はエネルギー関数 $\beta H(\boldsymbol{s})$ を示している．複数の温度を用いることで容易に局所解から抜け出すことができる．

2.3　温度交換マルコフ連鎖モンテカルロ（ExMCMC）法

マルコフ連鎖モンテカルロ法（MCMC）法とは，ある想定した不変分布に従う乱数を発生させサンプリングする手法である．対象のエネルギー関数が多谷構造をしている場合，MCMC 法のサンプリングは局所解にとらわれてしまう可能性がある．この問題を解決した手法が温度交換 MCMC 法（ExMCMC）である [3][4]．ExMCMC 法ではエネルギー関数の制約を緩めるために，温度変数 $T > 0$ の逆変数 $\beta = 1/T$ を考える．逆温度 β を重みとして考えると，新たな遷移確率 $p(\boldsymbol{s}) \propto \exp(-\beta H(\boldsymbol{s}))$ を定義することができる．温度 $\beta = 1$ は元のエネルギー関数を意味しており，これより小さな β では $H(\boldsymbol{s})$ の影響が低下する．

図 2 は ExMCMC 法のコンセプト図である．ExMCMC 法によるサンプリングでは M 個の MCMC システムを準備し，適切な逆温度 $0 < \beta_0 < \beta_1 < \cdots < \beta_{M-1} = 1$ を設定する．そして，m 番目のシステムが確率 $p(\boldsymbol{s}_m) \propto \exp(-\beta_m H(\boldsymbol{s}_m))$ に従うようなサンプリングを行い，定期的に隣接温度の状態を交換する．隣接温度の状態交換では以下の操作を行う．

(1) 時間 τ における 1 組の隣接温度 β_j と β_{j+1} を選ぶ．
(2) 次式によって交換確率 r' を計算する．

$$r' = \min\left(1, \exp(-\Delta)\right) \tag{1}$$

$$\Delta = (\beta_j - \beta_{j+1})\left(H(\boldsymbol{s}^{(\tau)}{}_j) - H(\boldsymbol{s}^{(\tau)}{}_{j+1})\right) \tag{2}$$

(3) $[0, 1]$ 上で発生させた一様擬似乱数 u' を r' と比較し，$u' < r'$ の場合には状態 $\boldsymbol{s}^{(\tau)}{}_j$ と $\boldsymbol{s}^{(\tau)}{}_{j+1}$ を交換する．

この操作により，ExMCMC 法は単一の MCMC 法より容易に局所解から脱出することができ，広域のサンプリングを行うことが可能になる．

3.　実験

3.1　対象データ

本研究では，徳島大学医学部附属病院より提供された CT 画像を対象データとした評価実験を行った．この CT 画像の詳細は次の通りである．撮影機器：東芝製 “Aquilion 16”，画像サイズ：512×512 [pixels]，画素サイズ：0.546 ∼ 0.820 [mm]，スライス厚：1.0 [mm]．実験では，対象データを医師の指導の下で 36 × 36 [pixels] の関心領域（Region of Interest: ROI）に切り取ったグレースケール画像を入力データとした．各診断クラスの典型的な ROI 画像は図 3 に示しており，各クラスの枚数は次の通りである．Consolidation (CON): 38, GGO: 76, Honeycomb (HCM): 49, Reticular (RET): 37, Emphysema (EMP): 54, Nodular (NOD): 48, Normal (NOR): 58.

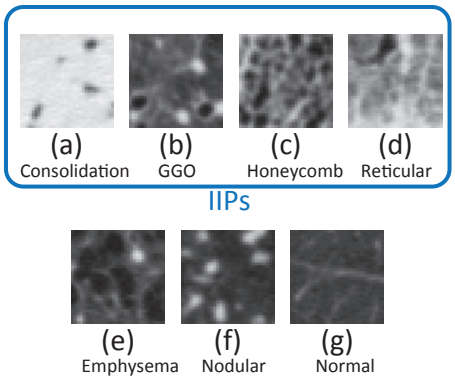

図 3　7 クラスの ROI 画像の典型例．(a) から (d) が間質性肺炎（IIPs）に関わりのあるクラスであり，(e) と (f) は異なる疾患，(g) は正常状態のクラスである．

3.2　手順

準備した ROI 画像に対し菅田らの手法に基づいたテクスチャ解析を適用し，39 個の特徴量を獲得した [6]．テクスチャ特徴には濃淡ヒストグラム，差分統計量，同時生起行列，ランレングス行列とフーリエパワースペクトルに関する特徴量が含まれている．本実験では，これらの特徴量の中から 7 クラスそれぞれの分類に有効な組合せを ExMCMC 法を用いて探索した．ここで ExMCMC 法のエネルギー関数として 10 分割 CV スコアを使用し，温度数は 7 個，サンプリング回数は 2,000,000 回とした．

4.　結果

図 4 は ExMCMC による $H(\boldsymbol{s})$ の密度を示している．CON クラスでは，CV スコアが 0 になる状態が多いため，識別が簡単であると考えられる．一方 GGO クラスや NOD クラスでは，CV スコアの最小値が他のクラスより大きいため，識別が特に難しいと考えられる．

図 5 は CV スコア上位 5 つで選択された特徴量の組合せを示している．この図から，CON クラスではランレングス行列の特徴量の殆どが選択されていることが明らかである．同様に，GGO クラスでは差分統計量とフーリエパワースペクトルの方位に関する特徴量が重要であり，NOD クラスでは同時生起行列と濃淡ヒストグラムに関する特徴量が重要だと考えられる．

5.　考察

本研究では，全状態探索に基づいている Nagata らの手法を用いた特徴量選択を行った．全状態探索では，特徴量数が少ない場合は最善の組合せを見つけ出すことができるが，特徴量数が多い場合は計算コストが膨大になってしまうため探索が困難

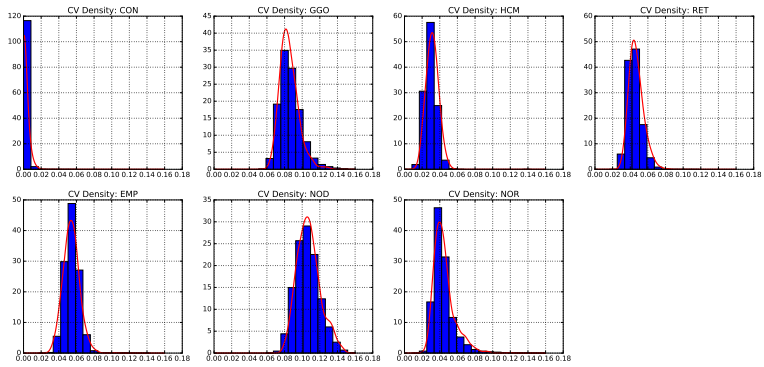

図 4 各クラスにおけるエネルギー関数 $H(\boldsymbol{s})$ の密度．横軸は CV スコアを，縦軸は密度をそれぞれ示している．各図において縦棒はヒストグラムを，赤線は推定された密度分布をそれぞれ表している．上行の図は CON, GGO HCM, RET クラスの，下行の図は EMP, NOD, NOR クラスの結果にそれぞれ対応している．

図 5 CV スコア上位 5 つで選択された特徴量の組合せ．横軸は特徴量番号，縦軸は順位をそれぞれ示している．黒マスは各クラスにおいて選択された特徴量の場所を表している．左列の図は CON, GGO HCM, RET クラスの，右列の図は EMP, NOD, NOR クラスの結果にそれぞれ対応している．

となってしまう．この問題を解決する方法として用いたサンプリング手法が ExMCMC 法である．これにより，準最適な密度曲線を獲得することができ，分類問題としての難しさを推測することができた．

今後の展望としては，スパース推定による識別法を用いた特徴量選択手法の結果との比較を行うべきである．スパース推定は強力な手法であるが，しばしば異なるスパース推定手法間で特徴量抽出の結果が異なる場合がある [4]．そのため，これらの取扱いは慎重に行うべきであり，ExMCMC 法の結果が良い指標になると考えられる．

参考文献

[1] Duda, R. O., Hart, P. E. and Stork, D. G.: *Pattern Classification (2nd edition)*, Wiley-Interscience, 2nd edition edition (2000).

[2] Gangeh, M. J., Sorensen, L., Shaker, S. B., Kamel, M. S., de Bruijne, M. and Loog, M.: A Texton-Based Approach for the Classification of Lung Parenchyma in CT Images, *MICCAI*, LNCS 6363, No. 3, Springer-Verlag Berlin Heidelberg, pp. 595–602 (2010).

[3] Koji, H. and Koji, N.: Exchange Monte Carlo Method and Application to Spin Glass Simulations, *Journal of the Physical Society of Japan*, Vol. 65, No. 6, pp. 1604–1608 (online), DOI: 10.1143/JPSJ.65.1604 (1996).

[4] Nagata, K., Kitazono, J., Nakajima, S.-i., Eifuku, S., Tamura, R. and Okada, M.: An exhaustive search and stability of sparse estimation for feature selection problem, *IPSJ Transactions on Mathematical Modeling and Its Applications*, Vol. 8, No. 2, pp. 23–30 (2015).

[5] Shouno, H. and Okada, M.: Bayesian Image Restoration for Medical Images Using Radon Transform, *Journal of the Physical Society of Japan*, Vol. 79, p. 074004 (online), DOI: 10.1143/JPSJ.79.074004 (2010).

[6] Sugata, Y., Kido, S. and Shouno, H.: Comparison of two-dimensional with three-dimensional analyses for diffuse lung diseases from thoracic CT images, *Medical Imaging and Information Sciences*, Vol. 25, No. 3, pp. 43–47 (online), available from ⟨http://ci.nii.ac.jp/naid/130000097652/en/⟩ (2008).