様式 C-19

科学研究費補助金研究成果報告書

平成２１年５月１５日現在

研究種目：若手研究（B）
研究期間：２００６　～　２００８
課題番号：１８７０９００６
研究課題名（和文）　ブースティング法とカーネル法の統合と遺伝子データ解析への応用

研究課題名（英文）　Synthesis of boosting methods and kernel machines and its application to genomic data analysis

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

研究成果の概要：
　ブースティング法とカーネルマシンを統合して両法の特徴を持つ方法の開発と遺伝子データへの応用を行った。　最初にブースティング法とカーネルマシンを弱学習機カーネルという概念を通じて統一的に解釈した。また両者を組み合わせた判別法をいくつか開発した。それらの理論的な性能評価と実データの上での性能評価を行った。また情報幾何による解析でブースティングによる改悪について調べた。また上で開発した判別法により癌研究会より提供された遺伝子発現データの解析を行った。

交付額

（金額単位：円）

	直接経費	間接経費	合　計
2006 年度	1,100,000	0	1,100,000
2007 年度	600,000	0	600,000
2008 年度	500,000	150,000	650,000
年度			
年度			
総　計	2,200,000	150,000	2,350,000

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：　ブースティング法、カーネル法、遺伝子発現データ

１．研究開始当初の背景

(1) ブースティング法とカーネル法について

　ブースティング法とカーネル法は 1990 年代後半から 2000 年代にかけて多くの注目を集めた強力な判別法である。開発された当初、両者は以下のように異なる背景から提案されたものであった。ブースティング法は精度が悪い判別機を多数組み合わせることで精度のよい判別機を構成する方法である。一方でカーネル法の発端となったSVMはVapnikの原理に基づいて尤度の推定をすることなく、尤度比を直接カーネルトリックとマージン最大化に基づいて推定する方法である。このように一見背景が異なるように見えるため、従来両手法は別々に研究が行われてきた。両手法の関係性について知られている重要な結果は SVM と同様にブースティングもマージン最大化を行っていると解釈できることである。この関係性は理論的な性能評価に関するものであることに注意する。当研究及び近年の研究により両者の関連性はより深く直接的に示されることになる。

(2) 遺伝子データ解析について
　従来同時に発現を測定できる遺伝子数は数十のオーダーであった。しかし近年マイクロアレイなどの技術開発により、数万の単位の遺伝子についてその発現量を同時に測定することができるようになった。このように遺伝子数、つまり共変量の次元がかなり大きくなるのに比して、その一方で測定できる検体数（標本数）は必ずしも増えているとは限らない。この理由として例えば測定費用の高価さ、測定にかかる労力、また測定対象の検体数そのものが少ない、などが挙げられる。共変量の数 p より標本数 n が著しく少ないデータは従来の統計手法の適用が困難であり、このような状況を n<<p 問題と呼ぶ。マイクロアレイデータの解析は典型的な n<<p 問題であり活発に研究されている。

２．研究の目的
　当研究の目的はブースティングとカーネルマシンという異なる二つの強力な判別法を統合することにより、両者の長所をあわせもつような手法を開発することである。またその性能評価を理論的、実験的に明らかにすることを目指す。さらに開発した方法を遺伝子発現データに適用し、表現形に関連する遺伝子の発見を目指すことが目的である。

３．研究の方法
(1) ブースティング法とカーネル法の統合及び性能解析について
　ブースティングとカーネル法をカーネル指数型分布族の観点で統一的に解釈することを試みた。この目的のために弱学習機カーネル関数という概念を導入した。この弱学習機カーネル関数を通じてブースティングの弱いモデルを組み合わせるというアイデアと、カーネルトリック及び正則化を兼ね備えた判別法の開発を検討した。また理論的な性能評価の方法として PAC(probably approximately correct)学習を用いた解析と、情報幾何を用いた解析を行った。
(2) 遺伝子データ解析への応用について
　　どんなデータを用いてどんな方法を開発したいかを書く。

４．研究成果
(1) ブースティング法とカーネル法の統合について
①．弱学習機カーネルの導入とカーネル指数型分布族による統一的解釈　ブースティング法とカーネル法を統一的に解釈する枠組みを与えた。ブースティング法とカーネル法は背景で述べたように共に異なる定義のマージンを最大化する方法とみなせることが知られている。当研究ではより直接的な関係を導くために離散弱学習機カーネル関数を導入した。この弱学習機カーネル関数の定義と近いものが別の研究者によって既に提案されていたが、我々の定義は弱学習機の集合の上に定義される任意の確率分布（以下パラメータ分布と呼ぶ）を導入したより一般的な定義になっていることに注意する。弱学習機カーネルを通すとブースティング法とカーネル法は以下のように関連付けることができる。すなわち、ある弱学習機のセットを用いたブースティング法が達成できる判別関数の全体は、その弱学習機のセットから構成できる弱学習機カーネルを用いたカーネル法が達成できる判別関数の全体と等しくなる。このことはブースティング法における弱いモデルをカーネル法に持ち込むことができることを示している。また両者をさらに統一的な視点でみるためにカーネル指数型分布族 (canu and smola 2006)の視点で両者を書き下した。このように関連を除いた両者の違いは最適化の方法と正則化の有無のみといえる。すなわち普通のブースティング法は関数勾配降下法を用いており正則化は行わないのに対して、カーネル法は二次計画法などの最適化法を用いており、さらに通常正則化を行う点が異なるだけである。最適化法の違いは必ずしも本質的な差を生じないが、正則化の有無は本質的な違いである。なぜならオーバーフィットを防ぐためにブースティング法は弱いモデルから徐々に強化していくという戦略をとるのに対して、カーネル法は強いモデル（カーネル）を選び正則化を行うことでモデルを弱めるという考え方だからである。しかし後で述べるように正則化を用いたとしても強すぎるモデルはオーバーフィットするし、弱いモデルを用いても正則化が必要な状況も多い。当研究では以下でカーネル法とブースティング法を統合して弱いモデルを用いてかつ正則化を行う方法を提案した。
②．連続弱学習機カーネルの提案　離散弱学習機カーネルをカーネル法に適用すれば弱学習機をカーネル法に導入することができる。しかし弱学習機の数が膨大になれば計算量も膨大になる。ここでは離散弱学習機カーネルの定義を連続的にパラメトライズされた学習機の場合への拡張を行った。結果としてブースティング法でよく用いられる弱学習機である、決定スタンプ、線形判別機、深さ d の決定木から導かれるカーネル関数を導出した。興味深いことに決定スタンプから一様なパラメータ分布を仮定して導かれたカーネルは既存の三角カーネルと同じ形になったが、三角カーネルを用いたカーネル法とと決定スタンプを用いたブースティングが判別関数のモデルとしては等価となることは著者の知る限り知られていないはずであり、驚きである。

③．正則化ブースティング法の提案　ブースティング法の各ステップにおいて表現定理を用いてカーネル法のように正則化を行うアルゴリズムを提案した。このとき前節で提案した連続弱学習機カーネルを用いると計算量を削減できる。またすでに知られている導出法に基づいて汎化誤差の上界を与えた。また実データを用いた解析により従来の様々な方法よりしばしばよい汎化誤差を達成することを示した。
④．ブースティングカーネルマシンの提案　前節の考え方を拡張してカーネルマシンのカーネルをブースティングするアルゴリズム（以下 BKM と記す）を提案した。結果としてこのアルゴリズムの用いるモデルはうえで提案した正則化ブースティングよりもさらに弱いモデルを用いることに相当する。従ってよりオーバーフィットに耐性を持つ。また単一特徴量に基づく弱学習機を用いたブースティング法が変数選択の機能をもったように、BKM も同様に変数選択の機能を持つ。下ではこの性質を用いて遺伝子選択を行う。
⑤．ブースティングによる改悪　ブースティングの汎化誤差を予測分布の視点から解析した。ただし弱学習機の集合をパラメータについて微分可能なモデル全体と仮定した。このときブースティングは弱学習機のモデルの異なる二点を結ぶ e 測地線によってモデルの外に飛び出すことで改良を行っていると解釈できる。従って e 平坦なモデルはブースティングによって改良されないことがわかる。モデルから汎化誤差を最小にする意味での飛び出し方向とその量は komaki (1996)により導出されている。この解析をブースティングに適用するために確率から非負値測度の空間で行った結果として以下のことを明らかにした。もし弱学習機モデルがすでに真の判別関数を含んでいるならば、ブースティングを行うことで komaki (1996)とほぼ同じ量だけ（厳密にはわずかに異なる）改悪する方向に飛び出す。すなわち弱学習機として強いモデルを使うことはオーバーフィットを起こして平均的には精度が悪化することがわかる。
⑥．局所ブースティングとダブルカーネルブースティング　弱学習機が弱いことが望ましいが、弱すぎれば近似誤差が大きく精度が悪化する。ここでは局所尤度法のアイデアを用いてブースティングの局所化を行った。従来の局所尤度法は計算量の観点から高次元のデータへの適用が困難であったが、ブースティングと組み合わせることでこの計算可能となっていることが重要である。理論的な評価としてベイズリスク一致性を証明した。また UCI データセットの上での有効性を確認した。この方法はカーネル法の観点から見ればダブルカーネル（カーネルの積もまたカーネルとなる）を用いていることに相当する。すなわちカーネル法で同じモデルを達成するには一つを離散弱学習機カーネルとして、もう一つをガウスカーネルとしたことに相当する。
(2) 遺伝子データへの適用について
①．kvs の提案　副産物としてブースティングとクロスバリデーションを組み合わせた変数選択法を提案した。従来の情報量基準による変数選択などは n<<p 問題を持つデータには適用できない。弱学習機を適切に選択したブースティングは高次元の場合でも変数選択が可能である。また変数選択のオーバーフィットを防ぐためにクロスバリデーションと組み合わせることを考えた。しかしクロスバリデーションによる汎化誤差の推定量は高い分散を持つ。これを回避するために k-fold クロスバリデーションの k 個の推定量のばらつきが一定値以下の条件のもとでの推定値のみを用いて変数選択を行うのが提案手法 kvs である。この方法の有効性は今後の興味ある課題である。
②．遺伝子発現データの解析　高悪性度神経・内分泌系肺癌(HGNT)には小細胞性肺癌(SCLC)と大細胞性神経・内分泌系肺癌(LCNEC)の二種類がある。癌研究会より提供された 25 名 HGNT 患者について測定されたマイクロアレイによる遺伝子発現データを用いて、上述の癌腫の判別及び判別に有効な遺伝子を探索した。BKM 及び OCV を適用した結果いくつかの有効と思われる遺伝子の候補を特定した。今後はこの研究計画の枠組みを越えて、新たに独立に測定された検証用のマイクロアレイデータの上での有効性の確認、及び癌研究会の共同研究者とともに病理学的な調査が予定されている。

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 1 件）

① Kawakita, M. and Eguchi, S. "Boosting method for local learning in statistical pattern recognition", Neural Computation, vol. 20, issue 11, 2008　査読有り

〔学会発表〕（計９件）

① **江口　真透, 川喜田　雅則 "アレイデータ解析からの統計学の新しい方向", 2006 年度統計関連学会連合大会, 東北大学, 9 月, 2006**

② 川喜田 雅則 "カーネルマシンとしてのブースティング", 2006 年度統計関連学会連合大会, 東北大学, 9 月, 2006

③ Kawakita, M., Eguchi, S. "Boosting method for local learning", the Institute of Statistical Mathematics, Research Memorandum, No. 1005, 13 September, 2006.

④ 川喜田 雅則 "カーネルマシンとしてのブースティング法", 第 9 回 情報論的学習理論ワークショップ (IBIS), 大阪大学中之島センター, 10月, 2006

⑤ 川喜田 雅則 "機械学習による高次元データにおける変数選択法", 2007年度統計関連学会連合大会, 神戸大学, 9 月 6-9 日, 2007

⑥ Kawakita, M. "Study of gene selection based on machine learning on Microarray data", Pacific Simposium of Biocomputing 2008, the Fairmont Orchid, the Big Island of Hawaii, Hawaii, 2008.

⑦ Kawakita, M. and Eguchi, S. "Boosting method for local learning in statistical pattern recognition", Neural Computation, vol. 20, issue 11, 2008.

⑧ 川喜田 雅則, 竹内 純一 "予測分布の観点からのブースティングの性能解析", 2008 年度統計関連学会連合大会, 慶應義塾大学, 9 月 7-10 日, 2008

⑨ 川喜田 雅則, 竹内 純一 "ブースティングによる改悪について", IBIS 2008, 仙台国際センター (仙台市), 10月29－31日, 2008.

６．研究組織

(1)研究代表者

川喜田　雅則（KAWAKITA MASANORI）

九州大学・大学院システム情報科学研究院・助教

研究者番号：９０４３５４９６

(2)研究分担者

なし　　（　　）

研究者番号：

(3)連携研究者

なし　　（　　）

研究者番号：