# 科学研究費助成事業　　研究成果報告書

平成　２６　年　　５　月　２４　日現在

機関番号：　１３９０３
研究種目：　若手研究(B)
研究期間：　2011～2013
課題番号：　２３７００１６５

研究課題名（和文）機械学習による超高次元データの統計的マッチングとマイクロアレイ解析への応用

研究課題名（英文）Machine-learning based statistical data matching and its application to microarray data analysis

研究代表者
竹内　一郎（Ichiro, Takeuchi）

名古屋工業大学・工学（系）研究科（研究院）・准教授

研究者番号：４０３３５１４６

交付決定額（研究期間全体）：（直接経費）　　3,300,000円、（間接経費）　　990,000円

研究成果の概要（和文）：本研究では、マイクロアレイデータなどに代表される網羅的遺伝情報から抽出した情報の統計的信頼性を評価するための多変量二標本検定のアルゴリズム構築と実装を行った。開発したアルゴリズムは機械学習技術に基づいており，高速に抽出された知識の信頼性を計算することができる。本研究の成果は、生命科学研究の推進に有益である。

研究成果の概要（英文）：We developed an algorithm and a software that can be used for evaluating the credibility of the knowledge taken from genome-wide biological data such as DNA expression microarray. Our algorithm is based on machine learning technology and it can compute the credibility of the knowledge efficiently. Our results are useful for evidence-based bio-medical studies.

研究分野：　機械学習

科研費の分科・細目：　知能情報学

キーワード：　machine learning　bioinformatics　statistics

## １．研究開始当初の背景

生命科学分野ではマイクロアレイ発現解析データに代表されるような網羅的遺伝情報の計測技術が発展し、数千～数万次元の超高次元データが得られるようになった．これらの網羅的遺伝情報を有効活用することは、生命科学研究の推進や個別化医療の発展に重要な役割を果たすと考えられている．

網羅的遺伝情報を情報科学・統計科学の技術を用いて解析する際には、これらの超高次元データを比較し、違いを定量化することが必要となる．高次元データの類似度を統計的に評価するには多変量検定を利用することができるが、従来の多変量検定法はたかだか数十次元のものを想定したものであったり、特定の分布形（例えば、多変量正規分布）を仮定したものが多く、超高次元の網羅的遺伝情報の処理に用いるのは適切でない場合があった．

網羅的遺伝情報は、超高次元データであるとともに複数の因子が複雑な依存関係を有している．したがって、網羅的遺伝情報の性質に即した多変量検定の方法論を構築することが急務となっている．

図 1 は、本研究で取り組む課題を模式的に表したものである．マイクロアレイ発現解析データのような超高次元データをマッチングし、統計的に類似度を評価する．

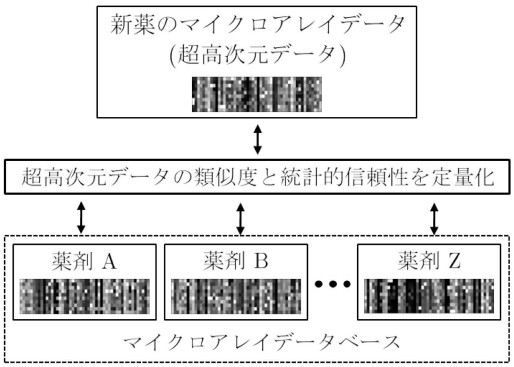


図 1. 本応募課題で考察する問題の基本概念図(マイクロアレイ解析の例): 超高次元データをデータベースとマッチングし, 類似度およびその統計的信頼性を定量化する.

## ２．研究の目的

本研究では、高次元データ処理のために開発された様々な機械学習アルゴリズムを高次元データの統計的マッチングに利用することにより上述の課題を解決することである．

機械学習アルゴリズムは、データの分布を特に仮定する必要がない、超高次元のデータを扱うことができる、といった特長を持つ．しかしながら、機械学習研究の主な標的は予測問題であり、統計的信頼性を定量化する枠組みはほとんど議論されてこなかった．本研究では、予測問題のために開発された機械学習アルゴリズムを統計的データマッチングに利用するための方法論を開発し、その成果を生命科学のデータ解析に利用することである．

図 2 は、本研究のアプローチを模式的に表したものである．機械学習分野では数多くの２クラス分類アルゴリズムが開発されている．２つの高次元データを比較する問題を２クラス分類問題の分類精度で評価することができる．しかし、予測を目的とした従来の２クラス分類問題とは異なり、分類精度の評価尺度が統計的に有意かどうかを検証する必要がある．

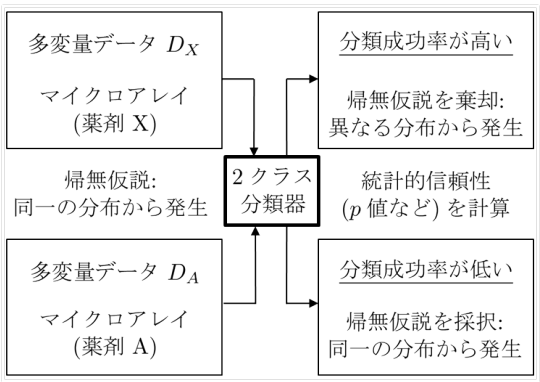


図 2. 機械学習を用いた多変量 2 標本検定の基本概念図: 多変量二標本検定とは, 多変量データ DX, および, DA が与えられたとき, これらが同一の多変量分布から発生したのか否かを統計的に定量化する問題である.

## ３．研究の方法

本研究では、機械学習分野で開発されたアプローチを多変量検定に利用するためのアルゴリズムと理論を構築することである．研究の方法は大きく、以下の 3 つのステップから構成される．

(1) 機械学習アルゴリズムを利用した検定統計量に関する理論的・実験的考察を行う．機械学習分野で研究が進んでいる分類アルゴリズムを用いると多変量データからクラスの違いを特徴づける要素を検出し、分類することができる．本研究では、機械学習アルゴリズムの特長を活かし、分類器の分類性能を検定統計量として利用するかどうかを検討した．具体的な分類アルゴリズムとしては、多くの分野で高性能な分類を行えることが保証されているサポートベクトルマシン（Support Vector Machine）を利用した．

(2) 検定統計量の帰無分布を推定するための理論的・実験的考察を行った．統計的信頼性を評価するため、帰無分布を推定する必要があるため、クラスラベルをランダムシャッフルしたデータに対する分類器を繰り返し学習し、検定統計量の帰無分布を推定するアプローチを検討した．ラベル並べ替え演算のアルゴリズム構築、帰無分布の推定精度に関する理論解析を行った．ラベル並べ替えによる帰無分布の推定では、数多くの（1000～10000 個程度）の２クラス分類問題を解かなくてはならない．サポートベクトルマシンなどの２クラス分類問題は凸最適化問題として定式化されるため、これは、数多くの凸最適化問題を解かなければならないことを意味する．本研究では、帰無分布の推定を効率的に行うための方法論を開発した．

(3) 超高次元データの統計的マッチング技術をマイクロアレイ解析へ適用した．遺伝子群解析と呼ばれるマイクロアレイデータ解析のタスクに上記の方法論を適用した．遺伝子群解析とは、特定の遺伝子のグループを同定する問題で、多重多変量検定問題として定式化されるものである．従来の遺伝子群解析に比べ、検出力が高いことを確認した．また、本研究で構築する統計的マッチング法は、ゲノムコピー数の異常領域を同定する問題にも利用することができる．リンパ腫や白血病を含む血液の癌患者から採取されたゲノムコピー数の網羅的データ解析に本手法を適用した．

## ４．研究成果

網羅的遺伝情報のための多変量検定統計量として３つのアルゴリズム（最近傍分類誤差に基づくもの、サポートベクトルマシン分類器の分類誤差に基づくもの,サポートベクトルマシン分類器の平均マージンに基づくもの）それぞれにおいて、検出力、及び、帰無分布の（ラベル並べ替え検定による）推定の計算効率を比較した．

また、サポートベクトルマシンの学習誤差（目的関数値）を検定統計量として利用できるかについても考察した．サポートベクトルマシンの学習は二次計画凸最適化問題として定式化できるため、主双対ギャップの概念を利用すれば、並べ替えサンプルにおける最適化問題を途中で打ち切ることによる計算の効率化が可能なことがわかった（例えば、並べ替えサンプルの双対目的関数値が並べ替え前のサンプルの目的関数値を上回っていれば打ち切りを行なっても p 値の計算が可能となる）．

さらに、パラメトリック計画法と呼ばれる数理最適化アルゴリズムを用いると、並べ替えられたサンプルに対する最適解をより効率的に解くことができることを確認した．提案法をマイクロアレイ発現解析データとゲノムコピー数異常データに適用した．

## 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 2 件）


① M. Karasuyama and I. Takeuchi, Nonlinear regularization path for quadratic loss support vector machines, IEEE Transactions on Neural Networks, vol. 22, pp. 1513-1625, 2011.

② M. Karasuyama, N. Harada, M. Sugiyama and I. Takeuchi, Multi-parametric solution-path algorithm for instance-weighted support vector machines, Machine Learning, vol. 88, pp. 297-330, 2012.


〔学会発表〕（計 8 件）


① I. Takeuchi and M. Sugiyama, Target neighbor consistent feature weighting for nearest neighbor classification, 25th Annual Conference on Neural Information Processing Systems (NIPS2011), 2011 年 12 月.

② 小川晃平, 竹内一郎, 杉山将, パラメトリック計画法を用いた S3VM の最適化手法に関する一考察, 電子情報通信学会 IBISML 研究会, 2012 年 6 月.

③ 石原直樹, 久留美里織, 竹内一郎, パラメトリック計画法を用いたマルチインスタンス SVM, 電子情報通信学会 IBISML 研究会, 2012 年 11 月.

④ M. Sugiyama, T. Kanamori, T. Suzuki, M. Plessis, S. Liu and I. Takeuchi, Density-difference estimation, 26th Annual Conference on Neural Information Processing Systems (NIPS2012), 2012 年 12 月.

⑤ K. Ogawa, I. Imamura, I. Takeuchi and M. Sugiyama, Infinitesimal annealing for training semi-supervised support vector machines, The 30th International Conference on Machine Learning (ICML2013), 2013 年 6 月.

⑥ K. Ogawa, Y. Suzuki and I. Takeuchi, Safe screening of non-support vectors in pathwise SVM computation, The 30th

International Conference on Machine Learning (ICML2013), 2013年6月.

⑦ I. Takeuchi, T. Hongo, M. Sugiyama and S. Nakajima, Parametric task learning, The 27th Annual Conference on Neural Information Processing Systems (NIPS2013), 2013年12月.

⑧ S. Nakajima, A. Takeda, S. D. Babacan, M. Sugiyama and I. Takeuchi, Global solver and its efficient approximation for variational Bayesian low-rank subspace clustering, The 27th Annual Conference on Neural Information Processing Systems (NIPS2013), 2013年12月.

〔図書〕（計0件）

〔産業財産権〕

○出願状況（計0件）

○取得状況（計 0件）

〔その他〕

特になし

## ６．研究組織

(1)研究代表者
竹内一郎(TAKEUCHI, Ichiro)
名古屋工業大学・工学研究科・准教授
研究者番号：40335146