高次元遺伝子データ解析理論 (Theory2) の完成報告 2

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

1. はじめに

1971 年から 4 年間大阪府立成人病センターで、30 群以上の心電図異常所見群の診断論理を、正常と各異常所見の 2 群判別として LDF で開発した。しかし医師の開発した多群の階層化構造をもつ ECG データの分類に対応できる枝分れ論理（DT）に敵わなかった。これが医学診断に役立つ判別理論（**Theory1**）研究の開発動機である。また介護保険システムの分類木による開発の成功にも多少貢献した。これら以外の分類手法は研究対象であっても実用システムの開発に用いるのは注意が必要と考える。

本発表では LSD を中心とした新しい成果を示す。

2. 判別理論は Fisher の仮説でなく組合わせ論

Gauss は 2 地点間の繰り返し測定値が Gauss (正規) 分布することを見つけた。これを聡明な Fisher が注目し、2 群が平均だけ異なる同じ 2 つの正規分布と仮定し、 p 次元の LDF が $f(\mathbf{x})=y_i*({}^t\mathbf{b}*\mathbf{x}_i+1)$になることを見つけ判別理論を提唱した。今正常 1 例 ($y_1=1$)と 2 例の癌 ($y_2=y_3=-1$)の 2 変数の計測値を考える。統計ソフトに (3＊2) のデータを入力し判別係数 **b** が求まる。MP で **b** を求めるために 2 次元の係数空間で次の 3 個の $\mathbf{x}_i$ に対応した判別超平面 H_i を考える。

H_i: $f(\mathbf{x}_i)=y_i*({}^t\mathbf{x}_i*\mathbf{b}+1)$　　for i =1,2,3.　　(1)

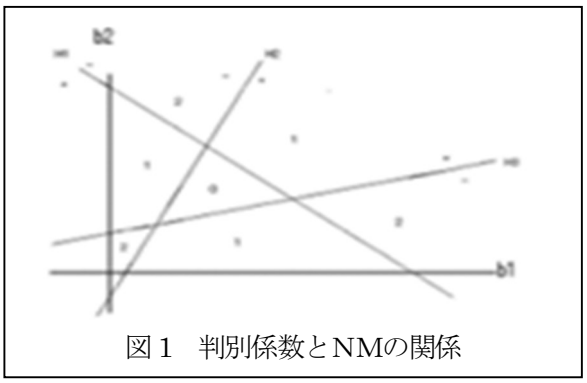

図 1　判別係数とNMの関係

これを $f(\mathbf{x}_i)=0$ で図示すると図 1 になる。$f(\mathbf{x}_1)>0$ は H_i のスコアが正の半平面で、$f(\mathbf{x}_1)<0$ は負の半平面で、この 1 点を選ぶと $\mathbf{x}_1$ を誤判別する LDF になる。3 つの H_i で 7 つの凸体ができる。例えば内点が 1 個の負と 2 個の正の半平面にあれば NM=1 で同じ $\mathbf{x}_i$ を誤判別する。全ての三角形の内点が、＋半平面で MNM=0 の LDF になる。N=3 であり NM は 0, 1, 2 しかないので、Gauss 分布でなく「組合わせ論」で考えて誤分類例の研究を行うべきだ。しかし ${}_pC_n$ 個以上の組み合わせを考える必要があり、真の信号（癌遺伝子）を選択するために 5 段階の検証法を考えた。

3. 3 つの通常の LSD と 169 の Array で Fact3 を確認

RIP と **Fact1** と **Fact2** で 4 つの**問題**を解決し 2015 年に **Theory1** を完成した。そして 169 種の Array が Fact3 という判別分析に重要なデータ構造を見つけた。

スイス銀行紙幣データが 6 変数で LSD で、(X4, X6) が最小次元の **BGS** である。**Fact2** の MNM の単調減少性から 6 変数から 2 変数の BGS まで 24 組のモデルが LSD で **Nest 構造 1** を持ち、残りの 39 組が LSD でない。ただしこの成果は、他の偽札紙幣に適用できない。この 24 組は、**Theory2** では癌と正常を分ける「**多変量の癌遺伝子の候補**」と呼ぶ。24 遺伝子から山中 4 因子までの万能細胞ができる遺伝子の組も多変量の関係を示す。

一方合否判定は、2 個の得点合計が 50 点を合格とすれば f=T1+T2-49.5 という LDF になる。合格と不合格は SV1≧50 と SV2≦49 の 1 点差で得点が[0, 100]であれば、RatioSV=1＊100/100=1%(**検証 2**）が線形分離度を表す。また日本車の普通車と小型車は、排気量と座席数で規格が決まり 2 個の 1 変数の BGS である。この 2 つの LSD は、いずれも人間が判別規則を変数の一次式で定義したので、常に LSD になる。

一般に生物に関する計測値は、より曖昧な精度である。しかし発現量にかぎって、筆者が初めて Array と SM と BGS と DF 遺伝子集合が LSD で、第 2 世代の Array が 10 重 CV で 10 組の検証標本の平均 ER の **M2 が 0**（**検証 1**）になる多くの例を見つけた。これで膨大な「多変量の癌遺伝子の候補」から真の信号である発癌遺伝子を選択できる。しかし「癌は不均質な病気である」という事実に反する。そこで正常細胞が癌細胞になると「**エピジェネチックな変化**」が遺伝子に起きて発現量が正確に癌と正常が分かれることを **Fact3** が示していると考えた。即ち「発現量は人類が出会った最高品質の計測値」である。

4. 癌の症例設計の 3 原則

169 の Array の分析を通して、医師が非線形のクラスター分析で視覚的によく分かれていると主張している。しかし階層型のクラスター分析であっても、誤

分類症例や癌の亜種の位置関係は分からない。それを PCA に取り込んで散布図(**検証 3**）を見ればある程度の位置関係が分かる。

NIH の Liver3 は、181 例の肝癌と 176 例の正常とほぼ同数で、正常例を含み、統計的には 357 例と一番多いので、鈴木医師と癌の疫学研究で議論した従来の「**症例設計の３原則**」を満たす一番良いデータである。しかし**図 2** に示すように散布図上で多くの誤分類例があり、M2 の結果は悪かった。しかし 63 例の Colorectal6（31 例の肝癌と 32 例の正常）は中規模だが、M2=0 の SM や BGS が多く見つかり、散布図でも 2 群が **Prin1** の正と負で綺麗に分かれた。ほぼ同じ条件の Colorectal5（26 肝癌例と 26 正常例）は、**図 2** の下に示すように 4 例の誤分類症例があり、**M2＝0** になる SM や BGS は無かった。そこで散布図上の 4 例の誤分類症例は、高次元で考えると LSD であるが、辛うじて 3 次元以上の小さなバラツキで LSD になり真の信号でない可能性が高いと考えた。このような誤判別例を Test 標本として省けば、省いた Array は変動の大きな散布図上で分かれ、**M2=0** になる BGS も多い良好な結果になった。「発現量は高感度な測定値で僅か 4 例を省いただけで改善」されるので、誤分類例が「**癌の不均一性**」を表すと考えた。これは医学研究で検証してほしい。

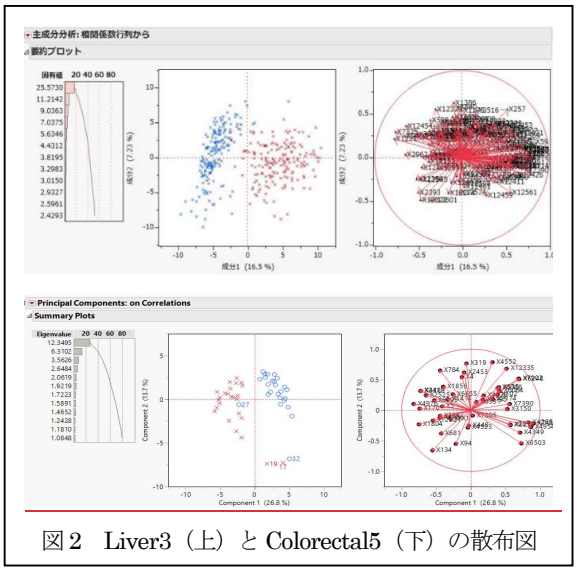

図２　Liver3（上）と Colorectal5（下）の散布図

そこで Liver3 を Prin1 のスコアで 2 群の距離が離れた 300 例、200 例、100 例、60 例と 40 例を選び、5 種類の Array を作った。**M2＝0** になる BGS があるかを調べたところ、60 例と 40 例だけで多く見つかった。即ち、「**癌症例に限って**癌と正常がほぼ同数で 100 例以下の中規模な Array が癌診断に有効な結果をもたらす傾向がある事」が分かった。

そこで研究者が LINGO で開発した Program の利用法さえ分かれば、MP の理論を知らなくても利用できる「スクリーニング法」を開発した。Step1 は Array の全遺伝子を Program2 で SM に分割する。操作ミスを少なくするために 10 組の SM に含まれる遺伝子を、Program3 で BGS 分割する。しかしデータ解析で見つけた BGS が真の多変量の癌遺伝子か否かを調べる必要があるので Step2 で、癌診断に有効な BGS を選ぶ。それには、医学研究で見つけた 1 変量の遺産遺伝子が、BGS に含まれる調査を行い(**検証 4**)、含まれた遺産遺伝子の特徴で BGS の役割が分かる。そのため、症例数が少ないほど BGS に含まれる遺伝子も少なくなり関連性が分かりやすい。そこで **M2=0 で遺伝子数が 5 個以下**の「**重要な BGS**」を見つけることにした。医師がこの関係を調べる**検証法 4** と本庶先生が普及に尽力された Genome Cohort と生存時間分析の**検証法 5** を行えば良いと体系づけた。現在の医学研究では、医学的に選んだ遺産遺伝子の何組かの集合を、クラスター分析し直ぐに**検証法 5** を行っている。遺産遺伝子の組み合わせを用いているので、そこそこ良い結果になり満足しているようだ。しかし LR で判別すれば恐らく LSD でないことが分かるが、誰も行っていない。最初に正しい多変量の癌遺伝子の候補を選び、5 段階順に検証し、真の多変量の癌遺伝子で研究すべきことを怠っているようだ。

5. おわりに

まさか退職後に日本で無視された RIP で、Theory2 の大輪の花が咲くとは思わなかった。多くの研究者はしつこく目標を定めて退職後も研究すべきだ。癌研究者には、「医学の専門知識がなくてもデータ解析は合否判定と同じく発現データから正しい結果を導き出せる学際的な学問である」と知ってもらう必要がある。しかし掲載料が 30 万以上の医学誌が多く研究費の捻出の問題がある。現在年 10 万円程度で米国の CSCE という夏と冬の学会で年２本以上の 7 頁以上の Proceedings で発表し、その後に Springer の研究叢書や IEEE で出版された論文が 12 篇以上ある。契約上問題のある物を除いて Research Gate からダウンロードできる。頭は冴えきっているが、目や腱鞘炎に悩まされて研究をいつまでできるか不安である。それ以上に「和田先生の食事療法は守っている」が癌等で死ぬ可能性もあり、その前に普及させたい。企業や大学で遺伝子解析を行いたい場合、2 時間程度の ZOOM の講習で直ぐに使える。研究より普及が重要で、私以外に同じ分析方法は再現できないので、誰かに継承してほしい。