様式 C-7-2

自己評価報告書

平成　23 年　5 月　6 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：(1)ゲノムデータ(2)予測(3)機械学習(4)パタン認識(5)高次元小標本(6) ROC 曲線(7)マイクロアレイ(8)ブースティング

１．研究計画の概要

本研究は，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の開発を目指す。実際には，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と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。このような観点から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に，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目指す。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研究分野の創造を目指す.

２．研究の進捗状況

以下のように 3 つの細目に分けられる.
(1) 統計的パタン認識：ゲノムデータに基づく表現系の予測スコアーを構成するために更に考察を加え、実用化に向けて幾つかの検討をした。特に ROC (Receiver Operational Characteristic) 曲線の下側面積の最大化について改良を加えた。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ものである。また、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た。
(2) メタ解析の援用：メタ解析ついては逸見助教（統数研）・J. B. Copas 教授（ウォーリック大）との長期に渡る研究を基礎にする。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,このような問題を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。Copas 教授を 2008 年，2010 年に招聘して活発なディスカッションに基づく共同研究を行った.
(3) これらの統計的な方法論の開発を通して、国立がんセンターの田村グループと乳がん治療の効果予測のための共同研究を推進した。これらの中から最も実用性の高いモデルを近々に得られる検証用のデータによって決定するプロジェクトが順調に進められている。江口・藤澤と松浦チームが共同開発した乳癌患者の抗癌剤の治療効果を予測するための共通ピーク法を本データに適用した。三菱化学との共同研究も血液情報から脳梗塞の予後予測のプロジェクトも 22 年度から小森理が中心となって活発に遂行されている.

３．現在までの達成度

②おおむね順調に進展している
ゲノムデータの膨大な情報から予測・発見・推論の統合が具体的な目的であるが，その結果，統計学と機械学習の融合による新たな研究分野の構築に寄与することが最終的な目標であり，前節で説明されたように着実な進

捗が成されている．しかしながら，実際のゲノム科学の著しい発展によって更なる統計的思考を拡大する必要がある．現時点では，個人化医療のためにゲノムデータによる予測モデルは実用化まで更に長い道のりが必要と思われる．具体的には以下のように推進方策を提示したい．

４．今後の研究の推進方策

現実のゲノム科学から大量にかつ，次世代シーケンサーなどに代用されるように新しい形式のデータが産出されてきている．これによって適切な予測・発見・推論を行うことは更に困難さが増してきている．このような現状を下で特に深刻な問題である「高い性能を示す予測モデルが構築されても異なる施設のデータセットでテストをしたときに必ずしも再現しない」ことに集中して研究したい．理由の一つとして被験者の表現形に関するデータの背景に起因する．多くの場合が観察データであり，臨床的な背景の異質性，医師の治療の変更などから生じる選択性バイアスが問題となる．共同研究で扱われている疾病は乳がん，脳梗塞などの重度の症例であり，薬剤の重篤な副作用のリスクもあるので，ランダムデザインは不可能である．このような問題は検証的な問題の中では活発に考察されてきているが，予測・発見の探索的な問題の中ではあまり議論されていない．探索と検証をつなぐための統計方法が予測モデルの発見と再現性の強化のためには必要になってきている．今後，知識発見ベースのアプローチの過程の中で，その予測性能の再現性を高める方法論の開発に重点を置いてプロジェクトを推進したい．

５． 代表的な研究成果

（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８件）

(1) A boosting method for maximizing the partial area under the ROC curve. O. Komori and S. Eguchi. BMC Bioinformatics (2010) 11:314. 査読有
(2) Entropy and divergence associated with power function and the statistical application. S. Eguchi and S. Kato. Entropy 12, 2 (2010) 262-274. 査読有
(3) Robust extraction of local structures by the minimum beta-divergence method. N. H. Mollah, N. Sultana, M. Minami and S. Eguchi. Neural Networks 23, 2 (2010) 226-238. 査読有
(4) Likelihood for statistically equivalent models. J. Copas and S. Eguchi. J. Royal Statistical Society B, 72, 2 (2010) 193-217. 査読有
(5) Robust kernel principal component analysis. S-Y. Huang, Y-R. Yeh and S. Eguchi. Neural Computation, 21, 11 (2009) 3179-3213. 査読有
(6) Boosting method for local learning in statistical pattern recognition. M. Kawakita and S. Eguchi. Neural Computation, 20, 11 (2008) 2792-2838. 査読有
(7) Robust parameter estimation with a small bias against heavy contamination. H. Fujisawa and S. Eguchi. J. Multivariate Analysis, 99, 9 (2008) 2053-2081. 査読有
(8) Robust boosting algorithm against mislabeling in multi-class problems. T. Takenouchi, S. Eguchi, N. Murata and T. Kanamori. Neural Computation 20, 6 (2008) 1596-1630. 査読有

〔学会発表〕（計 10 件）

(1) Eguchi, S. and Komori, O. Tutorial lecture on Learning with Information Divergence Geometry , Taipei , Taiwan , 2010.04.24-25
(2) Eguchi, S. , U-entropy and maximum entropy model, Information Geometry and its Applications Ⅲ, Leipzig, Germany, 2010.08.02
(3) Komori, O.* and Eguchi, S. , A Statistical Method for the Partial Area under the ROC Curve, 25th International Biometric Conference , Florianopolis , Brazil, 2010.12.07
(4) 江口 真透. ゲノム関連データを解析するための新しい統計方法と機械学習の方法. 日本計量生物学会，チュートリアルセミナー 11 月 2009 年.
(5) Eguchi, S., Maximizing t-values for all functions of a feature vector. Workshop on Geometric and Algebraic Statistics, 2009.7.13 at The Open University, Milton Keynes.
(6) Eguchi, S. and Osamu Komori, Boosting true positive and false positive rates for pattern recognition. IMS-APRM, Seoul, 28 June-2 July
(7) Eguchi, S., Projective Tsallis Entropy and its Application to Robust Statistics. Kyoto RIMS workshop on `Mathematical Aspects of Generalized Entropies and their Applications', July 7-9, 2009 held at Kyoto TERRSA
(8) Eguchi, S., Information Divergence Geometry and its application to machine learning. The Fifth Statistics and Machine Learning Workshop, 27-28 April, 2009 at National Cheng Kung University, Tainan