Vol.2020-BIO-64 No.10
2020/12/7

テンソル分解を用いた教師無し学習による変数選択法による miRNA/mRNA/プロテオームの統合解析

田口 善弘[1,a)]

概要：マルチオミックスデータの解析はいろいろ難しい問題があり、簡単ではない。ここではマルチオミックス解析パッケージである DIABLO のテストデータ（mRNA,microRNA, プロテオーム）を用いて、テンソル分解を用いた教師なし学習による変数選択法のパフォーマンスをデモンストレーションし、遥かに高速で初期値に依らない同等の精度の結果を出せることを示す。

1. はじめに

近年、マルチオミックスデータの計測が広く行われるようになってきた。従来から広く計測されてきたゲノムや DNA のメチル化、遺伝子発現プロファイルに加えて、ヒストン修飾、非コード RNA 発現量、転写因子などの DNA への結合、クロマチンの構造、さらに最近は RNA の修飾など、計測されるマルチオミックスデータの種類数は増えることはあっても減ることはない。

一方で、これらの多様なオミックスデータを統合的に解析することには困難が多い。それは以下の様な理由による

(1) マルチオミックスデータは次元数が大きく異なる。遺伝子の発現プロファイルの場合は次元数は遺伝子の数なので $\sim 10^4$（人間の場合)、非コード RNA、例えば、microRNA では $\sim 10^3$、一方で DNA の修飾などではゲノム長 $\sim 10^9$ になってしまうことも多い。これらをただ並列に扱ったのでは次元数の低いものがほぼ無視される。しかし、オミックスの種類ごとに同じ重みにすると、今度は DNA の修飾などは一箇所あたりの重みが microRNA 一個の発現量に比べてほぼ無視されることになってしまうので加減が難しい。

(2) オミックスデータごとのダイナミックレンジが大きく異なる。DNA の修飾などでは有無（0 か 1）の値しか基本取ら無いが、遺伝子発現プロファイルは対数をとってもガウス分布しているほどダイナミックレンジが大きい。この場合、例えば、値を単純に正規化した場合にはダイナミックレンジが少ないものはほぼ無視されてしまう

この様な問題を解決するのは簡単ではない。異なったオミックスデータをどう重み付けして統合解析するのが最適であるかが不明なため、そこに人間の恣意が入ってしまう。しかし、なんの重みもつけないと、上記の様な問題を回避できない。本研究ではこの問題を回避するためにテンソル分解を使うこと [3] を提案する。

2. 材料と方法

2.1 mRNA,miRNA, プロテオームの発現プロファイル

マルチオミックスデータとしては、DIABLO パッケージ [4] に付随しているテストデータを用いた。このデータは１５０個のサンプル（３種類の培養細胞、各々 Basal：４５個、 Her2：３０個、 LumA：７５個）に対し、２００種類の mRNA、１８４種類の microRNA、１４２種類のプロテオームが計測されたデータが提供されてる。

2.2 テンソル分解を用いた教師なし学習による変数選択

mRNA の発現プロファイルは $x_{i_1 j} \in \mathbb{R}^{200\times 150}$、mRNA の発現プロファイルは $x_{i_2 j} \in \mathbb{R}^{184\times 150}$、mRNA の発現プロファイルは $x_{i_3 j} \in \mathbb{R}^{142\times 150}$ という行列の形式で提供されているものとする。これらから、以下の形式でテンソルを作成する。$x_{i_1 i_2 i_3 j} \in \mathbb{R}^{200\times 184\times 142\times 150}$。$x_{i_1 i_2 i_3 j}$ に Higher Order Singular Value Decoposision (HOSVD) [3] を適用してテンソル分解

$$x_{i_1 i_2 i_3 j} = \sum_{\ell_1=1}^{200}\sum_{\ell_2=1}^{184}\sum_{\ell_3=1}^{142}\sum_{\ell_4=1}^{150} G(\ell_1\ell_2\ell_3\ell_4) u_{\ell_1 i_1} u_{\ell_2 i_2} u_{\ell_3 i_3} u_{\ell_4 j} \quad (1)$$

を得る。ここで $G(\ell_1\ell_2\ell_3\ell_4) \in \mathbb{R}^{200\times 184\times 142\times 150}$ はコアテンソル、$u_{\ell_1 i_1} \in \mathbb{R}^{200\times 200}, u_{\ell_2 i_2} \in \mathbb{R}^{184\times 184}, u_{\ell_3 i_3} \in$

[1] 中央大学
Chuo University
[a)] tag@granular.com

本研究は国際会議 ICIC2019 のプロシーディングの一部として刊行済みである [1], [2]。

$\mathbb{R}^{142\times142}, u_{\ell_4 j} \in \mathbb{R}^{150\times150}$ は特異値行列で、全て直交行列である。

テンソル分解を用いた教師なし学習による変数選択法 [3] では、まず、3 種類の培養細胞間で差がある特異値ベクトル $\boldsymbol{u}_{\ell_4}$ を選び、次に、mRNA,microRNA, プロテオームを選択するための特異値ベクトル $\boldsymbol{u}_{\ell_1}, \boldsymbol{u}_{\ell_2}, \boldsymbol{u}_{\ell_3}$ を選ぶために、選択した ℓ_4 に対して $G(\ell_1, \ell_2, \ell_3, \ell_4)$ の絶対値が大きくなるような ℓ_1, ℓ_2, ℓ_3 を選択する。

最後に、選択された特異値ベクトル $u_{\ell_1 i_1}, u_{\ell_2 i_2}, u_{\ell_3 i_3}$ がガウス分布することを仮定して、χ 二乗分布を用いて i_1 番目の mRNA、i_2 番目の microRNA、i_3 番目のプロテオームに P 値を

$$P_{i_1} = P_{\chi^2}\left[> \sum_{\ell_1} \left(\frac{u_{\ell_1 i_1}}{\sigma_{\ell_1}} \right)^2 \right] \tag{2}$$

$$P_{i_2} = P_{\chi^2}\left[> \sum_{\ell_2} \left(\frac{u_{\ell_2 i_2}}{\sigma_{\ell_2}} \right)^2 \right] \tag{3}$$

$$P_{i_3} = P_{\chi^2}\left[> \sum_{\ell_3} \left(\frac{u_{\ell_3 i_3}}{\sigma_{\ell_3}} \right)^2 \right] \tag{4}$$

という式で付与する。但し、$P_{\chi^2}[> x]$ は引数が x 以上になる場合の χ 二乗分布の累積確率であり、$\sigma_{\ell_1}, \sigma_{\ell_2}, \sigma_{\ell_3}$ は標準偏差である。

2.3 判別分析

前節で選んだ $\boldsymbol{u}_{\ell_4}$ を用いて 3 種類の培養細胞の線形判別分析を行う。ツールは R [5] の MASS パッケージに入っている `lda` 関数を用いる。`prior=rep(1/3,3)` をつけることで含まれるサンプル数が異なる 3 種類の培養細胞を同じ重みで扱い、`CV=T` をつけることで交差検定の方法として Leave One Out Cross Validation を選択する。

3. 結果

3.1 3 種類の培養細胞の判別

図 1 は u_{1j} と u_{4j} による、１５０サンプルの散布図である。3 種類の培養細胞が綺麗に別れていることが解る。これが完全な教師なし学習であり、クラスターの個数や、個々のサンプルのラベルの情報を用いていないことを考えると非常によい結果であると言えるだろう。同時に、第一、第二特異値ベクトルではなく、第一、第四特異値ベクトルの散布図に 3 種類の培養細胞のクラスターが反映していることから、mRNA,microRNA、プロテオームの 3 種類の培養細胞間での差が、必ずしもドミナントではないことを示している。第 2、第 3 特異値ベクトルが何を表現しているのかは不明である。**表 1** は線形判別の結果である。誤差はわずかに 5 %程度であり、ほぼ完璧に判別が出来ている。

3.2 mRBA,microRNA, プロテオーム選択

前節では、3 種類のオミックスデータ（mRNA,microRNA, プロテオーム）の統合解析によって 3 種類の培養細胞を高い精度で判別できる 2 つの特異値ベクトルが教師なし学習で構成できることを示した。しかし、生物学的な問に答えるには「どの」mRNA,microRNA, プロテオームが主に 3 種類の培養細胞の間で差があるのかを知ることもまた重要である。テンソル分解を用いた教師なし学習による変数選択法で、この様なことが可能だろうか？

これを調べるために、材料と方法で述べた方法で、mRNA,microRNA, プロテオームにそれぞれ P 値を付与して、上位（P 値が小さい方）から１０個ずつ選択することを試みた。これを実行するには $G(\ell_1\ell_2\ell_3 1)$ と $G(\ell_1\ell_2\ell_3 4)$ の絶対値が大きい ℓ_1, ℓ_2, ℓ_3 を特定する必要がある。**表 2** は $G(\ell_1\ell_2\ell_3 1)$ と $G(\ell_1\ell_2\ell_3 4)$ の絶対値が大きい上位１０位までを表にした。ほぼ $1 \le \ell_1, \ell_2 \le 2, 1 \le \ell_3 \le 4$ しか出現していないことが解る。従って、P 値の付与にこれらの ℓ_1, ℓ_2, ℓ_3 に対応する $\boldsymbol{u}_{\ell_1}, \boldsymbol{u}_{\ell_2}, \boldsymbol{u}_{\ell_3}$ を用いることとした。

図 2 は選ばれた３０種類の mRNA,microRNA, プロテオームのヒートマップである。培養細胞（行）はこれらの３０種類のオミックスデータだけでも十分に判別が出来、

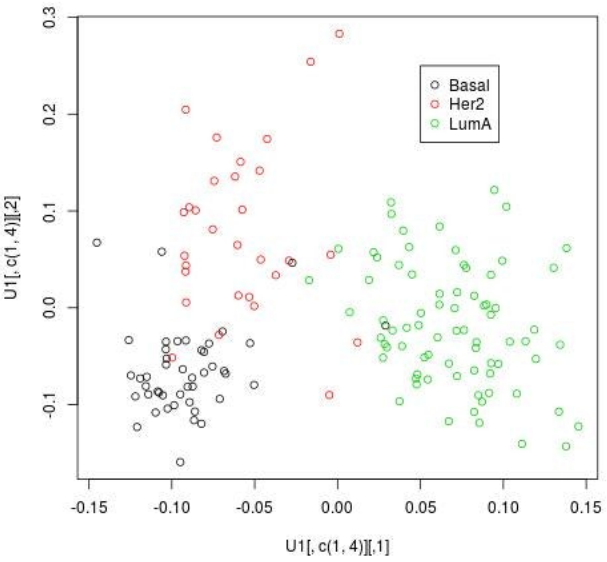

図 1 u_{1j}（横軸）と u_{4j}（縦軸）による、１５０サンプルの散布図。
Fig. 1 Scatter plots of 150 samples using u_{1j}(horizontal axis) and u_{4j}(vertical axis).

表 1 u_{1j} と u_{4j} を用いた 3 種類の培養細胞の判別分析の結果。列：培養細胞、行：予測。
Table 1 Results of linear discriminant analysis of three cell lines using u_{1j} and u_{4j}. Columns: cell lines, rows: predictions.

	Basal	Her2	LumA
Basal	42	4	0
Her2	2	25	2
LumA	1	1	73

表 2　$G(\ell_1\ell_2\ell_3 1)$ と $G(\ell_1\ell_2\ell_3 4)$
Table 2　$G(\ell_1\ell_2\ell_3 1)$ and $G(\ell_1\ell_2\ell_3 4)$

rank	ℓ_1	ℓ_2	ℓ_3	ℓ_4	$G(\ell_1\ell_2\ell_3\ell_4)$
1	1	1	1	1	-407857.582
2	1	1	4	4	-209720.615
3	2	1	1	4	-20452.480
4	2	1	3	1	-11677.505
5	2	1	4	1	-10428.742
6	2	1	2	1	10157.467
7	1	1	2	1	-8973.774
8	1	2	1	4	8360.976
9	2	1	5	4	-6628.467
10	1	1	3	4	6623.046

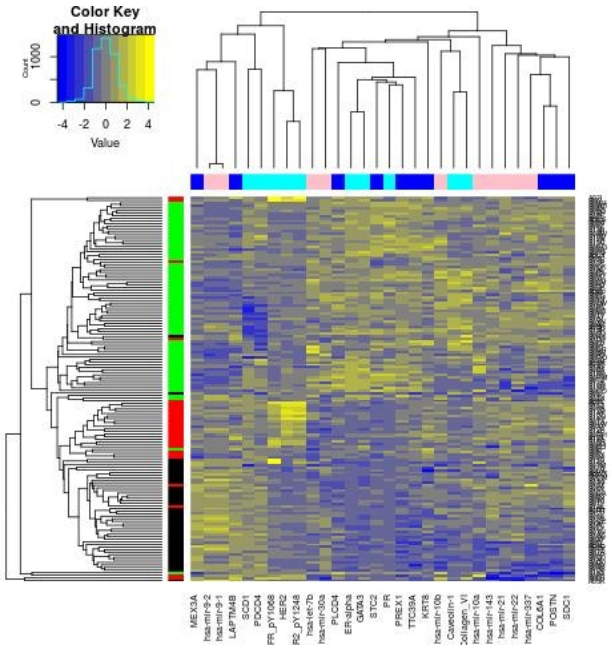

図 2　mRNA,microRNA, プロテオームのヒートマップ。行：サンプル（培養細胞）、列：青が mRNA, ピンクが miRNA, 水色がプロテオーム。
Fig. 2　Heatmap of mRNA, microRNA and proteome. Rows: samples (cell lines), columns: mRNA (blue), microRNA (pink) and proteome (cyan).

また、異なったオミックスデータ間で同じような培養細胞の種類依存性をもったものが選ばれていることもわかる。このことから、テンソル分解を用いた教師なし学習による変数選択法は、オミックスの選択にも有効に使えることが解る。

3.3　DIABLO との比較

次に DIABLO との比較を行う。DIABLO は作り込まれてはいるものの、典型的な教師あり学習の判別装置兼変数選択装置である。ほぼパラメーターフリーのテンソル分解を用いた教師なし学習による変数選択法と非常に対象的である。３つのオミックスデータの統合方法も単純に掛けるだけのテンソル分解を用いた教師なし学習による変数選択法とは異なり、どのような統合をするかをモデルベースで人間が指定しなくてはならず、詳述はさけるが非常に多数のモデルが想定しうる。ここでは、実行例ページ（http://mixomics.org/mixdiablo/case-study-tcga/）にあるモデルの計算結果との比較を行う。

まず、最初の判別性能の比較を行う。**図 3** は DIABLO に

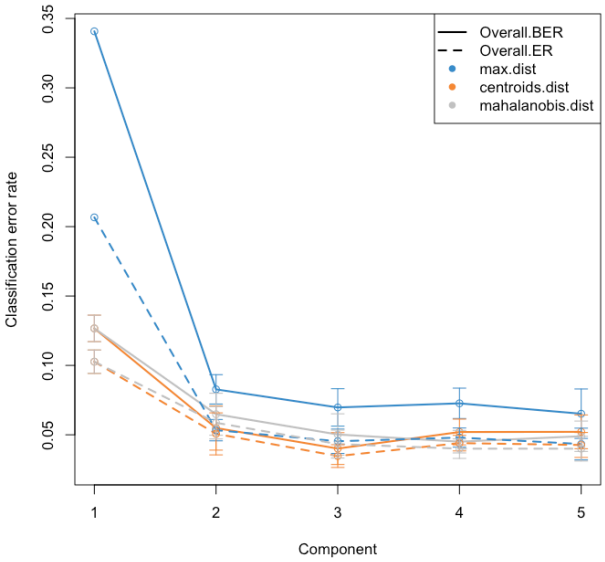

図 3　DIABLO による３種の培養細胞の判別のエラー率。横軸は判別に使用した合成ベクトル数。
Fig. 3　Error rates of discrmination of three cell lines by DIABLO. Horizontral axis: the number of generated vectors used for discrmination.

よる３種類の培養細胞の判別のエラー率である。DIABLO もテンソル分解を用いた教師なし学習による変数選択法の特異値ベクトルのような合成ベクトルを構成し、その空間内で判別を行うのは同じである。非常に興味深いことに DIABLO もまた２本の合成ベクトル、つまり、平面で３種類の培養細胞を区別することに成功している。一般に K 個のクラスターは K-1 次元の空間に埋め込むことが可能なので、３種類の培養細胞を空間的に分離した形で配置できれば、それがどのような高次元空間であっても、K 個のクラスターを判別できる K-1 次元の空間が存在するはずである。その意味では、DIABLO もテンソル分解を用いた教師なし学習による変数選択法と同様に、高次元空間内に３種類の培養細胞を分離した形で配置することに成功したと思われる。

それではどれくらいよく分離できたのであろうか？このパフォーマンスであるエラー率を見てみると５％程度と表 1 の判別性能とほぼ同じである。つまり、DIABLO とテンソル分解を用いた教師なし学習による変数選択法の判別能力はほぼ同等であったと結論付けられる。

次にオミックスの選択性能を見てみる。図 4 は DIABLO

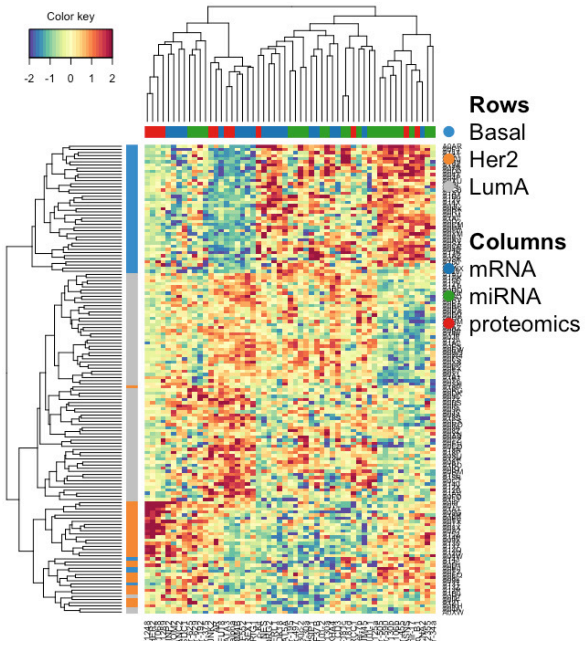

図 4　DIABLO で選択されたオミックスのヒートマップ。
Fig. 4　Heatmap of omics data selected by DIABLO.

で mRNA,microRNA, プロテオームを１０個ずつ選択した場合のヒートマップである。図 2 と、選択されたオミックスこそ異なるものの、非常によく似たヒートマップ（培養細胞（行）がよく別れており、また、異なったオミックス（列）から同じような培養細胞依存性があるオミックスを選べている、など）になっていることが解るだろう。

これらの結果からテンソル分解を用いた教師なし学習による変数選択法と DIABLO の性能はほぼ同等であると結論付けられる。

4. 議論

前節で DIABLO とテンソル分解を用いた教師なし学習による変数選択法はほぼ同等の性能があることが示された。それではどちらがより方法として優れているだろうか？

まず、第一に計算時間の面でテンソル分解を用いた教師なし学習による変数選択法が圧倒的に有利である。DIABLO は教師あり学習であるために最適のパフォーマンスを上げるために学習のための繰り返し計算を行わなくてはならない。これに対し、テンソル分解を用いた教師なし学習による変数選択法はたった一回のテンソル分解を行うだけである。ざっくり行ってしまえば、DIABLO で学習のために繰り返し計算でなされている一回分の計算ですんでしまう。実際、DIABLO が数十分を要するところ、テンソル分解を用いた教師なし学習による変数選択法は数秒で済んでしまい、まったく勝負にならない。

また、テンソル分解を用いた教師なし学習による変数選択法はテンソル分解を行っているだけであり、HOSVD は初期値を必要としないアルゴリズムなので、オミックスの選択に乱数依存性がない。これに対して、DIABLO は初期値依存性のある収束計算を行う必要があり、乱数によって全く異なったオミックスが選択されてしまう。生物学的には「どのオミックスが重要か」という問に答えることも大切なので、選ばれるオミックスが乱数依存性を持っているのは望ましいこととはいえない。この点でもテンソル分解を用いた教師なし学習による変数選択法の方が DIABLO より優れていると思われる。

テンソル分解を用いた教師なし学習による変数選択法はまた、モデルビルディングが不要だという点も有利である。掛け算を行っているので、個々のオミックスの間の比重は考えなくて良い。テンソルの個々の成分がすべて 3 つのオミックスの掛け算なので、最初から比率を考える余地がない。複数のオミックスをどう統合解析するかを考える上で重みの値は重要だからこれがないテンソル分解を用いた教師なし学習による変数選択法は非常に有利である。DIABLO は残念ながらこの部分がパラメーターフリーではなく、人間の関与が必要なのである。

一方、テンソル分解を用いた教師なし学習による変数選択法の欠点は必要とするメモリーが膨大になってしまうことである。ここでは DIABLO 用に用意されたテストデータを用いたため、個々のオミックスの次元数は数百だが、実際には遺伝子は数万個ある。テンソル分解を用いた教師なし学習による変数選択法は掛け算をしてテンソルを作っているので、個々のオミックスの次元数の累積になってしまい、膨大なメモリーが必要になってしまう。DNA のメチル化など、次元数がゲノムの塩基長と同程度のデータなどは扱うことが出来ず、部分和をとるなどして次元数を下げないといけないため、現実的ではない。

しかし、この問題は最近、テンソル分解を用いた教師なし学習による変数選択法にカーネルトリックを導入することで解決された [6]。カーネルトリックでは問題を双対空間で解くために、オミックスデータの次元数は計算量と無関係になり、サンプル数だけが問題になる。N 個のサンプルに K 種類のオミックスデータを計測した場合、N^{K+1} のオーダーのメモリーしか必要ではなくなった。このため、カーネルテンソル分解を用いた教師なし学習による変数選択法はより広範なデータに対して適用可能であることが期待される。

5. おわりに

マルチオミックスデータの統合解析の重要性は論を待たない。今後計測可能なオミックスデータの種類数は増えることがあっても減ることはない。テンソル分解を用いた教師なし学習による変数選択法がこの問題に対するデファクトスタンダードになってくれることを願って止まない。

謝辞　本研究は科研費番号 20K12067、20H04848、

19H05270 の科研費の支援の元に行われた。

参考文献

[1] Taguchi, Y.-H.: Multiomics Data Analysis Using Tensor Decomposition Based Unsupervised Feature Extraction, *Intelligent Computing Theories and Application*, Springer International Publishing, pp. 565–574 (online), DOI: 10.1007/978-3-030-26763-6_54 (2019).

[2] Taguchi, Y.-h.: Multiomics data analysis using tensor decomposition based unsupervised feature extraction – Comparison with DIABLO–, *bioRxiv*, (online), DOI: 10.1101/591867 (2019).

[3] Taguchi, Y.-h.: *Unsupervised Feature Extraction Applied to Bioinformatics, A PCA Based and TD Based Approach*, Springer International (2020).

[4] Singh, A., Shannon, C. P., Gautier, B., Rohart, F., Vacher, M., Tebbutt, S. J. and L Cao, K.-A.: DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays, *Bioinformatics*, Vol. 35, No. 17, pp. 3055–3062 (online), DOI: 10.1093/bioinformatics/bty1054 (2019).

[5] R Core Team: *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria (2019).

[6] Taguchi, Y.-h. and Turki, T.: Mathematical formulation and application of kernel tensor decomposition based unsupervised feature extraction, *bioRxiv*, (online), DOI: 10.1101/2020.10.09.333195 (2020).