3C-02

スパコンを用いた判別分析における判別関数の最適化とその問題点

石井　一夫[†]　　小林　拓嗣[‡]　　古崎　利紀[††]　　山形　洋平[‡‡]
東京農工大学　農学府農学部[†]　　東京農工大学　連合農学研究科[‡]
東京農工大学　農学府農学部[††]　　東京農工大学　連合農学研究科[‡‡]

はじめに

筆者らは、次世代シーケンサーやマイクロアレイなどの多次元データを用いて、判別分析を用いた判別関数による表現型予測を実施してきた。判別関数の説明変数が増加するにしたがって、その組合せは指数関数的に増えるため、総当たり法又は、無作為抽出による（いわゆるブートストラップ法）最適化を並列分散処理にて行ってきた。その過程で、判別関数を最適化する際に、ランク落ちや多重共線性の問題でエラーが起こり、その対応が必要になることが頻発した。今回、次世代シーケンサーやマイクロアレイによるそれらの特性の差、データの反復数の変化によるエラーの起こりやすさなど、判別分析による生物学的データの分析を扱う際の問題点についてまとめる。

１、背景

近年の技術進歩により、マイクロアレイや次世代シーケンサーなど、大量の説明変数を持つデータを用いて、多種類（高次元）の生物学的情報を扱って分析することが可能になってきた。これに伴い、我々は従来のクラスター分析などのデータの分類したりする教師なし学習以外に、複数の説明変数を組み合わせて数理モデルを作成し、生物の表現型を予測し、臨床診断や農作物の生育予測に用いるという試みを実施してきた[1)]。遺伝子発現定量分析においては、多数の遺伝子の発現量を組合せて、その表現型を予測する。定量的データに対しては重回帰分析を、カテゴリカルなデータに対しては判別分析やサポートベクトルマシン、ナイーブベイズ分類器などの分類方法を用いて、識別を検討してきた。

Optimization of discriminant function in discriminant analysis with HPC and its characterization
†Kazuo Ishii・Tokyo University of Agriculture and Technology
‡Takuji Kobayashi・Tokyo University of Agriculture and Technology
††Toshinori Kozaki・Tokyo University of Agriculture and Technology
‡‡Yohei Yamagata・Tokyo University of Agriculture and Technology

２、統計学的有意差検定の限界

次世代シーケンサーやマイクロアレイなどは多数の説明変数候補を持つの多次元データであるため、識別指標にどの説明変数を用いるかを選択することが一つの課題である。従来的な t 検定などを用いて p 値を求め統計学的な有意差検定により説明変数を選択することも検討してきた。しかし、次世代シーケンサーやマイクロアレイなどの場合、説明変数の候補は数百から多い時は数千個にもなる。これらの説明変数を組合わせて総当たりで判別分析による識別を行い、感度及び特異度を評価すると、必ずしも t 検定において p 値の低かった説明変数が、識別に効果的ではないことがわかってきた。

３、総当り法とモンテカルロ法による検討

したがって、表現型の識別を行うのに必要な説明変数の絞り込みにはやはり、多数の説明変数の組合せにより総当たりで識別を行って評価する必要があると考えられた。しかし、現実に全ての組合せを検討することは不可能である。それを克服するために、説明変数が少なく計算が可能な場合（２個ないし、３個）は総当りで、説明変数が多く総当り計算が不可能な場合は、モンテカルロ法（ブートストラップ）による無作為抽出により組合せを抽出して、識別性能を判定することを試みた。これらは計算量が多く、通常の計算機環境では困難であるので、メニーコア CPUs のスパコン（HP Integrity Superdome X, 12 TB メモリ、240 コア CPUs）により並列処理により計算を行った（表１、図１）。　今回、大腸がんの１８検体（SRR975551-SRR975568; SRA）、健常者の１８検体（SRR975569-SRR975586; SRA）[2)]の次世代シーケンサーデータを用いてその効果を検討した。具体的には、上記の次世代シーケンサーデータをダウンロードし、これをヒトの cDNA データ[3)]に対してマッピングした。この各データから RPKM を求め、大腸がんと健常者の間で有意差のあった 39254 個の発現バリアントを選択した（両者で、ステューデント t 検定による FDR が 0．05 以下のもの）。この 39254 個の中から、２個、３個、４

個と説明変数の組み合わせを作成し、判別分析を行い、感度、特異度、およびウィルクスラムダを計算して評価した（表１）。例えば、39254 個の中からの２個の説明変数の組合せは 770418631 通りある、スパコンを用いてすべての組合せで判別分析を行い最適の組合せをウィルクスラムダを指標に選択した。39254 個の中からの２個の説明変数の組合せは $1.01x10^{13}$ ありこれについてはすべてを検討することは不可能であるので、モンテカルロ法（ブートストラップ）による無作為抽出により組合せを抽出し、スパコンによる並列計算により最適化を行った。（この、モンテカルロ法（ブートストラップ）と並列計算を組合せた数式モデルの最適化法を PMC-ML（Parallel Monte Carlo Machine Learning）法と名づけた。）表１および図１からわかるように、この方法による近似的最適化により、ステューデント t 検定でトップにランクされた説明変数の組合せによる判別分析よりも、はるかに高感度、高特異度で、低いウィルクスラムダを示す組合せが見出された。この方法により、迅速な数理モデルの最適化が可能であることが示された。

表１　スパコンを用いた総当り法とモンテカルロ法による判別分析の最適化

39254 variants	モンテカルロ法＋並列処理による近似的最適化					t検定のp値の上位から順に使用		
変数の数	最適化された感度	最適化された特異度	最適化されたウィルクスラムダ	組み合わせの数	計算数	感度	特異度	ウィルクスラムダ
2	100	100	0.174751126	770418631	25000000	83.33	100	0.211519
3	100	100	0.13968498	1.01E+13	25000000	83.33	100	0.2112648
4	100	100	0.109800966	9.89E+16	25000000	88.89	100	0.2019975
5	100	100	0.110306313	7.76E+20	25000000	88.89	100	0.2019975
6	100	100	0.087722262	5.08E+24	25000000	88.89	100	0.2019748
7	100	100	0.087623086	2.85E+28	25000000	88.89	100	0.1686441
8	100	100	0.081694177	1.40E+32	25000000	99.44	100	0.158924
9	100	100	0.054680448	6.09E+35	25000000	99.44	100	0.1587317
10	100	100	0.061702636	2.39E+39	25000000	99.44	100	0.1578518
11	100	100	0.05506847	8.53E+42	25000000	100	100	0.1567522

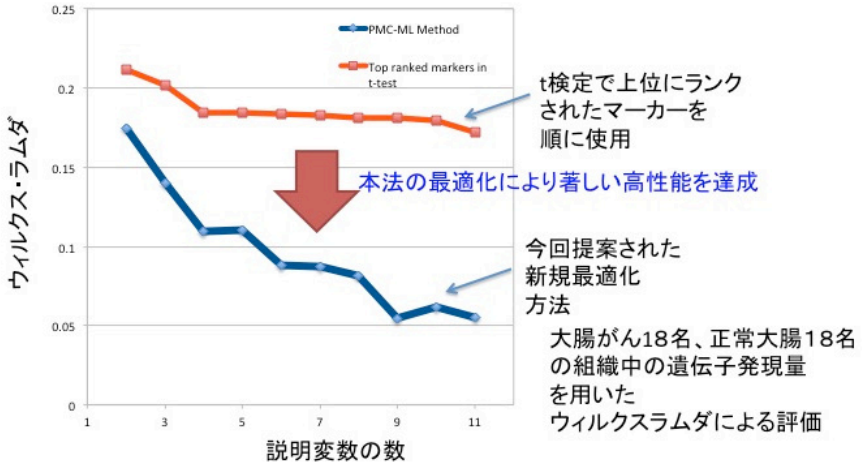

図１　スパコンを用いた総当り法とモンテカルロ法による判別分析の最適化

４、考察と顕在化してきた問題点

上記方法により、変数最適化に関するいろいろな検討が可能になった。これらの検討過程でいろいろな問題点が顕在化してきたので以下に列挙する。

１）次世代シーケンサーデータは離散型データであり、マイクロアレイは連続型データである。このため次世代シーケンサーデータの解析に置いて、検体数やリード数が少ない場合は、多変量解析の計算過程でランク落ちのエラーが生じやすい。したがって、判別分析などの計算を行う場合は、ランク落ちが起こるデータはあらかじめ除いておく必要がある。

２）同様の理由から、次世代シーケンサーデータの解析に置いて、検体数やリード数が少ない場合は、多変量解析の計算過程でや多重共線性のエラーが生じやすい。したがって、計算を行う場合に正則化処理をしておく必要がある。

３）上記エラーを避けるためには、特に次世代シーケンサーの場合は、多くの検体数やリード数を必要とし、エラーを克服するためにはより多くの計算を必要とすることが明らかとなった。

今後、これらの問題点を克服するための検討が必要となる。

文献など

１）石井一夫、沼田周助、木下　誠、大森哲郎ビッグデータ分析による精神神経系疾患診断系の検討、日本計算機統計学会 第 28 回大会特別セッション「ビッグデータ特別セッション」、東京（2014）

２）SRA http://www.ncbi.nlm.nih.gov/sra

３）Homo_sapiens.GRCh38.cdna.abinitio.fa.gz ftp://ftp.ensembl.org/pub/release-83/fasta/homo_sapiens/cdna/