様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２５年　６月１１日現在

機関番号：６２６０３
研究種目：基盤研究（Ａ）
研究期間：２００８～２０１２
課題番号：２０２４００２８
研究課題名（和文）　ゲノムデータからの予測・発見・推論の統合化のための統計学と機械学習の融合
研究課題名（英文）　Integration of Statistics and Machine Learning for Combining Prediction, Knowledge Discovery and Inference.
研究代表者
江口　真透（EGUCHI SHINTO）
統計数理研究所・数理・推論研究系・教授
研究者番号：10168776

研究成果の概要（和文）：ゲノム・オミクスデータから導かれる科学的成果を得るための統計的方法の開発を行った。特に表現形（病型、治療奏功性、予後）の予測に適切な情報を抽出するために統計学と機械学習の方法の融合的な活用を実用化に向けて推進してきた。表現形の予測のために ROC カーブの下側面積の最大化によるブースト法を開発した。乳がんサブタイプを決める有効な遺伝子選択のための LASSO クラスタリングを提案し、良好な成果が得られつつある。
研究成果の概要（英文）：Statistical methods to achieve scientific results conducted from genomics and omics data have been exploited. In particular, for prediction of phenotypes (disease type, treatment effect, prognosis) we enhanced to implement several proposals by a fusion between statistics and machine learning. The boosting method by maximization of Area Under Curve is applied to gene expression data. LASSO clustering is proposed to select genes for detecting subtypes in breast cancer patients.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００８年度	7,900,000	2,370,000	10,270,000
２００９年度	5,500,000	1,650,000	7,150,000
２０１０年度	6,000,000	1,800,000	7,800,000
２０１１年度	6,900,000	2,070,000	8,970,000
２０１２年度	6,200,000	1,860,000	8,060,000
総　計	32,500,000	9,750,000	42,250,000

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：分類・パターン認識

１．研究開始当初の背景

(1) 20 世紀の終りの頃から急速な発展を遂げたバイオテクノロジーで多種の観測技術が様々な形式のデータを蓄積することとなり、研究開始当初にはその蓄積スピードは更に加速していた。特にマイクロアレイによる遺伝子発現やプロテオームによるタンパク発現の網羅的で特異的な観測が可能になり、メタボロームによる代謝機構の観測も盛んになっていた。このように膨大なスケールで蓄積されたゲノム・オミックスデータから、有用な情報を取り出す方法論を提案し関連する諸科学の発展へ貢献することが、統計科学の重要な課題の一つに挙げられていた。
(2) 高次元データ小標本問題は、各国の有力な研究グループによって数理的な研究から実用的な研究まで幅広く活発な研究がなされていたが、未だにバイオインフォマティクスにおいて重要な遺伝子が発見されたこと

はほとんどないという極論もあった。当時の研究方向の弱点の一つに知識発見が単発のプロジェクトで成されている点が挙げられる。単発のプロジェクトといえども大規模な予算で組まれた研究では、発表成果はその新規性や独自性を強調されがちで、過去の幾つかの相同な成果との相対化、その知識の結合には力が注がれていなかった。要するにゲノム研究やオミックス研究において SNPs、マイクロアレイ、プロテオーム、メタボロームのデータは互いに強い相関性があるはずであるが、各々のデータから得られた知識や仮説を統合して知識発見の強化を図った研究は、研究開始当初において進んでいなかった。

２．研究の目的

(1)　主目標：　ゲノム研究やオミックス研究において各々の観測で帰結された結論を統合するための方法論を開発することを目指す。SNP と発現遺伝子と発現タンパクと代謝産物はゲノムから始まる転写、翻訳さらに代謝へとつながり、生殖を通して次世代のゲノムへの橋渡しを繰り返す一個のサークルになっている。このサークルの上に立ち、それぞれの観測によって得られた結果をより強めあう統計的な方法を開発したい。 この観点に立ち、ゲノム研究やオミックス研究に関する予測・発見・推論の統合化のための新たな方法を統計学と機械学習の融合によって開発することを主目標とする。特にゲノム・オミックスデータと表現形データへの相関研究において、マイクロアレイデータ、プロテオームの中から別々に開発された統計的パタン認識の方法を総合化して、より精度の高い予測・発見の方法を開発したい。 これにより機械学習と統計学の融合に寄与する新たな方向を見つける。

３．研究の方法

遺伝子情報の網羅的な計測性に対して、被験者数は少数に限られている点が大きな問題となっている。個人化医療のために治療効果を予測するキットの実用化までには解決しなければいけない多くの困難な課題が見つかっている。この課題の解決のため、ゲノム・オミクスデータの特徴である『p>>n』問題に適切に対処し、より高性能でロバストな表現型（病型、治療奏功性、予後）予測の方法の開発を以下の方法で行った。また統計学と機械学習を融合させた新しい方法の開発も並行させて行った。

(1)　問題の原因を探る：　ゲノム研究やオミックス研究で生産されるデータの特徴は、生物の複雑の機構を網羅的に特異的に観測できる点である。例えば、マイクロアレイによって測られた発現遺伝子とある疾病との関連を発見する問題を考えよう。現在では被験者の遺伝子発現が一度の実験でヒトの持つ全ての遺伝子に対して計測できるようになっている。しかし問題は結果の再現性である。これは次の理由から今後の観測のテクノロジーが開発されても本質的に回避不能な問題であると言える。

①　1 番目の理由は、遺伝子は互いにネットワークでつながれて強い相関を持っていることが想定され、関連性に有意な発現遺伝子が単一ではなく複数あると考えられていることである。しかも遺伝子ネットワークの研究は単純な遺伝子構造を持つ生物にしか適用できず、ヒトのような種内に遺伝的多様性を持つ生物には、ほとんど不可能である。

②　2 番目の理由は、「正解＝関連性に有意な発現遺伝子」が情報のない膨大なデータの中に埋没されて、見せかけの正解の発見によって真の正解を見逃すことが増えてしまうことである。これがゲノム研究やオミックス研究において結果の再現性が本質的に困難であるという理由である。更にバイオテクノロジー技術の拡大がもたらすデータ次元の巨大化によって、この困難さに拍車がかかると思われる。

(2)　上の再現性欠如の本質的な問題に対し、統計方法から機械学習の方法までありとあらゆるアプローチを結集して解決して行く。特にゲノム・オミックスデータと表現形データへの相関研究において、マイクロアレイデータ、プロテオームの中から別々に開発された統計的パタン認識の方法を総合化してより精度の高い予測・発見の方法を開発する。疾病の種類、程度や薬剤の奏功性、感受性に対する予測の方法を確立する。

(3)　個人化医療のために治療効果の予測するキットの実用化までには未だ解決しなければいけない多くの困難な課題が見つかっている。この課題の解決のキーとして医学的な知識を積極的に統計方法に取り入れる。ゲノム・オミクスデータに内在している不均一性に対して医学の新しい知見を積極的に援用して、より精度の高い予測モデルの構築を目指す。乳がんのサブタイプのクラスタリングの有効な方法を開発して、より高性能でロバストな治療予測方法の開発につなげる。

(4) 重合する仮説の集合から適合する仮説を選択すると多重性の問題が起こり、間違った見せかけの結論を導く危険性があるが、このような問題について高次元小標本の状況下で有効に働く適切な推論を提案したい。そのために相関構造のある下での多重性の調整法を考察する。特に、パタン認識で知識発見された表現形との相関研究において、治療効

果や薬剤感受性についてのパラメータの信頼領域と仮説のP値に対する標準的な方法から最悪評価による補正を導く。

４．研究成果

ゲノムデータから導かれる科学的成果を得るための統計的方法の開発を目指し、特に表現形予測に適切な情報を抽出するために統計学と機械学習の方法の融合的な活用を実用化に向けて推進してきた。個人化医療のために治療効果を予測するキットの実用化のために解決しなければいけない多くの困難な課題の解決のキーとして、医学的な知識を統計方法に積極的に取り入れた。ゲノム・オミクスデータに内在している不均一性に対して医学の新しい知見を積極的に援用して、より精度の高い予測モデルの構築を目指した。特に最終年度には、乳がんのサブタイプのクラスタリングの有効な方法を開発し、より高性能でロバストな治療予測の方法の開発につなげつつある。教師なし学習のための機械学習の方法として、創発クラスタリングの完成や、ブースティング方法による密度関数や回帰関数の推定法を実用化した。これらは表現形予測の前処理として適切な指針を与えるものとして期待される。

(1)　統計的パタン認識アルゴリズムの開発：　オミクスデータを予測変数とする表現型の予測のためのパタン認識の困難さの本質について徹底的な考察を行い、性能の良い予測キットを構築する際に大きな妨害となる点として予測法の多重解の問題を指摘した。この点について注意深く慎重な考察を行い、幾つかのオミクスデータの学習 アルゴリズムを以下のように開発した。

①　AUCBoost：　予測の性能を測るROC曲線の上側面積（AUC）を最大化するブースト法を開発した。AUCBoostの開発については、主に江口、小森によって研究を進めた。次に、表現形によって定めるクラスラベルについて、順序カテゴリーを成す場合、例えば、疾病の程度や治療の効果を表す場合についてブースティングを提案した。国立がんセンターの田村グループとの共同研究において、乳がんの治療効果を予測するためのマイクロアレイデータの解析が急速に進み上記の方法の実用化に目処がついた。また既存の方法との比較についても詳細な研究ができ、幾つかの特徴的な結果が得られている。

②　pAUCBoost：　医療の現場では擬陽性確率を一定の低い値にして正陽性確率をできるだけ高くする予測が広く受け入れられていることを勘案して、特に擬陽性確率が低い値より小さな領域に対応するROC曲線の部分下側面積の最大化について新たな機械学習の方法論を提案した。これをpAUCBoostと呼び、論文発表を行った。

③　t-Boost::　マイクロアレイによる遺伝子発現による予測問題に対して古典的な2標本検定による変数選択の問題に対して考察した。この問題に対して遺伝子選択から予測まで、一貫してｔ検定を使うことを検討している。そのためにｔ検定量をブースティングの観点から見直し、新たにｔ-Boostを発表した。

④　Lasso Clustering：　がん研究所と松浦正明研究グループとの共同研究で表現系の中に複合的なサブクラスが内在することが示唆され、このクラス内の異型性が予測の信頼度を低下させていることが特定された。この研究をもとに乳がんサブタイプのクラスタリングについて研究した。L1正則化を用いたLassoタイプの統計的方法によって遺伝子選択を行いながらサブタイプ分類を行うクラスタリングをK平均法をベースに開発した。これを遺伝子発現の実データ解析に適用しサブタイプの予測に有効な方法として実用化した。

⑤　Gamma-clustering：　K平均法など従来のクラスタリング法は予めクラスター数を決めておかないと実装はできない。この従来法の欠点を改良する自動的にクラスター法を決めるモード探索型クラスター法をガンマーベキエントロピーの援用によって提案した。理論的な考察とべきガンマーの選択についても簡便な方法が提案された。

(2) 機械学習と統計学の融合：ブースティング法、カーネル法について統計的な観点から幾つかの新しい方法を提案した。

①　U-Boost density estimator：　U-エントロピーを用いて密度推定のためのブースティング法を考案して論文発表を行った。

②　Robust kernel PCA：　カーネル法による従来の主成分分析に対してロバストな性能を持つ方法を提案した。関数データに対しても良好な性能を持つことが実データ解析より示された。

③　UAUC-Boost：　上記のAUCBoostやpAUCBoostをより一般的な形で定義したブースティングアルゴリズムを提案した。これにより従来の幾つかの方法との関連が明らかにされた。

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 21 件）以下すべて査読有

① P. Chen, H. Hung, O. Komori, S-Y. Huang, and S. Eguchi, Robust Independent Component Analysis via Minimum γ-Divergence Estimation, To appear in IEEE □
② K. Naito and S. Eguchi, Density estimation with minimization of U-divergence, Machine Learning 90, (2013) 29-57
③ T. Takenouchi, O. Komori and S. Eguchi, An extension of the Receiver Operating Characteristic curve and AUC-optimal classification, Neural Computation, 24, 10 (2012) 2789-2824
④ O. Komori and S. Eguchi, Boosting learning algorithm for pattern recognition and beyond, E94-D, 10 (2011) 1863-1869 DOI 10.1587/transinf.E94.D.1863
⑤ S. Eguchi, O. Komori and S. Kato, Projective power entropy and maximum Tsallis entropy distributions, Entropy, 13 (2011) 1746-1764 DOI 10.3390/e13101746
⑥ O. Komori and S. Eguchi, A boosting method for maximizing the partial area under the ROC curve, BMC Bioinformatics, 11:314 (2010) online
⑦ S. Eguchi and S. Kato, Entropy and divergence associated with power function and the statistical application, Entropy 12 (2010) 262-274
⑧ J. Copas and S. Eguchi, Likelihood for statistically equivalent models, J. Royal Statistical Society B, 72 (2010) 193-217
⑨ N. H. Mollah, M. Pritchard, O. Komori and S. Eguchi, Robust hierarchical clustering for gene expression data analysis, Communications of SIWN, 6 (2009) 118-122
⑩ プリチャード真理, 江口真透, 関連遺伝子セットの多重解の存在, 日本統計学会誌, (シリーズ J 2 号), 38 (2009) 199－212
⑪ H. Fujisawa, Y. Horiuchi, Y. Harushima, T. Takada, S. Eguchi, 他 4 名, SNEP: Simultaneous detection of nucleotide and expression polymorphisms using Affymetrix GeneChip, BMC Bioinformatics, 10:131 (2009)
⑫ S-Y. Huang, Y-R. Yeh, and S. Eguchi, Robust kernel principal component analysis, Neural Computation 21 (2009) 3179-3213
⑬ T. Takenouchi, Robust Boosting Algorithm Against Mislabeling in Multiclass Problems, Neural Computation, 20 (2008) 1596-1630
⑭ H. Fujisawa, S. Eguchi, Robust parameter estimation with a small bias against heavy contamination, Journal of Multivariate Analysis, 99 (2008) 2053-2081
⑮ M. Kawakita, S. Eguchi, Boosting method for local learning in statistical pattern recognition, Neural Computation, 20 (2008) 2792-283
⑯ S. Eguchi, Asymptotical improvement of maximum likelihood estimators on Kullback-Leibler loss, Journal of Statistical Planning and Inference, 138 (2008) 3502-3511

〔学会発表〕（計 38 件）

① 江口真透，遺伝子発現による表現形の予測のための統計的方法，久留米大学バイオ統計センター10 周年記念第 11 回久留米大学バイオ統計学フォーラム, 2013.1.19，福岡県久留米市
② O. Komori, S. Eguchi and J. Copas, Maximization of the generalized t-statistics for two-class discrimination problem, Statistical Seminar, 2012.10.15, Academia Sinica, Taipei, Taiwan
③ 小森 理，江口真透，コーパス ジョン, t 統計量の一般化とその判別解析への応用，統計関連学会連合大会，2012.9.10, 札幌，北海道
④ O. Komori, S. Eguchi and J. Copas, Maximization of a Generalized t-Statistic for Two-Class Discrimination Problem, XXVI th International Biometric Conference, 2012.8.28, Kobe, Japan
⑤ S. Eguchi, A. Notsu and O. Komori, Projective power cross entropy and detectability for hidden structures: part II, International Workshop on Anomalous Statistics, Generalized Entropies, and Information Geometry, 奈良女子大学，2012.3.9 (招待講演)
⑥ S. Eguchi, Maximization of a generalized t statistic for linear discrimination in the two group classification problem, 2012.1, Centre

for Research in Statistical Methodology, University of Warwick, (Invited talk)

⑦ S. Eguchi, Projective power entropy and detectability for hidden structures, Workshop on Contemporary Statistics, 2011.12.20, Taida Institute for Mathematical Sciences, National Taiwan University (Invited talk)

⑧ S. Eguchi, O. Komori, A. Notsu, Projective power entropy based learning for unsupervised data, Joint Meeting of 2011 Taipei International Statistical Symposium, 2011.12.17, Academia Sinica, Taiwan (Invited talk)

⑨ O. Komori, S. Eguchi, A Statistical Method for the Partial Area under the ROC Curve, 25th International Biometric Conference, 2010.12.07, Florianopolis, Brazil

⑩ S. Eguchi, U-entropy and maximum entropy model, Information Geometry and its Applications III, 2010.08.02, Leipzig, Germany（招待講演）

⑪ S. Eguchi, Boosting finite mixture model, International Conference on Robust Statistics 2010, 2010.06.29, Prague, Czech Republic（招待講演）

⑫ O. Komori and S. Eguchi, Pattern recognition from genome and omics data, Tutorial Workshop on Learning with Information Divergence Geometry, 2010.04.25, National Taiwan University, Taipei, Taiwan（招待講演）

⑬ S. Eguchi and O. Komori, Boosting leaning algorithm and U-loss functions I and II, Tutorial Workshop on Learning with Information Divergence Geometry, 2010.04.25, National Taiwan University, Taipei, Taiwan（招待講演）

⑭ S. Eguchi and O. Komori, Information geometry on model uncertainty, Tutorial Workshop on Learning with Information Divergence Geometry, 2010. 04. 24, National Taiwan University, Taipei, Taiwan（招待講演）

⑮ S. Eguchi and O. Komori, Information divergence class and robust statistical methods I,II, Tutorial Workshop on Learning with Information Divergence Geometry, 2010.04.24, National Taiwan University, Taipei, Taiwan（招待講演）

⑯ S. Eguchi, Maximizing t-values for all functions of a feature vector, Workshop on Geometric and Algebraic Statistics, 2009.07.13,Milton Keynes, U.K. (Invited Talk)

⑰ S. Eguchi, Projective Tsallis Entropy and its Application to Robust Statistics, Mathematical Aspects of Generalized Entropies and their Applications, 2009.07.08, Kyoto, Japan (Invited talk)

⑱ S. Eguchi, Boosting true positive and false positive rates for pattern recognition, Institute of Mathematical Statistics Asia Pacific Rim 1st Meeting, 2009.06.29, Seoul, South Korea (Invited Talk)

⑲ S. Eguchi, Information divergence geometry and its application to machine learning, The Fifth Statistics and Machine Learning Workshop, 2009.04.28, Tainan, Taiwan (Invited Talk)

⑳ 江口 真透, タンパク質構造と進化と情報幾何，数理研短期共同研究集会「離散力学系の分子細胞生物学への応用数理」, 2009. 1. 8，京都

㉑ O. Komori, A Boosting Method for Maximizing the Partial Area under the ROC Curve, International Association for Statistical Computing 2008, 2008.12.7, Yokohama, Japan

㉒ S. Eguchi, On the bound of statistical inference for observational data, International Association for Statistical Computing 2008, 2008.12.6, Yokohama, Japan

㉓ 江口 真透, バイオインフォマティクスにおける統計的課題について，科研費研究集会「高次元データの統計解析」, 2008.11.21，博多，福岡

㉔ 江口 真透, 累積分布関数のエントロピーと順序ラベルの判別への応用，統計関連学会連合大会，2008.9.9，横浜，神奈川

㉕ 江口 真透, モデル不確定性と不完全観測バイアス，統計関連学会連合大会，2008.9.9，横浜，神奈川，(受賞記念講演)

㉖ 小森 理，1クラスラベルに注目したブースティング，統計関連学会連合大会，2008.9.8，横浜，神奈川

㉗ S. Eguchi, Information divergence geometry and its application to machine learning, The 1st MSJ-SI, Probabilistic Approach to Geometry, 2008.8.4, Kyoto, Japan

〔図書〕（計 1 件）

① S. Eguchi, Eds. Frank Emmert-Streib Information Theory and Statistical Learning, Springer, New York 2008

〔産業財産権〕
○出願状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等
http://www.ism.ac.jp/~eguchi/

６．研究組織
(1)研究代表者

江口　真透　（EGUCHI　SHINTO）
統計数理研究所・数理・推論研究系・教授
研究者番号：10168776

(2)研究分担者

栗木　　哲　（KURIKI　SATOSHI）
統計数理研究所・数理・推論研究系・教授
研究者番号：90195545
（平成 22 年度まで）

藤澤　洋徳　（FUJISAWA　HIRONORI）
統計数理研究所・数理・推論研究系・准教授
研究者番号：00301177

逸見　昌之　（HENMI　MASAYUKI）
統計数理研究所・数理・推論研究系・准教授
研究者番号：80465921

松浦　正明　（MASAAKI　MATSUURA）
（公財）がん研究会・ゲノムセンター情報解析部門・部門長
研究者番号：40173794

間野　修平　（MANO　SHUHEI）
統計数理研究所・数理・推論研究系・准教授
研究者番号：20372948
（平成 23 年度から）

小森　　理　（KOMORI　OSAMU）
統計数理研究所・統計思考院・特任助教
研究者番号：60586379
（平成 22 年度から）

(3)連携研究者

竹之内　高志（TAKENOUCHI TAKASHI）
はこだて未来大学・複雑系情報学科・准教授
研究者番号：50403340

川喜田　雅則（KAWAKITA MASANORI）
九州大学・システム情報科学研究科・助教
研究者番号：90435496