ビッグデータから科学的発見を導く統計手法

実験科学における幅広い利用が期待

津田 宏治
つだ こうじ
koji.tsuda@aist.go.jp

生命情報工学研究センター
主任研究員
（臨海副都心センター）

機械学習・データマイニングの手法開発と、その生命科学への応用を行っています。増加の一途を続ける複雑なデータから有用な知識を発見することは容易なことではありません。この記事で紹介しましたように、対象が複雑になると発見された知識の信頼性を示すことが難しくなります。今後も数理的な立場から、広く応用可能な統計手法の開発を目指していきます。

関連情報：

● 共同研究者

寺田 愛花、瀬々 潤（東京工業大）、岡田 眞里子（産総研）

● 参考文献

A. Terada *et al.*: *Proceedings of the National Academy of Sciences* (2013).

● 用語説明

＊ FDR：False Discovery Rate の略。発見された対象のうち誤っているものの割合を指す。

● プレス発表

2013年7月23日「ビッグデータから新たな科学的発見をもたらす統計手法を開発」

● この研究開発は、JST ERATO 湊離散構造処理系プロジェクトの支援を受けて行っています。

ビッグデータのパラドックス

自然科学では新しい現象を見つけたとき、誤発見の確率を示す検定値（P値）が計算され、あるしきい値（一般には、0.05）以下の場合にのみ、信頼しうる科学的発見として認められます。観測できる対象が増えると誤発見の確率も高くなるため、発見の基準を厳しくしなくてはなりません。多重検定法の中で最もシンプルでよく用いられるボンフェローニ法では、n個の対象があれば、P値にnを掛けて補正し、それでも0.05以内であれば、発見として認めます。その結果、観測対象が増えたのに、科学的発見が減るという奇妙な現象「ビッグデータのパラドックス」が起きる場合があります。

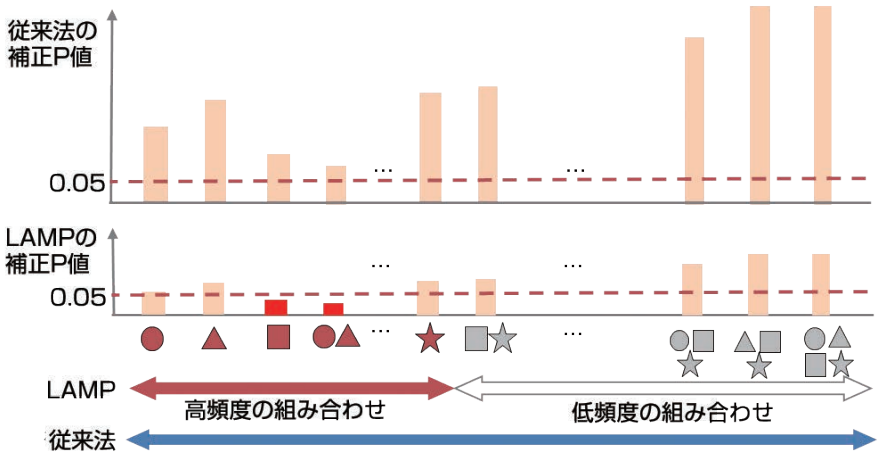
超高速アルゴリズムにより発見力を大幅に改善

私たちは今回、これまでよりも格段に正確な補正P値を計算できるアルゴリズムLAMP（Limitless-Arity Multiple testing Procedure、無限次数多重検定法）を開発しました。ボンフェローニ法では、全ての組み合わせ因子の数を補正係数として用いるのに対し、LAMPでは、出現頻度の低い組み合わせは誤発見率を変化させないという数理的性質に注目し、超高速アルゴリズムを用いて無為な出現頻度の低い組み合わせを特定し取り除くことによって、補正係数を大幅に削減しています（図）。またLAMPでは通常のボンフェローニ法と比べて、統計的な検定の精度を保ったままで、補正係数を十分に低くすることができます。この手法を用いて、ヒトの乳がん細胞株の遺伝子発現データを再解析したところ、これまで見過ごされてきた、最大8個の転写因子の組み合わせが乳がん細胞の増殖に関与していることを発見できました。

出現頻度の低い組み合わせが誤発見率を変化させないという事実は、1990年に米国のタローネによって明らかになっていましたが、アルゴリズムを用いて、それらを実際に数えあげて、生命科学データに適用したのは世界初です。生命科学で広く用いられているFDR*による方法では、誤発見率については妥協することで、発見力を高めていますが、この手法ではそのような妥協をせず、アルゴリズムのみによって発見力を大幅に高めることに成功しました。

今後の予定

この成果により、転写因子の組み合わせ効果の研究をはじめ、複数の遺伝子が原因となっている疾患の同定や多数の部位が関わる脳の高次機能の解明など、複合要因に起因する現象の解明が加速されることが期待されます。



従来法と LAMP の比較
赤色で示した組み合わせ因子のみが発見として認められる。