様式 C-19

# 科学研究費補助金研究成果報告書

平成 23 年 3 月 31 日現在

機関番号：１７１０２
研究種目：基盤研究（B）
研究期間：2007～2010
課題番号：１９３０００９６
研究課題名（和文）超高次元データの分類手法の導出とその理論的性質の解明および実データへの応用の研究
研究課題名（英文）Research of classification methods of hyper-spectral data, elucidation of the theoretical nature and applications to real data
研究代表者
西井 龍映（NISHII RYUEI）
九州大学・大学院数理学研究院・教授
研究者番号：４０１２７６８４

研究成果の概要（和文）：超高次元データにも適応可能な分類手法が近年求められている．そこで柔軟な判別境界を表現できて過学習となりにくい bagging 型 AdaBoost を提案した．また地球環境空間データの解析のため，空間依存性をマルコフ確率場でモデル化し，森林被覆率の判別問題や回帰問題，土地被覆割合の推定手法を考察した．なお統計モデル選択のための情報量基準についての専門書を出版した．

研究成果の概要（英文）：Recently, classification methods applicable to hyper-dimensional data are required. We proposed a bagging-type AdaBoost method, which can give complicated decision boundaries and avoid over learning. Classification methods and regression problems related to geo-spatial data, and unmixing of land-cover categories were also studied through modeling of spatial dependency by Markov random fields. Furthermore, we published a book discussing information criteria for statistical model selection.

交付決定額

（金額単位：円）

| | 直接経費 | 間接経費 | 合　計 |
|---|---|---|---|
| 2007 年度 | 2,700,000 | 810,000 | 3,510,000 |
| 2008 年度 | 2,300,000 | 690,000 | 2,990,000 |
| 2009 年度 | 2,300,000 | 690,000 | 2,990,000 |
| 2010 年度 | 2,300,000 | 690,000 | 2,990,000 |
| 年度 | | | |
| 総　計 | 9,600,000 | 2,880,000 | 12,480,000 |

研究分野：総合領域
科研費の分科・細目：情報学・統計科学
キーワード：パターン認識，学習理論，時空間現象，統計モデリング，多重分光画像

## 1．研究開始当初の背景

判別分析はアヤメの種を4次元データにより判別する Fisher の考察から始まった．判別には4次元正規分布を用いたため， 未知母数の数は総計22である．もし200次元の変数を観測したときにも同様の手法を用いたとすると 20,700 個の未知母数を推定する必要がある．莫大な数の教師データが必要になるばかりか， 大きな教師データを用いても， ある観測次元数を超えると判別効率が悪化することが知られている（次元ののろい）．

さてデータ取得・転送・蓄積技術の発達により，超高次元データ（コンビニでの POS データ，ゲノムデータ，人工衛星による地球観測データ等）が観測されるようになった．このような膨大なディジタルデータ・画像に対する新しい判別手法の展開とその実データへの応用研究が求められている．

## 2. 研究の目的

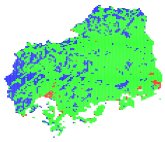
高次元データの場合にも適応可能なパターン認識の方法論の展開，およびリモートセンシングによる環境データの画像判別等の実データへの応用について考察する．考察した手法の一つが，高性能で知られる AdaBoost である．柔軟な判別境界を得るための弱学習機の選択，および過学習を避けるための Baggingの利用について考察する．さらに提案手法を多重分光画像の画像判別に応用する．また画像判別のために空間依存性をモデル化し，混合正規分布やロジスティック判別による判別手法について考察し，実データに応用することを目的とする．また森林被覆率の判別問題や非線形回帰モデリングについても考察する．

## 3. 研究の方法

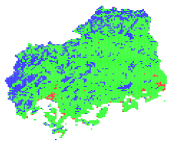
AdaBoost について，Bagging により過学習を回避し，弱判別機を工夫して柔軟な判別境界を記述できる判別手法を考察する．また選ばれた弱判別機の頻度に応じて，変数の重要度を評価する指標を導入する．また画素判別やカテゴリ比率の推定のための方法論を展開し，人口衛星や航空機から地表面を観測した多重分光画像により判別性能の評価を行う．

## 4. 研究成果

変数とその重みをランダムに与えて作成した一次結合を基底関数とするアダブーストを提案した．また過学習を回避するため，教師データを判別機の生成用と評価用にランダムに分割し，判別機を生成･評価する bagging



(a) 真のカテゴリ



(b)空間的に独立　　(c) 空間的に従属

図:広島県の森林被覆率の判別

を用いた．これにより安定した高性能の判別機を得ることに成功した．提案手法は，各変数が判別にどの程度貢献するかを数値として評価できるため，変数選択に用いることもできる．シミュレーションデータや実データで検証したところ，提案手法は高性能として知られるSVM，ANN 等を常に上回った．

さて航空機搭載のセンサーからは200次元以上の多重分光画像が観測できる．この画像にもとづいて各画素の土地被覆(カテゴリ)を推定する問題を考察した．各画素を一つのカテゴリに分類する解析が判別分析(画像判別)である．一方解像度が粗い場合には，一つの画素が複数のカテゴリで被われている．そこでカテゴリ比率を推定する問題 unmixing について考察した．

空間的に隣接している画素のカテゴリ比率ベクトルはお互いに近いものになっているはずである．この事前情報をマルコフ確率場としてモデル化した．また各カテゴリでの特徴ベクトルは正規分布に従い，複数のカテゴリで被われている場合は，混合正規分布で与えられると仮定した．このもとで局所事後密度の最大化により被覆割合の推定方式を提案した．また高次元のケースでも適応可能なロジスティック判別についても，空間依存性を取り入れて得られる手法を提案した．いずれもシミュレーションデータや実データに適応したところ，十分な性能を有することが示された．

また森林面積比率について，判別問題や回帰問題を考察した．地表面の観測領域において，森林の被覆率を人口密度と土地の起伏量(最高標高 － 最低標高)で説明する非線形回帰モデルを考察した．まず観測領域について 1) 森林が全くない， 2) 部分的に森林で被われている，3)完全に森林で被われている，の 3 つのケースについて，当該領域の人口密度および起伏量から判別するロジスティック型判別モデルを考察した．判別には周辺の観測領域の影響も取り込んだモデルが優れていることや，説明変数の巾変換が有効であること等の知見を得た．

推定したモデルから，森林が全く無くなる原因は周辺領域の影響によるものではなく当該領域で決定されること，逆に完全に森林で被われるのは周辺領域からの影響が大きいことが判明した．このように森林減少過程や地域差の考察が可能となった．

前ページの図は広島県における森林被覆率を表したもので，(a)真の 3 群，(b)空間依存性を用いない判別結果 ,(c) 空間依存性を用いた判別結果である．明らかに (c)が (a)に近いことが分かる．

その他の研究として，時間の経過にともなっ

て高頻度で観測されたデータや時空間データを，観測時点の不均一性や欠測，あるいは個体差や内在するノイズを考慮して関数化処理し，処理した関数集合に基づく非線形現象解明のためのモデリングと理論・方法論の研究に取り組んだ．さらに AIC を研究の端緒とする統計モデル選択のための情報量基準について，この分野の研究の集大成的書籍を出版した．

また高次元配列(テンソル)データの応用が画像解析、脳波解析，WEB解析など様々な分野で研究されている．テンソルデータ解析における基本的な問題として階数決定問題があり，いくつかの結果を得た．また比較的サイズの小さな3次元分割表に対して、ある時点における条件つき推論のFRAMEから次の時点における条件つき推論のFRAMEへの全射をアルゴリズム的に構成することができることを示した．

なお，本科研費と科研基盤(A):統計科学における数理的方法の理論と応用（研究代表者:谷口正信）との共催で，研究集会「高度情報抽出のための統計理論・方法論とその応用」(九大附属図書館視聴覚ホール 2008/11/20-22)を開催した．また「Symposium on Remote Sensing for Environmental Sciences」(休暇村 志賀島 2010/8/29-31) を開催した．いずれも本研究の分担者や研究協力者をはじめ海外からの招待講演者，大学や企業研究者が参加し，有意義な研究集会となった．

5．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕(計 51 件)


①P. Qin & R. Nishii, Selection of ARX models estimated by the penalized weighted least squares method, 査読有, Bulletin of Informatics and Cybernetics, Vol. 42, 2010, 35-43

②P. Qin & R. Nishii, T. Nakagawa & T. Nakamoto, ARX models for time-varying systems estimated by recursive penalized weighted least squares method, 査読有, Journal of Math-for-Industry, Vol. 2, No. 2(A), 2010, 109-114

③ J. Copas & S. Eguchi. Likelihood for statistically equivalent models, J. Royal Statistical Society B, Vol. 72, No. 2, 2010, 193-217

④ T. Abe, K. Shimizu & A. Pewsey, Symmetric unimodal models for directional data motivated by inverse stereographic projection, Journal of the Japan Statistical Society, 査読有, Vol. 40, No.1, 2010, 45-61

⑤ M. Kayano, K. Dozono & S. Konishi, Functional cluster analysis via orthonormalized Gaussian basis expansions and its application, 査読有, Journal of Classification, Vol. 27, 2010, 211-230

⑥H. Masuda, Approximate self-weighted LAD estimation of discretely observed ergodic Ornstein-Uhlenbeck processes, Electronic Journal of Statistics, Vol. 4, 2010, 525-565.

⑦ S. Tateishi, H. Matui & S. Konishi, Nonlinear regression modeling via the lasso-type regularization, 査読有, Journal of Statistical Planning and Inference, Vol. 140, 2010, 1125-1134

⑧G. Kitagawa & S. Konishi, Bias and variance reduction techniques for bootstrap information criteria, 査読有, Annals of the Institute of Statistical Mathematic, Vol. 62, 2010, 209-234

⑨T. Sumi, M. Miyazaki & T. Sakata, About the maximal rank of 3-tensors over the real and the complex number field, 査読有, Annals of the Institute of Statistical Mathematic, Vol. 62, 2010, 807-822

⑩Y. Sawamura, R. Nishii, A. Nakamoto, S. Kawaguchi and T. Ozaki, Contextual clustering and unmixing of geospatial data based on Gaussian mixture models and Markov random fields, 査読有, Bulletin of Informatics and Cybernetics, Vol. 41, 2009, 39-49

⑪ S. Tanaka and R. Nishii, Non-linear regression models to identify functional forms of deforestation in East Asia, 査読有, IEEE Transactions on Geoscience and Remote Sensing, Vol. 47, 2009, 2617-2626

⑫K. Aizawa and S. Tanaka, A constant time algorithm for finding neighbors in quadtrees, IEEE Trans. Pattern Analysis and Machine Intelligence, Vol.31, No.7, 2009, 1178-1183

⑬ S. Iacus, M. Uchida, & N. Yoshida, Parametric estimation for partially hidden diffusion processes sampled at discrete times, Stochastic Processes and their Applications, Vol. 119, No. 5, 2009, 1580-1600.

⑭Y. Araki, S. Konishi, S. Kawano & H. Matsui, Functional regression modeling via regularized Gaussian basis expansions, 査読有, Annals of the Institute of Statistical Mathematics, Vol. 61, 2009, 811-833

⑮ M. Kayano & S. Konishi, Functional principal component analysis via regularized Gaussian basis expansions and its application to unbalanced data, 査読有, Journal of Statistical Planning and Inference, Vol. 139, 2009, 2388-2398

⑯ S. Tateishi, H. Matsui and S. Konishi,

Nonlinear regression modeling via the lasso-type regularization, 査読有, Journal of Statistical Planning and Inference, Vol. 140, 2009, 1125-1134

⑰T. Sumi, M. Miyazaki & T. Sakata, Rank of 3-tensors with 2 slices and Kronecker canonical forms, 査読有, Linear Algebra and its Application, Vol. 431, 2009, 1858-1868

⑱ M. Uchida, Approximate martingale estimating functions for stochastic differential equations with small noises, Stochastic Processes and their Applications, Vol. 118, No. 9, 2008, 1706-1721

⑲ T. Tanaka, S. Torii, I. Kabuta, K. Shimizu & M. Tanaka, Pattern classification of nevus with texture analysis, 査読有, IEEJ Trans. Vol. 3, 2008, 143-150

⑳Y. Ninomiya & A. Yoshimoto, Statistical method for detecting structural change in the growth process, 査読有, Biometrics, Vol. 64, 2008, 46-53

㉑ M. Ichikawa & S. Konishi, Constructing second-order accurate confidence intervals for communalities in factor analysis, 査読有, British Journal of Mathematical and Statistical Psychology, Vol. 61, 2008, 361-378

㉒ H. Matsui, Y. Araki, S. Konishi, Multivariate regression modeling for functional data, 査読有, Journal of Data Science, Vol. 6, 2008, 313-331

㉓ K. Hirose, S. Kawano & S. Konishi, Bayesian factor analysis and information criterion, 査読有, Bulletin of Informatics and Cybernetics, Vol. 40, 2008, 75-87

㉔S.H. Ong, K. Shimizu & C.M. Ng, A class of discrete distributions arising from difference of two random variables, 査読有, Computational Statistics & Data Analysis, Vol. 52, 2008, 1490-1499

㉕S. Kawaguchi & R. Nishii, Hyperspectral image classification by Bootstrap AdaBoost with random decision stumps, 査読有, IEEE Transactions on Geoscience and Remote Sensing, Vol. 45, 2007, 3845-3851

㉖Y. Ninomiya & H. Fujisawa, A conservative test for multiple comparison based on highly correlated test statistics, 査読有, Biometrics, Vol. 63, 2007, 1135-1142

㉗T. Kanamori, T. Takenouchi, S. Eguchi & N. Murata, Robust loss functions for boosting, 査読有, Neural Computation, Vol. 19, 2007, 2183-2244

㉘ M. Henmi, R. Yoshida & S. Eguchi, Importance sampling via the estimated sampler, 査読有, Biometrika, Vol. 94, 2007, 985-991

㉙ H. Masuda, Ergodicity and exponential $¥beta$-mixing bound for multidimensional diffusions with jumps, 査読有, Stochastic Processes Appl., Vol. 117, 2007, 35-56

〔学会発表〕(計 59 件)

① R. Nishii & S. Tanaka, Statistical Deforestation modeling based on zero-one inflated distributions with spatial dependence (招待講演), Forum for Interdisciplinary Mathematics, Dec. 20, 2010, Patna University (India)

②T. Sumi & T. Sakata, The set of 3×4×4 contingency tables has 3-neighborhood property, COMPSTAT' 2010, Aug. 23, 2010, CNAM (France)

③R. Nishii and T. Ozaki, Contextual unmixing of geospatial data based on Markov random fields and conditional random fields, Whispers 2009, Aug. 28, 2009, GIPSA-LAB (France)

④ R. Nishii, T. Ozaki & Y. Sawamura, Semi-supervised contextual classification and unmixing of hyperspectral data based on mixture distributions, IGARSS 2009, July 16, 2009, Univ. of Cape Town (South Africa)

⑤T. Sakata, T. Sumi & M. Miyazaki, The evaluation of the maximal rank of tensors simply by row and column operations and symmetrization (招待講演), IASC 2008, Dec. 6, 2008, Pacifico Yokohama (Japan)

⑥S. Tanaka & R. Nishii, Non-linear regression models to identify functional forms of deforestation, IEEE IGARSS 2008, July 10, 2008, Hynes Convention Center (USA)

⑦ R. Nishii, Contextual image classification based on statistics and machine learning (招待講演), ICTS 2008, Aug. 5, 2008, Institut Teknologi Sepuluh Nopember (Indonesia)

⑧ R. Nishii, Y. Sawamura & T. Ozaki, Semi-supervised contextual unmixing of geospatial data (招待講演), IASC 2008, Dec. 7, 2008, Pacifico Yokohama (Japan)

⑨S. Tanaka and R. Nishii, Spatial logit models of deforestation due to population and relief energy in East Asia, IEEE IGARSS 2007, July 24, 2007, Centre Convencions Intl. Barcelona City (Spain)

⑩S. Eguchi, Boosting methods for association studies in bioinformatics, International Conference on Multiple Decision Theory, Statistical Inference and Applications, Dec. 28, 2007, Fu Jen Catholic University (Taipei)

〔図書〕(計 6 件)

①小西 貞則, 岩波書店, 多変量解析入門 － 線形から非線形へ －, 2010, 306
② S. Konishi & G. Kitagawa, Springer, Information Criteria and Statistical Modeling, 2008, 273
③内田 雅之, 東京大学出版会, 確率微分方程式の母数推定. 21 世紀の統計科学 III, 北川源四郎・竹村彰通 編, 2008, 179-206 (分担執筆)

6. 研究組織

(1)研究代表者
西井 龍映(NISHII RYUEI)
九州大学・数理学研究院・教授
研究者番号:40127684

(2)研究分担者
小西 貞則(KONISHI SADANORI)
中央大学・理工学部・教授
研究者番号:40090550
坂田 年男 (SAKATA TOSHIO)
九州大学・芸術工学研究院・教授
研究者番号:20117352
秦 攀(QIN PAN)
九州大学・数理学研究院・学術研究員
研究者番号:40532718

(3)連携研究者
二宮 嘉行 (NINOMIYA YOSHIYUKI)
九州大学・数理学研究院・准教授
研究者番号:50343330
増田 弘毅 (MASUDA HIROKI)
九州大学・数理学研究院・准教授
研究者番号:10380669
田中 章司郎 (TANAKA SHOJIRO)
島根大学・総合理工学部・教授
研究者番号:00197427
清水 邦夫 (SHIMIZU KUNIO)
慶應義塾大学・理工学部・教授
研究者番号:60110946
江口 真透 (EGUCHI SHINTO)
統計数理研究所・教授
研究者番号:10168776
内田 雅之 (UCHIDA MASAYUKI)
大阪大学・大学院基礎工学研究科・教授
研究者番号:70280526