論文の内容の要旨

論文題目　ヒト腸内細菌組成データを用いた関連解析手法に関する研究

奥井 佑

・序文、目的

ヒト腸内細菌のゲノム情報を網羅的にシークエンスするメタゲノム解析のうち、全ゲノム領域の中で各菌種の識別が可能である 16S rRNA 領域のみをシークエンスする 16S rRNA 解析により各菌種の組成に関するデータ(以降、16S rRNA データ)が得られるようになった。16S rRNA データは統計学的には各個体において各菌種がいくつ検出されたかを示す多変量のカウントデータであり、個体ごとに総カウント数は大きな違いがあるため、解析を行う際には一般的に総カウント数の違いを補正する正規化を施す。近年、形質情報と関連する菌種を特定する関連解析手法が多く提案されてきている一方で、複数の菌種変数を同時に扱う多変量データ解析手法を用いる場合の適切な正規化の方法は検討されていない。正規化の方法として、各菌種変数値を各個体の総カウント数で割った組成割合のデータを用いた場合には菌種変数値の和が 1 となり、変数間に擬似的な相関関係が生じるという compositional data の問題が生じる。他の正規化法として、各個体の総カウント数から一定数のカウント量をサブサンプリングし、各個体の総カウント数を等しくする rarefying 法がある。rarefying 法を用いることで変数間に擬似的相関を生じさせることなく各個体の総カウント数を等しくすることが可能であるが、使用可能なデータの一部を用いないという問題がある。

多変量データ解析手法を用いる際には、一つのデータに対して複数のモデルを適用しそれらの結果を統合して最終的なモデルとする集団学習を行うことでモデルの予測能を高めることが可能である。集団学習を応用して同一データに対して rarefying 法を適用したモデルを複数生成し、それらの結果を統合することでデータの一部のみを用いる rarefying 法の問題は解消することができる可能性がある。そこで本研究では、関連解析において多変量データ解析手法を用いる際の正規化法に着目し、rarefying 法に集団学習法を併用した手法を提案して他の正規化方法とともにシミュレーション実験により性能を評価する。

・方法

1. 提案法

本研究では多変量解析データ解析手法として実際に 16S rRNA データ解析に用いられている Random forest(RF), Elastic Net(EN), Sparse partial least squares discriminant analysis(SPLSDA)の 3 手法について検討した。提案法である rarefying 法と集団学習法を併用した方法については 2 通りの方法を検討した。1 つ目の方法では、rarefying 法によりデータを正規化して多変量データ解析

手法を適用したモデルを同一データに対して複数回作成し、それらの結果を統合したモデルを最終的なモデルとする方法を検討した(以降、rarefying +サブサンプリング法と表記)。2 つ目の方法では、データに対してブートストラップ法を適用して複数のブートストラップ標本を生成し、各ブートストラップ標本内で rarefying 法による正規化を行って多変量データ解析手法を適用しモデルを作成した。それらモデルの結果を統合して最終的なモデルとした (以降、rarefying +バギング法と表記)。これら 2 つの方法をまとめて rarefying 法に集団学習法を併用した方法とする。

2. 本研究で検討したモデルとその評価法

rarefying 法に集団学習法を併用した方法の性能を評価する方法として、各多変量データ解析において、正規化法として組成割合を用いた場合、rarefying を用いた場合、rarefying 法に集団学習法を併用した方法を適用した場合、組成割合とバギングを併用した場合の 4 通りの解析を行い、各モデルの性能を評価した。比較の公平性を保つため、各ブートストラップ標本について組成割合による正規化を施したモデルを作成し、それらのモデルを統合する組成割合とバギングを併用した方法も比較に含めた。これらより、本研究では正規化法、多変量データ解析手法、集団学習法の組み合わせにより計 14 種類の方法について検討した。表 1 に各モデルの詳細とそれぞれの手法の変数重要度の計算方法、判別力の評価方法を示した。組成割合データを用いる場合について RF では計算過程でバギングを利用しているため、バギングとの再度の併用は行わなかった。モデル 1~4、モデル 5~9、モデル 10~14 と多変量データ解析手法内において正規化法と集団学習有無を変化させた際の性能に基本的には着目する。各説明変数の応答変数との関連度を RF にならい本研究では変数重要度とよび以降の検討を行う。

2. シミュレーション実験

実際の 16S rRNA データを模したデータを発生させ、各モデルの形質関連菌種の同定力、形質の判別力を評価した。シミュレーションデータは、総カウント数のばらつき、応答変数に対する関連変数以外の要因の効果の有無、説明変数の数をそれぞれ 2 通り設定し、表 2 のように計 8 パターンのシミュレーション実験を行った。データはモデルの訓練用と検証用で 2 通り生成した。各モデルの評価方法として、各モデルの変数重要度の順位と真の関連変数との一致度に関する AUC 値を算出した。また、各モデルによる応答変数の判別力についても、検証データにおいて訓練データで作成したモデルの予測値と実際の応答変数との一致度に関する AUC 値を算出し評価した。それに加え、各モデルについて、関連変数の真の係数値と推定された変数重要度のスピアマン順位相関係数の値を算出した。シミュレーション回数は 200 回とした。

・結果

シミュレーション実験の結果、表 3 のように SPLSDA,EN については集団学習法と rarefying 法を併用した方法(モデル 8,9,13,14)で変数重要度の AUC 値が他の正規化法より大きな値を示した。一方、RF(モデル 3,4)に関しては同様の傾向は認められず、RF を用いたモデル(1,2,3,4)はモデル

8,9,13,14 よりも常に値が低くなった。判別力については、各モデルで判別力に大きな差が生じることはなかったが、集団学習法と rarefying 法を併用することで他の正規化法よりも AUC 値が若干大きくなった。相関係数については、SPLSDA,EN については集団学習法を用いたモデルで用いない場合よりも値が大きくなったが、RF に関しては同様の傾向は認められなかった。集団学習法を用いない場合には、組成割合のデータを用いた解析(モデル 1,5,10)は、rarefying 法(モデル 2,6,11)と比較して変数重要度の AUC 値や判別力について劣るということはなく、EN,SPLSDA については変数が 300 の場合にはバギングを併用する方法(モデル 7,12)で変数重要度の AUC 値が上昇した。

・考察

集団学習法と rarefying 法を併用することで SPLSDA,EN については変数重要度の AUC 値が他の正規化法より大きな値を示した。RF では提案法により変数重要度の AUC 値が上昇しなかったが、その要因として、RF ではすでに内部で集団学習を用いているという点と、変数の相関関係を考慮して学習を行うわけではないため組成割合を用いる際の compositional data の影響を他の一般的な回帰モデルよりは受けにくいことが考えられる。RF は実際のデータ解析において多く用いられるが、EN と SPLSDA に提案法を用いることで RF の性能を上回る可能性が示唆された。判別力については、各モデルで大きな違いがみられなかったが、本研究では訓練データと検証データで各説明変数の存在比率が同等であるという前提を置いたため、変数重要度の正確性が判別力に反映されにくかったと考えられる。

他の手法の性能と比較し、提案法は変数が比較的少ない科(family)以上の上位階層のデータを用いる場合や、総カウント数のばらつきが大きいデータでは有用な方法となる。本研究では 3 手法のみの検討にとどまったが、説明変数間の関連を考慮して推定が行われる一般的な回帰モデルであれば本研究の結果は一般化可能性を持つことが想定される。ただ、本研究で行ったシミュレーション実験のパターンは限定的であるとともに、多変量データ解析手法もほかに多く存在するため、提案法の有効性を調べるためには今後より広範な実験や実データにおける検証が必要となる。

・結論

腸内細菌 16S rRNA データを用いて形質情報と関連する菌種を同定する関連解析手法において、集団学習法と rarefying 法を併用した方法を提案するとともに、各種多変量データ解析手法を用いて正規化法の性能評価を行った。結果、個体間での総カウント数のばらつきが大きい場合や変数が 100 程度の菌階層データでは、説明変数間の関連を考慮する一般的な回帰モデルに rarefying 法と集団学習法を併用する方法を用いることで応答変数と関連する菌種をより正確に特定できる可能性が示唆された。

表 1 本研究で検討した各種モデル

モデル	多変量データ解析手法	集団学習法	正規化法	変数重要度	判別力の評価法
1	RF	なし	組成割合	RFの変数重要度	形質有の予測確率
2	RF	なし	rarefying	RFの変数重要度	形質有の予測確率
3	RF	バギング	rarefying	RFの変数重要度の平均	形質有の予測確率の平均
4	RF	サブサンプリング	rarefying	RFの変数重要度の平均	形質有の予測確率の平均
5	EN	なし	組成割合	係数値	線形予測子
6	EN	なし	rarefying	係数値	線形予測子
7	EN	バギング	組成割合	係数値の平均	線形予測子の平均
8	EN	バギング	rarefying	係数値の平均	線形予測子の平均
9	EN	サブサンプリング	rarefying	係数値の平均	線形予測子の平均
10	SPLSDA	なし	組成割合	係数値	線形予測子
11	SPLSDA	なし	rarefying	係数値	線形予測子
12	SPLSDA	バギング	組成割合	係数値の平均	線形予測子の平均
13	SPLSDA	バギング	rarefying	係数値の平均	線形予測子の平均
14	SPLSDA	サブサンプリング	rarefying	係数値の平均	線形予測子の平均

表 2　各シミュレーションパターンの説明

パターン	総カウント数のばらつき	応答変数に対する他の要因の有無	説明変数の数
1	大きい	有	100
2	大きい	有	300
3	大きい	無	100
4	大きい	無	300
5	小さい	有	100
6	小さい	有	300
7	小さい	無	100
8	小さい	無	300

表 3　各モデルの変数重要度に関する AUC 値の平均値(標準偏差)

モデル		パターン1	パターン2	パターン3	パターン4	パターン5	パターン6	パターン7	パターン8
RF	1	0.681(0.083)	0.673(0.073)	0.679(0.083)	0.67(0.074)	0.698(0.071)	0.686(0.07)	0.688(0.081)	0.681(0.078)
	2	0.647(0.085)	0.635(0.081)	0.645(0.088)	0.638(0.069)	0.677(0.077)	0.663(0.071)	0.676(0.083)	0.664(0.091)
	3	0.648(0.08)	0.635(0.085)	0.649(0.085)	0.637(0.079)	0.675(0.078)	0.664(0.076)	0.678(0.082)	0.671(0.079)
	4	0.661(0.078)	0.65(0.081)	0.659(0.083)	0.65(0.075)	0.687(0.076)	0.679(0.073)	0.686(0.082)	0.682(0.077)
EN	5	0.745(0.072)	0.69(0.072)	0.699(0.068)	0.653(0.082)	0.754(0.066)	0.71(0.062)	0.711(0.071)	0.653(0.079)
	6	0.726(0.071)	0.667(0.075)	0.685(0.074)	0.629(0.075)	0.746(0.069)	0.692(0.067)	0.703(0.075)	0.647(0.084)
	7	0.702(0.07)	0.711(0.066)	0.678(0.076)	0.687(0.071)	0.725(0.063)	0.717(0.072)	0.681(0.08)	0.699(0.075)
	8	0.753(0.074)	0.746(0.082)	0.725(0.069)	0.724(0.077)	0.763(0.068)	0.749(0.073)	0.723(0.081)	0.73(0.073)
	9	0.761(0.07)	0.756(0.07)	0.727(0.071)	0.722(0.068)	0.769(0.067)	0.751(0.063)	0.731(0.071)	0.715(0.076)
SPLSDA	10	0.651(0.091)	0.641(0.097)	0.641(0.084)	0.632(0.09)	0.658(0.084)	0.634(0.086)	0.634(0.08)	0.632(0.092)
	11	0.676(0.08)	0.643(0.096)	0.643(0.082)	0.629(0.08)	0.67(0.08)	0.655(0.087)	0.658(0.08)	0.639(0.095)
	12	0.703(0.074)	0.724(0.067)	0.678(0.069)	0.694(0.071)	0.715(0.066)	0.724(0.073)	0.684(0.076)	0.708(0.069)
	13	0.745(0.072)	0.744(0.075)	0.719(0.076)	0.716(0.073)	0.756(0.067)	0.747(0.068)	0.718(0.082)	0.729(0.071)
	14	0.74(0.076)	0.742(0.078)	0.715(0.072)	0.724(0.071)	0.746(0.07)	0.746(0.07)	0.712(0.072)	0.721(0.076)