Vol.2017-MPS-114 No.12
2017/7/17

びまん性肺疾患画像における特徴選択手法の比較

小野 賢志[3]　小岩井 誠[4]　庄野 逸[1,a)]　木戸 尚治[2]

概要：びまん性肺疾患は，診断に有効とされる CT 画像においても多様な陰影を示し，しばしば診断に困難をきたす場合がある．このような問題の計算器支援診断においておいては，パターン認識が使われることが多いが，パターン認識においてどのような特徴を入力として与えるのかがしばしば問題になる．本研究では，びまん性肺疾患画像における特徴選択手法にフォーカスをあて，提案されている様々な手法の比較を行った．特徴選択問題は，多種類の特徴量から，識別に都合の良い組合せを選択する，組合せ選択の問題であるため，しばしば計算困難を引き起こす．近年では，この特徴選択問題に関し，スパースモデルに基づいた最適化手法や，マルコフ連鎖モンテカルロ法を用いた列挙法を適用するケースが議論されている．そこで我々は，これらの特徴選択手法の幾つかの候補を取り上げ，比較を行った．

Comparison of Feature Selection method for Diffuse Lung Disease

1. はじめに

びまん性肺疾患とは，肺の広範囲にわたって異常な陰影が観察される病気の総称である．びまん性肺疾患の陰影は，淡い陰影パターンを呈し，広範囲にわたって観察されることが多く，X 線 CT 画像などが有効な診断手法として考えられている．びまん性肺疾患部の X 線 CT 画像は，陰影パターンが複雑多様であるため，医師の経験等によって診断が左右される場合がある．このような診断に対して計算機科学の観点から医師にサポートを行っていく計算機診断支援 (Computer Aided Diagnosys: CAD) システムの構築が望まれている [1][2][3]．特にパターン解析等の手法によってもたらされる知見は，医用従事者に対する知識のフィードバックなどが期待できる．本研究では，びまん性肺疾患画像の診断支援の為の基礎として，びまん性肺の識別システムに於ける特徴選択手法について取り扱う．機械学習の分野では，パターン識別のシステムは大枠で，特徴抽出などの前処理と，抽出した特徴の学習を行う機械の組合せが論じられることが多い．特徴抽出は画像などから識別に有効な特徴量と呼ばれる量を抽出する操作である．このため，画像診断においては各種の画像の統計量などを用いることが多い．一般に，このような特徴量を集めてきた場合すべての特徴を同時に使ったほうが，よりよい識別器ができると考えられがちであるが，識別器を構成してみると全ての特徴を用いるよりも，識別に効果的な特徴量を選んで，選択的に用いた方が，性能が高い識別器を構成するのに有利なケースが多い [2][4]．このような操作を特徴量選択と呼ぶ．

特徴量選択は，特徴量の中から有効な組合せを選ぶという組合せ最適化に属する問題であるため，特徴の量（次元数）が増えるほど，有効な選択を行うことが困難になってくる．これに対して，様々なアプローチが提案されている．古典的な手法としては，組合せを考えるのを諦めて，特徴量を一つ選んで増やしたり，減らしたりした時に性能がもっとも良い方向に変化する特徴量を逐次的に選択していくといった手法が提案されてきた．再帰的特徴除去 (Recursive Feature Elimination:RFE) などは典型的な，アプローチである．一本，近年のスパースモデリングの躍進により，スパースモデリングと識別器を組合せて，より少ない特徴から識別器を構成するという手法も提案されてきている．これには $L1$ 制約条件付きのロジスティック回帰や，サポートベクターマシン (SVM) といった識別器を構成することが，このアプローチにあたる．さらに，あらたなアプローチとして，

[1] 電気通信大学 大学院情報理工学研究科
Graduate School of Informatics and Engineering, University of Electro-Communications, Chofugaoka 1–5–1, Chofu 182–8585, Japan
[2] 山口大学 大学院応用医工学研究科
Graduate School of Applied Medical Engineering Systems, Yamaguchi University
[3] JAL インフォテック
[4] NTT コムウェア
[a)] shouno@uec.ac.jp

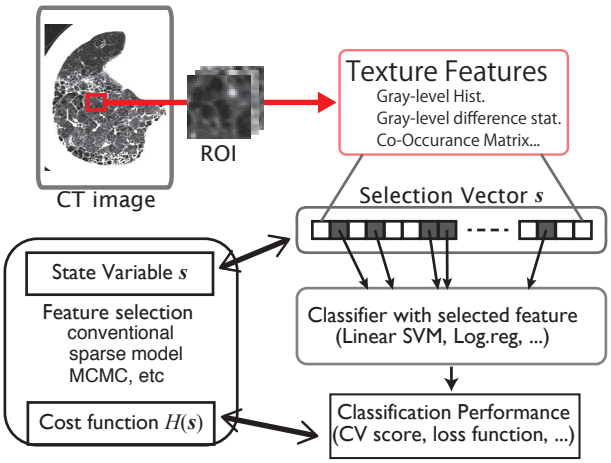

図 1: 本研究に於ける特徴選択の模式図．まず，対象である画像から特徴量を選択肢，これらの特徴から，選択すべき特徴セットを $\boldsymbol{s}$ で表すことを考える．選択ベクトル $\boldsymbol{s}$ の要素が 0 である場合は，特徴が選ばれなかったことを表す．この選択ベクトル $\boldsymbol{s}$ を与えた時の識別器による識別性能を $H(\boldsymbol{s})$ として，最適化を行う枠組みを考える．

全ての組合せを選出するかわりに，サンプリング手法を用いて解候補を列挙していくアプローチが挙げられる．この手法にはマルコフ連鎖モンテカルロ (Markov Chain Monte Carlo:MCMC) 法や，MCMC 法を拡張した温度交換 MCMC (Exchange Temprateure MCMC:ExMCMC) 法などを用いて列挙する手法などが提案されている [6][5]．特に ExMCMC 法は，脳神経活動の有効なニューロンの選択問題や，自閉症スペクトラムの検出問題などにおいて，全数列挙の手法とくらべても遜色のない解の選択を，実現可能であることが示されている [5]．そこで本研究でも，これらの特徴選択手法を微慢性肺疾患画像の識別問題に適用するために比較を行うことを試みた．

2. 特徴選択問題の定式化

ここでは特徴選択問題の定式化を行う．図 1 は，特徴選択全体の枠組みを表す．びまん性肺疾患画像の識別には，画像のテクスチャが重要であると考えられているため，ここではテクスチャ特徴と呼ばれる統計量を用いることを考える．まず，これらの特徴量 $\boldsymbol{x}$ を各画像から計算する．次に特徴量 $\boldsymbol{x}$ に対して，どの特徴を選ぶのかを決定するベクトル $\boldsymbol{s}$ を考える．$\boldsymbol{s}$ は従来手法の枠組みや，MCMC 手法などでは，どの特徴成分を選択したかといった変数にあたるため，$0-1$ の 2 値をとる変数として考えられる．一方，スパース表現手法においては，$\boldsymbol{s}$ を変数に対する重みとして線形和を計算し，これの値をもって識別関数を構成する．この場合 $\boldsymbol{s}$ に 0 成分が多くなるようにスパースな解を構成させることを考える．

2.1 従来手法

本研究内で取り扱う，従来手法は，特徴量のデータセット内で大きなブレ幅をもつものを選択する Univariate 手法と，識別に無効と考えられる特徴量を再帰的に消去していく RFE 手法を考える．Univariate 手法は，特徴選択の中でも，特に識別精度を評価にいれず，変化の無い変数を無意味な変数として削っていく，前処理的な選択手法である．一方の RFE 手法は，初期状態として全ての特徴量を選択し，$s_i = 1$ (for all i) 識別器を構成した後，ある一個の i 番目の要素を $s_i = 0$ として除去することを考える．このときの $\boldsymbol{s}$ を $\boldsymbol{s}_i$ と表すことにする．この際に，汎化誤差 $H(\boldsymbol{s}) > H(\boldsymbol{s}_i)$ であれば，i は要らない特徴となるので，除去し探索をおこなっていく．最終的に汎化誤差が小さくならなくなった時にアルゴリズムとして停止する．

2.2 スパースモデリングに基づく選択手法

スパースモデリングに基づく手法として，L1 ノルム制約を付けた識別器を考える．ここでは識別モデルとしてロジスティックシグモイド回帰 (Logistic-Sigmoid Regression: LR) による線形識別モデルと線形サポートベクターマシン (Support Vector Machine: SVM) とを対象とする．

LR を実現するコスト関数は交差エントロピー関数

$$\tilde{H}_{\text{cross}}(\boldsymbol{s}) = -\sum_p t_p \log y_p + (1-t_p)\log(1-y_p). \quad (1)$$

である．但し p はパターンの番号, y_p は p 番目のパターンに対する学習機械の出力で，t_p は学習のためのラベルである．この交差エントロピー関数に対して L1 制約を導入した

$$H_{\text{cross}}(\boldsymbol{s}) = \lambda \tilde{H}_{\text{cross}}(\boldsymbol{s}) + \sum_i |s_i|, \quad (2)$$

をコスト関数として考える．このコスト関数を $\boldsymbol{s}$ に関して最適化を行うことによって L1 スパースロジスティック回帰 (L1-SLR) を実現できる．

次にサポートベクターマシン同様に考える．は以下のようなヒンジロス関数

$$\tilde{H}_{\text{hinge}}(\boldsymbol{s}) = \sum_p \max\{1 - t_p y_p, 0\}, \quad (3)$$

に対して，同様の L1 制約を考えれば良いので

$$H_{\text{hinge}}(\boldsymbol{s}) = \lambda \tilde{H}_{\text{hinge}}(\boldsymbol{s}) + \sum_i |s_i|. \quad (4)$$

をコスト関数として考える．

2.3 マルコフ連鎖モンテカルロ法による選択手法

マルコフ連鎖モンテカルロ (Markov Chain Monte Carlo:MCMC) 法は，コスト関数をエネルギー関数とみなしたときに導出される確率分布（ボルツマン分布）からのサンプリングを行うことで解候補を列挙する手法である．MCMC 法は，前述の二つの最適化アプローチと異なり，解候補を列挙するアプローチを用いることで平均値な

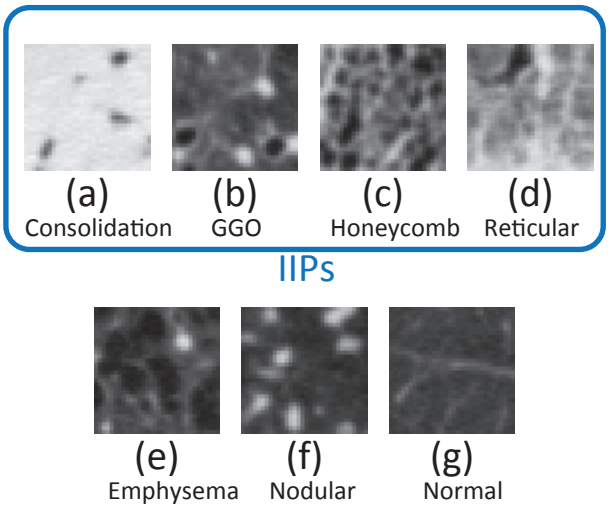

図 2: 各識別クラスに典型的な ROI 画像の例．(a) ～ (d) は，びまん性肺疾患画像を表し，(e), (f) は，それぞれ気胸と粒状影といった病変を表す．(g) は正常例を表す．

どを計算させることが可能になっている．

MCMC 法は，解候補を計算する際に，現在の状態を少しづつ変化させながら列挙していくために，局所解の周辺を探索するケースが往々に起こる．そこで，物理学的見地から温度と呼ばれるパラメータを導入し計算を行う．温度は，局所解の粒度をなますために用いられれ，高温状態であれば，より遠くの場所を大域的にランダム探索できる．その一方で，低温状態であれば，局所解の周りを重点的に探索する．温度交換　MCMC (ExMCMC) 法は，このような複数の温度状態の MCMC を並列に実行し，一定の感覚で，各温度間で状態を交換することで，大域的探索と，局所的な探索を効率的に行うような手法である [6].

Nagata は，この手法の特性を活かし，特徴選択手法に ExMCMC 法を導入している [5]．Nagata らの実験によれば，識別問題においても ExMCMC 手法は有効で，全数選択をした場合に比べてかなり少ない繰り返し回数で，実用上十分な解がえられることを示している．

3. 計算機実験

本節では，実験に用いた材料と，具体的な計算手法について述べる．

3.1 実験材料

本研究では徳島大学医学部から提供された 360 症例のイメージを用いて実験を行った．分類したいクラス数は 7 クラスで，それぞれは，浸潤影 (Consolidation: CON)，すりガラス状影 (Ground-Grass-Opacity: GGO)，蜂巣状影 (Honyecomb: HCM)，網状影 (Reticular: RET)，気胸 (Emphysema: EMP)，粒状影 (Nodular: NOD)，正常影 (Normal: NOR) である．各クラスのサンプルデータ数は CON:38 例, GGO:76 例, HCM:49 例, RET:37 例, EMP:54 例，NOD:48 例, NOR:58 例である．各クラスのデータは対象領域 (Region of Interest:ROI) を医師の指導の下，選択している．

表 1: 1 対他識別器を用いた識別精度比較

	CON	GGO	HCM	RET	EMP	NOD	NOR
Whole	1.000	0.915	0.972	0.958	0.915	0.380	0.888
Univar.	1.000	0.817	0.958	0.901	0.901	0.873	0.845
RFE	1.000	0.930	0.972	0.958	0.986	0.521	0.888
L1-SLR	1.000	0.915	0.972	0.958	0.972	0.648	0.859
L1-SVM	1.000	0.958	0.972	0.944	0.915	0.775	0.915
ExMCMC	1.000	0.939	0.989	0.980	0.967	0.925	0.981

3.2 計算機実験

特徴量としては Sugata らの特徴量を用いた [1]．この特徴量はテクスチャ画像解析一般に用いられる特徴量で，画像ピクセル値の同時生起行列 (Co-Ocurence Matrix: COM), ランレングス行列 (Run-Length Matrix:RLM), 濃淡ヒストグラム (Gray Level Histogram: GLH), ピクセル間差分量 (Gray Level Difference:GLD), フーリエパワースペクトル (Fourier Power Spectrum:FPS) を情報源として，これらの情報源から要約統計量を抽出する．ようやく統計量としては，平均，分散，などの統計量を用い，最終的に各情報源からの統計量を一列に並べることで 39 個の要素をもつベクトルを構成する．この画像から得られる 39 次元のベクトルを特徴量として取り扱う．

識別器は，2 節に述べた，従来手法，スパースモデリングに基づく手法，MCMC 法に基づく手法によって構成させる．この時の識別器の評価手法としては，交差検証法 (Cross Validation:CV) 手法を用いて，汎化誤差を計測する．計算機実験においては，学習データを 5 分割した 5-CV を用いて評価している．

また本研究で取り扱う，識別クラスは 7 クラスの分類問題であるため，ここでは識別器を一対多 (One-versus-Rest:OVR) 手法によって構成している．このため，各クラスに付随する識別器は，与えられたクラスかどうかを判定する識別機となる．

4. 実験結果

表 1 は，各手法の識別精度を，各クラスごとに算出した結果である．表中の各業は特徴選択手法を表しており，上から，変数全部を用いた (Whole) 手法, Univariate 手法, 再帰的変数除去 (RFE), L1-スパースロジスティック回帰 (L1-SLR), L1-SVM, 温度交換 MCMC (ExMCMC) 手法に基づいた結果である．この表から，クラス CON の識別問題は比較的優しい課題であるのに対し，クラス NOD の識別問題は比較的難しいことが見て取れる．特に，全ての特徴量を使ったとしても クラス NOD の識別精度は芳しくない．Univarite 手法や，RFE では，クラス GGO や，クラス NOD でやや効果的であるかのように見えるが，全体的に

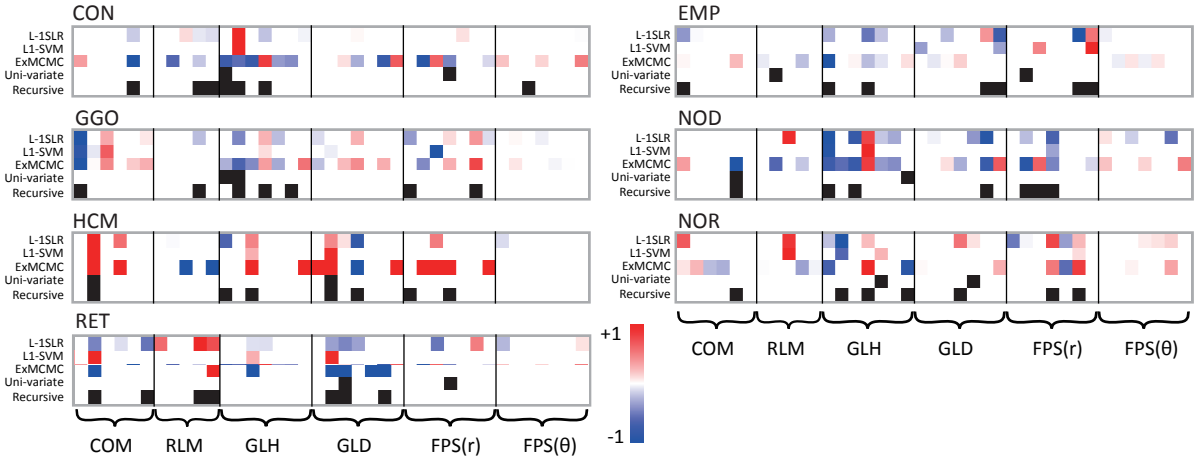

図 3: 選択された特徴量の可視化図．各クラスにおける識別に有効な特徴量を表す．各図はクラス選択に必要な特徴を表しており，線形識別器の重みを色を付けて表示している．図中赤色が正の重み，青色が負の重みを表す，重み強度が強いもの程，濃い色で表している．各図の上から，スパースモデリング手法 (L1-SLR 法，L1-SVM 法)，MCMC 法に基づく手法を表している．

は評価を全体をやや下げてしまう．これに対して，スパースモデリングの手法は，比較的良好であることが見て取れる．特に NOD の改善が，全体のパフォーマンスに影響することが見て取れる．一方，ExMCMC 手法では，ほぼ全体に渡って好成績を収めているのがわかる．特に NOD クラスや NOR クラスに関しては，他の特徴選択手法と比べて，良好な成績を収めていることがわかる．

図 3 は，各クラスごとの各特徴選択手法によって得られた選択特徴を図示かしたものである．39 種類の特徴を横方向に並べ，図中の各クラス，各特徴選択手法で選ばれた特徴は色が濃くなるような表示をしている．各クラスの上段から L1-SLR, L1-SVM, ExMCMC, Univariate, RFE 手法を表している．性能が良かった NOD クラスや NOR クラスで比較すると L1 ベースの手法は，やや特徴を絞りすぎている用に見受けられる．GGO, HCM, RET クラスを見ても，スパース手法をベースにした特徴選択 (L1-SLR や L1-SVM) は ExMCMC と同じような傾向を示しているが，やはり特徴を絞りすぎている印象を受ける．

5. まとめ

本研究では，びまん性肺疾患識別問題において変数選択手法に関する調査と議論を行った．特徴選択手法としては，従来法，スパースモデリングベースを用いた特徴選択手法，MCMC を用いた手法の比較を行った．MCMC 法を用いた選択手法が良い性能を示すことがわかった．スパースモデリングベースの手法は，比較的良い性能を示すが，やや変数を絞り込みすぎる．この絞込の調整は，式 4, 2 における λ パラメータを調整することで回避できるが，これを調整するためにさらに交差検証を行う必要が出てくるため，計算コストが増大する．このため，この交差検証手法と MCMC 法の計算コストとを比較する必要が出てくる．その一方で，MCMC 法は，比較的計算コストが高いものの，コンスタントに良好な性能を示す手法であることが理解る．

謝辞 本研究は，徳島大学医学部教授上野淳二先生にはびまん性肺疾患診断にまつわる様々な協力を頂きました．ここに感謝の意を表します．また医学データに関しましても徳島大学医学部病院の協力を頂き，感謝致します．また，本研究は科学研究費補助金，新学術領域研究 16H01542, ならびに基盤研究 (C) 16K00328 の一部からサポートを得て行われた．

参考文献

[1] Y. Sugata, S. Kido, and H. Shouno, “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, 2008. [Online]. Available: http://ci.nii.ac.jp/naid/130000097652/en/

[2] M. Wada, H. Shouno, and S. Kido, “An idiopathic interstitial pneumonia classification for ct image by use of a semi-supervised learning”, in *Intl. Forum on Medical Imaging in Asia (IFMIA)*, November 2012, pp. P1–34.

[3] H. Shouno and S. Kido, “Semi-supervised based learning for Idiopathic Interstitial Pneumonia on High Resolution CT images”, in *In Proc. PDPTA*, Jul. 2015, pp. 270–275.

[4] M. Koiwai, N. Iida, H. Shouno, and S. Kido, “Feature selection for diffuse lung disease using exchange markov chain monte-carlo method”, in *PDPTA*. PDPTA 2016, 7 2016, pp. 381–386.

[5] K. Nagata, J. Kitazono, S.-i. Nakajima, S. Eifuku, R. Tamura, and M. Okada, “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.

[6] H. Koji and N. Koji, “Exchange monte carlo method and application to spin glass simulations”, *Journal of the Physical Society of Japan*, vol. 65, no. 6, pp. 1604–1608, 1996. [Online]. Available: http://dx.doi.org/10.1143/JPSJ.65.1604