サブグループ同定について

水田　正弘　　大学統計教員育成センター　特任教授

1. サブグループ解析とサブグループ同定
   最近、「本医薬品は、50歳以上で過度な飲酒をする人に対する効果が低い」など、「特定の属性」を有する人の薬効などに関する報告を耳にする。このような主張を統計的に証明する問題は比較的明確に設定できる。これは、サブグループ解析として、個別化医療の研究の一環で注目されている。しかし、「特定の属性」を見つけ出すのは、別の課題であり、サブグループ同定法(Subgroup Identification)と呼ばれている。サブグループ同定法として、複数の手法が提案されているが、想定するサブグループの族の中で、効果(Treatment effect)が最大となるサブグループが導出される保証は全くない。そこで、想定する全てのサブグループを列挙し、それぞれの効果を計算することで、「厳密な意味で最適なサブグループ」を決定する方法を研究している。

2. 「良い」サブグループとは？
   ある属性を有するサブグループにおいて、治療を受けた人の、受けない人に対する効果(Treatment effect)の差が、その属性を有しないサブグループにおける効果の差と比べて高い場合、「良い」サブグループと言える。このこと自体は、差の検定やGLM、LASSOなどで調整した効果の差として、統計的な検定やベイズ的アプローチにより評価できる。また、Pmtree (model4you.Compute model-based tree from model)、glmtree (partykit. Logistic Regression Trees)、QUINT (Qualitative INteraction Trees)などが「良い」サブグループを同定する手法（ソフトウェア）として発表されている。

3. 最適なサブグループの同定法
   サブグループ同定の問題を、想定したサブグループの属の中から、治療等の効果が最も高いものを導き出すと設定することができる。先に紹介した既存の手法は、残念ながら近似最適解の検出に過ぎない。そこで、想定する全てのサブグループについて、効果を計算することで、（近似ではない）最適な解を同定することを試みた。2020年の統計関連学会連合大会では、p=5,n=100の人工データにおいて、43,199,128,758個(432億個)のサブグループが列挙されたことを報告した。しかし、実際のデータ解析を想定した場合、サンプルサイズn、変量数pともに少なすぎる。そこで、データ構造の特徴抽出を実施するなどにより、適用可能な状況の拡大および計算時間の短縮方法を検討している。

4. まとめ
   病気になったとき、その治療法の選択は人生において大きな岐路となることが多い。すなわち、1つの疾病に対して、複数の治療法があるとき、どれを選ぶかは重大な問題である。特定の治療法が全ての人に有効であるとは限らないので、個別化医療が注目されている。これの基本として、特定の治療方法の有効性を集団として検討するサブグループ解析がある。ある治療方法が有効であるサブグループと、そうではないサブグループが存在することは自然である。適切な「特定の属性」を設定し、治療の対象者を限定すれば、良好な治療成績が期待できる。例えば、COVID-19関連で、一般化した表現に、「特定の年齢の人に対するワクチン接種の推奨」がある。すなわち、あるグループに対してワクチンは損失よりも利得が大きいという意味である。同様に、ある療法が有効なグループの特定、ある政策に対する賛成が多いグループの特定などは重要な事項であり、データサイエンスにおいて、普遍的な課題である。サブグループを与えたとき、療法が有効や賛成が多いなどの数値的な評価は、統計的または社会学的に定義できることが多い。そこで、最適なグループの同定法、すなわちサブグループ同定法が研究されてきた。これにより、近似的に最適なグループを見つける方法は、いくつか開発されているが、最適解が得られる保証はない。厳密な意味での最適解の導出法があれば、それ自体、有効な手法であるとともに、これまでに提案されてきたサブグループ同定法の手法を評価することができる。

参考文献

水田正弘：Subgroup identificationのための全列挙について, 2020年度統計関連学会連合大会講演報告集

水田正弘：Subgroup Identificationにおける条件設定と最適解について, 2021年度統計関連学会連合大会講演報告集

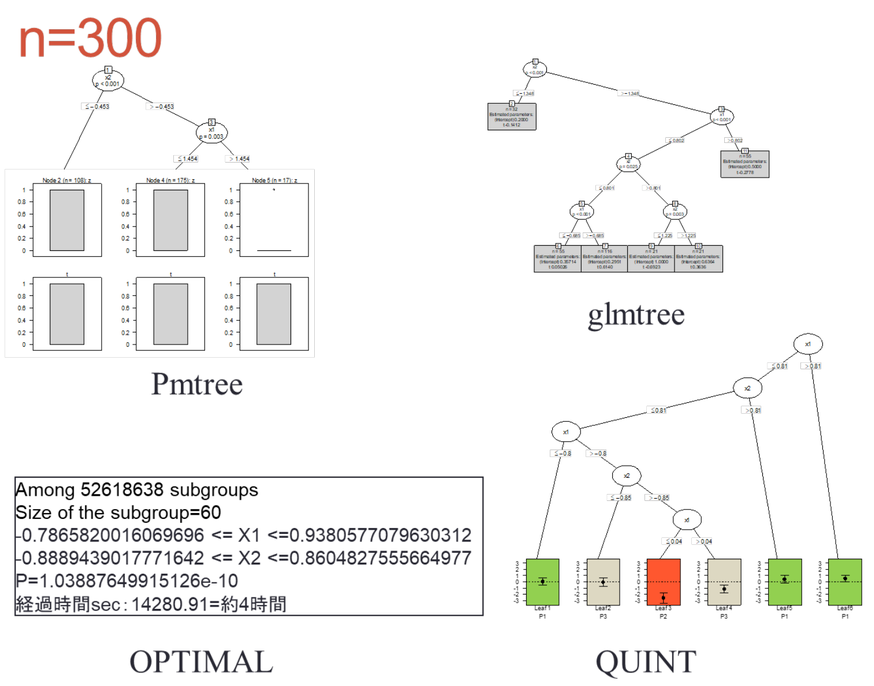

図1．人工データ（P=2, n=300）に対するPmtree, glmtree, QUINT, OPTIMALの結果

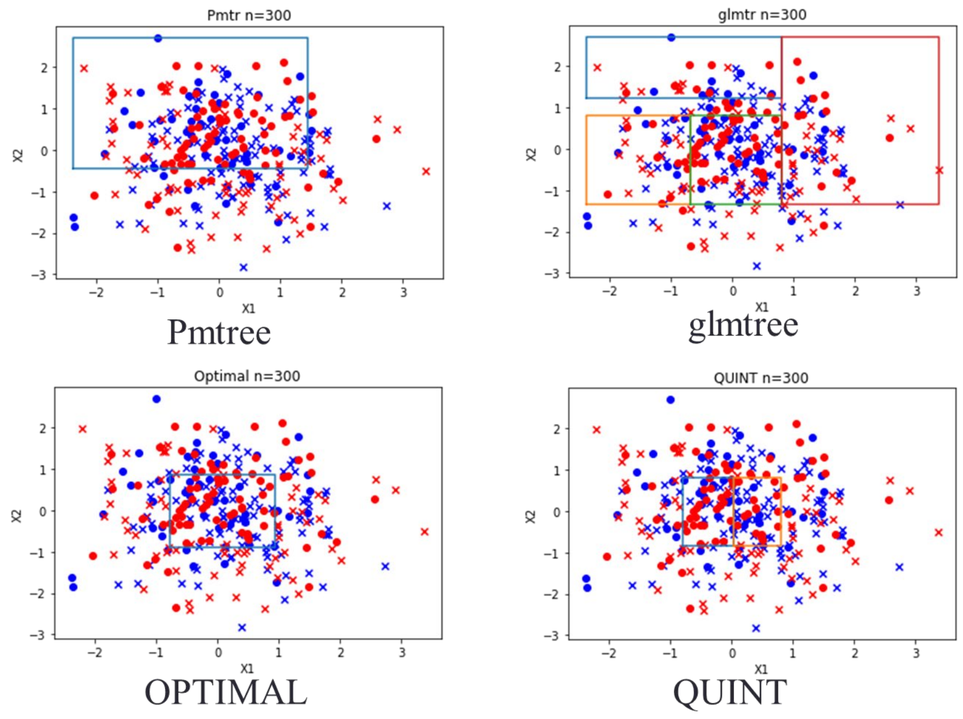

図2．人工データ（P=2, n=300）に対するPmtree, glmtree, QUINT, OPTIMALの結果の図示

Intel i7-6700 @3.40GHz 8GB Windows 10 Pro

Subgroupsの個数

n	p=2	p=3	p=4	p=5	p=6	p=7	p=8
30	7367	42359	323141	1095138	2095445	5502020	10257059
40	23730	315814	1642258	8541526	45209914	267994983	465289475
50	55383	763918	5855067	40492538	252925892		
60	109477	2564629	41869673	421171790	2205633436		
70	212738	7325518	124508171	1205050284			
80	274545	13980721	422780297	1018942170			
90	526690	31487613	924279288	17294722054			
100	840135	57023452	1884625234	43199128758			
200	11228118	2603313451					
300	53993551						

43GB個

計算時間(秒)

n	p=2	p=3	p=4	p=5	p=6	p=7	p=8
30	0.012	0.092	0.771	3.270	10.749	32.221	65.828
40	0.039	0.544	3.633	21.455	128.414		
50	0.101	1.428	14.297	113.929	841.001		
60	0.172	4.824	86.062	956.226	5217.815		
70	0.352	13.942	274.310	2699.425			
80	0.510	27.940	947.914	20971.274			
90	1.017	64.442	2106.096	40507.423			
100	1.799	123.689	4133.203	96078.766			
200	41.477	16424.585					
300	279.094						

ほぼ1日

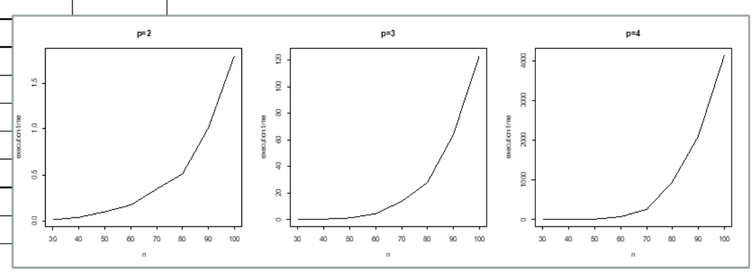

図3．列挙の個数と計算時間