高次元遺伝子データ解析理論 (Theory2) の完成報告 1

01202720　成蹊大学名誉教授　新村　秀一　　SHINMURA Shuichi

1. はじめに

1995 年に Microarray で動物の遺伝子が出す蛋白量（**発現量**）が計測できるようになった。人の遺伝子は 3 万個以上と言われ、生命現象を制御している。山中教授は Rat の ES 細胞で活性化する 24 の遺伝子を遺伝子 DB で特定し、高次元の特徴抽出法（**FS**）の Np-hard (**困難 2**)を簡単に解決した。さらに医学常識に反する「多変量的な実験」で山中 4 因子を見つけた。

多くの医学研究が研究に用いた高価なデータを無償公開した。そして統計、機械学習（ML）、AI と言った理工学研究が「高次元データ解析（**Theory2**）[4]」の研究をしたが、全て間違っていて役に立たない。その理由は、例えば 100 例で変数が 1 万個程度の 169 種の 2 群の Array が、全て線形分離可能（**LSD**）という事実を指摘した研究がないという分かり易い事実である。全遺伝子が n 個以下の LSD の Small Matryoshka (**SM**)と最小次元の LSD の Basic Gene Set (**BGS**)に分割できる（**LSD の 4 つのデータ構造、Fact3**）。

筆者は IP で最小誤分類数(minimum NM, **MNM**)を見つける最適線形判別関数(Optimal LDF)の **RIP** を開発した。RIP で判別するだけで、p から n 個以下の判別係数が非零で残り(p-n)個が 0 になり、iPS 研究と同じく **FS の困難 2** を解決した。これは筆者が見つけた判別理論の **Fact1**（LDF の係数と NM の関係）と **Fact2** ($MNM_k <= MNM_{(k+1)}$)で説明できる。さらに、全ての LSD は、その中に SM から BGS 迄の Nest 構造（**データ構造 1**）をもち、**LINGO**[1][2]**の Program3** で Array を判別し最初の SM1 を見つけた後、これを省いて再度判別すれば 2 番目の SM2 が求まり、簡単に排他的な SM に分割できる(**データ構造 2**)。また逐次変数減少法の手順で山中 4 因子と同じ BGS に分割できる(**データ構造 3**)。ロジスティック回帰（**LR**）で自由度 DF がほぼ n の DF 遺伝子集合に分割できる(**データ構造 4**)。これが LSD の汎用的な 4 つのデータ構造（**Fact3**）であり、誰も発見していない。

医学研究は、生物学的知識と顕微鏡で 400 個以上の**遺産的な発癌遺伝子**を見つけた。これらは 1 変数の発癌遺伝子であるが、癌研究の大きな成果である。一方 iPS 研究の 4 因子や SM や BGS は発現データで見つけた**多変量の発癌遺伝子の候補**である。

そこで 10 重 CV の 10 組の検証標本の平均誤分類確率 ER（**M2**）を 0 とし、5 個以下の遺伝子をもつ「**重要な BGS**」を見つけることにした。また MP を知らない研究者が 2 時間程度の LINGO の Program の説明を聞けば直ぐに癌診断に役立つ「2Step スクリーニング法」を開発した。さらに重要な BGS を検証する 5 種の Validation を開発した。**Validation3** は、重要な BGS に遺産遺伝子が含まれれば、医学研究者はそれで多変量の癌遺伝子の特徴が分かるので特に重要だ。これで 2022 年に完全な **Theory**2 が完成した。

2. 多くの研究分野の全ての研究が間違った理由

以下の多くの研究領域の研究者が間違った研究をした理由は、専門家による検証の必要がある。

2.1　2006 年迄の第 1 世代の医学研究

第 1 世代の古い Array を用いた医学研究で、ハーバード大学の Golub (1999)、Shipp と Singh (2002)と Chiaretti (2004)はこの研究の中核である。特に 1970 年から研究している Golub 博士は、誰もが認める先達である。彼らは遺産遺伝子に注目し、各遺伝子がその ε 近傍に含まれる個数をある基準で選び、約 50 個の遺伝子一組を見つけ、残りを雑音として無視した。そして SOM（非線形のクラスター分析）で、2 群が視覚的によく分かれていると主張し、最後に行うべき生存時間分析を行った(**Validation5**)。LDF で NM や ER を検討していないのは、判別関数の ER が役に立たないこと(判別分析の**問題 1**)を知っていたためであろう。古来「癌は遺伝子の病気であり、不均質な病気」ともいわれている。Fact3 はこの常識に反しているが、僅か 50 個の一組の遺伝子で癌の診断ができると考えるのは大きな問題である。しかも遺産遺伝子を用いた「加重投票法」という FS は、ほぼ全てが遺産遺伝子の寄せ集めと考えられる。米国の医学研究はデータ解析の経験がない統計研究者が分析したと考える。このため NIH が Golub らの研究費を打ち切り、癌研究者に Golub ショックを起こした。これが原因か分からないが 2007 年以降の第 2 世代の医学研究では、1）FS で正しい発癌遺伝子の組を選び、2）ER が最小な判別手法を Array 毎で決める、というデータ解析の研究が姿を消した。

医学研究で選んだ**遺産遺伝子の組み合わせで診断しても悪くない結果が得られる**事に安住した状態である。また医学誌最高峰の NEJM 誌に掲載された Tien

(2003)は、ｔ検定で 56 個の遺伝子を一組求め、LR で全ての係数が 0.01%棄却されたので選んだ遺伝子が妥当と主張した。LSD では回帰係数が不安定になるという「多変量の関係」を知らない問題がある。

2.2 統計研究者の問題

統計研究者は、判別分析の NM が信用できない (**問題 1**)、LSD の研究がない (**問題 2**)、大学入試センター試験の数学 IIb を Fisher の LDF で判別すると ER が 30%を超えるという分散共分散行列の**問題 3**、判別分析は推測統計学でない**問題** 4 を知らない。MNM はデータに一意に決まり 100 重 CV（**Method1**）で解決し、新しい判別理論（**Theory1**）を 2015 年に完成した[2][3]。この他、SVD による高次元 LDF と LASSO という役に立たない研究を行ってきたが成果がない。

2.3 機械学習（ML）、AI、バイオ工学等の問題

Theory2 の「3 つの困難」という言い訳は、数学に基礎をおく研究者の基本を知らないことが原因だ。**困難 1** の高次元データの分析手法がないは、「連立方程式の解の条件」に気づいていない。また高次元という認識は間違いで普通のデータ(n>=p)と横長データ(n<p)の違いを理解していない。癌と正常が各 1 例の 10 個の計測値があれば、p から 2 変数を選ぶことで連立方程式やロジスティク回帰の DF が 2 で残り 8 個は無視する。即ち ${}_{10}C_2$ の組合わせ問題になる。**困難 2** は FS が Np-hard で、**困難 3** は高次元の雑音から癌遺伝子という信号を見つける事は難しいで、SM と BGS 分割という FS で簡単に両方を解決した。

ML 研究は、判別手法を含む分類手法をまとめて Classifier を体系化したが、Array 毎に分析して最小の ER をもつ手法とデータの組合わせを議論している。これは「癌は不均質な病気」の概念に反しない。しかし、RIP と Hard margin 最大化 SVM(**H-SVM**)と LR の３種の LDF だけが LSD を ER＝0 で判別でき、他の Classifier は必要としない。そして何故か多くの研究者は Kernel SVM を用いているが、間違いである。

開原東大名誉教授が厚生省で企画した介護保険システムを高校の１年先輩の都立病院の副院長が分析していた。彼の求めで分類木を使うようにアドバイスしたら、３ヶ月程で Chaid で分析した結果をｃでシステム化した。しかし国会で決めた日程の制約で事前のシステム検証を行わないで、分析に用いていない在宅の老人まで適用して当初問題が起きた。現在最大規模の統計システムが運用されている。即ち 2 群判別に 3 つの LDF が、多群の階層のある診断には決定木や DT が適していて使い分けが必要になる。他の Classifier の ER＝0 は、決して LSD であることを示していない「ER の信頼性」の理解不足が問題だ。

2.4 MP による判別理論の終焉 [5]

RIP による判別研究を始めた年に米国の OR 誌に Stam [5]が MP による 300 以上の判別理論の総括を行い OR で終焉した。これらの研究が利用されないのは、統計的な判別モデルを MP で改善しているだけで、根本的な NM の欠点を解決していない。現在 IP でしか NM を最小化できない事が重要だ。

3. おわりに

2015 年に、RIP と**事実 1** と**事実 2** で判別の 4 つの問題を解決し Theory1 を完成した。特にスイス紙幣の真札と偽札、試験の合否判定、小型車と普通車の 3 種の LSD の研究が重要だ。2015 年 10 月 26 日に科研費の高次元シンポで発表し、定年前に研究を纏めたと考えた。しかし筑波大学の院生が、Array を PCA で分析すると第 1 固有値がスパイク上に大きいという報告で、6 種の古い Array が無償で利用できることが分かった。28 日に Shipp を判別すると 1 秒で LSD であり、32 個の係数が非零で残りの 7087 個が 0 になり自然に 32 個の SM1 が見つかった。そこで LINGO の Program3 で、排他的な LSD である SM 分割を行った。後日 Program4 で、排他的な LSD である BGS 分割を行い、Theory2 の基本を完成した[4]。

Theory2 で重要なのは**事実 3** で「動物の発現量で簡単に癌の遺伝子診断と同じ発見ができる」。さらに、MNM はデータに一意に決まる。そこで RIP で例えば CPD (児頭骨盤不均衡)データで自然分娩か帝王切開かの分娩法を決める。240 例で 2 症例が誤判別するので Test 標本として省き、残りの 238 例の LSD を判別し、19 変数の SM1 と 14 変数の BGS1 が求まる。この分析結果は、日本医科大学の産婦人科に提出した結果に比べて遙かに質の高い結果が 2 日で得られた。これは、今後判別分析が多くのデータに適用できる福音になる。身近に適切な無償の Array があれば、誰でも簡単に癌の遺伝子診断と同じ発見ができる。

参考文献

[1] 新村秀一(2011)数理計画法による問題解決法。日科技連。

[2] 新村秀一(2010)最適線形判別関数。日科技連。

[3] Shinmura S (2016) New Theory of Discriminant Analysis After R. Fisher. Springer.

[4] Shinmura S (2019a) High-dimensional Microarray Data Analysis. Springer.

[5] Stam A(1997) Non-traditional approaches to statistical classification: Some perspectives on Lp-norm methods. Ann Oper Res 74:1-3