連載　統計的機械学習ことはじめ

廣野　元久　著

第１回　ビッグデータの可視化

今回から６回にわたって、統計的品質管理の立場から機械学習の基礎的な方法の考え方を紹介します。最初はビッグデータのグラフィカルな取り扱いの基礎です。

ビッグデータをグラフにするための注意点

データ分析の鉄則は「データをグラフにして観の目で眺めよう」です。そのためのパッケージがQC７つ道具です。しかし、ビッグデータを取り扱う場合はグラフの取り扱いを間違えると分析の見通しを曇らせる場合があります。以下に３つのリスクをまとめました。

１）得られたデータの出所に１つの正規分布を仮定することのリスク

２）個体数nが多い場合に散布状態をプロットすることのリスク

３）特徴量(変数)pが多い場合に散布図行列で俯瞰することのリスク

１）得られたデータの出所に１つの正規分布を仮定することのリスクへの対応

ビッグデータは実に様々な素性の結果系データを扱うので、データのクラス分けが必要となります。得られたデータは１つの母集団、たとえば、「正規分布から得られたもの」と考えるには無理があります。品質管理では「混ざったものは後から層別はできない」と教えられますが、場合によっては複数の集団が混在したカオスな状態から後知恵でも意味がありそうなパターンを見つけることは重要です。

ここでは、ヒストグラムの例を紹介します。ビッグデータではデータの中にいくつのクラスが含まれているのかを調べることがあります。このときに役立つのが正規混合分布という考え方です。図１はあるデータのヒストグラムで、左から順に、単一分布・２重正規混合分布・３重正規混合分布・５重正規混合分布をあてはめたものです。正規混合分布のあてはまりの良さは適合度指標で判断します。適合度指標はデータにモデルがどの程度フィットしているかを定量的に示す指標で、対数尤度やAICを改良したAICcなどが使われます。対数尤度やAICcは小さい値ほど当てはまりが良いことを示しています。この例では５重正規混合分布のAICcが最小ですから、５つの正規分布が混合したモデルがデータに一番フィットしていることがわかります。この正規混合分布という考え方は、ヒストグラムに限らず、折れ線グラフや散布図などにも応用できる方法ですが、それは別の機会があれば紹介したいと思います。

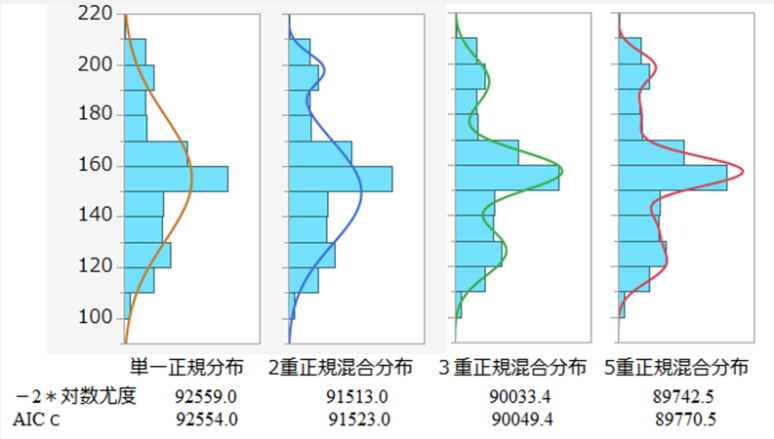

図1　あるデータのヒストグラムと正規混合分布の当てはめの様子

（次ページへつづく）

２）個体数nが多い場合に散布状態をプロットすることのリスクへの対応

散布図では個体数nが増えると多くの打点が重なり、黒塗りの絵のようになってしまい、データの少ない部分とデータが密集している部分の見分けがつきにくくなります。ビッグデータに隠れているパターンを見つけるには、個々の値を打点するよりも、領域内にあるデータの密度に着目したほうが見通しはよくなります。

図２左は相関を調べるために散布図を描いたものです。１万個のデータが打点されており、負の相関が読み取れます。しかし、散布図からは複数のクラスを発見することはできません。図２右はデータの密度で等高線を描いたグラフです。この地図のようなグラフは等高線図と呼ばれ、散布図では見えなかった２つのクラスを発見できます。なお、相関の有無を検定で評価することは危険です。個体数nが 多い場合は相関係数の絶対値が小さくても検定で有意になってしまうのです。

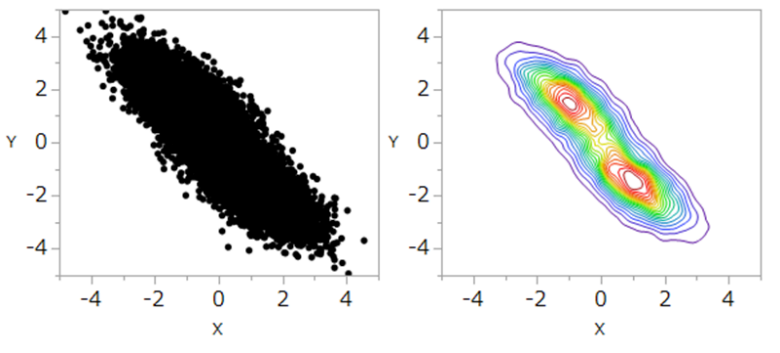

図２　散布図(左)と等高線図(右)

３）特徴量(変数)p が多い場合に散布図行列で俯瞰することのリスクへの対応

ビッグデータでは扱う特徴量pも多くなります。特徴量が増えると、特徴量間の相関の強さを一覧にして調べる散布図行列を作ると大変なことになり、たたみ一畳は優に超えてしまいます。膨大な特徴量間の相関の状態を俯瞰するには別のグラフ表現が必要です。その方法がカラーマップです。

カラーマップは特徴量の数 $p \times p$ 個のセルで構成され、各セルには相関係数の値で色分けしてグラデーションを作ります。こうすることで、正相関の強いところ、負相関の強いところ、相関の弱いところを色の違いで表すことができます。２つの特徴量の様子を散布図で表すことはできませんが、特徴量の数pが大きくなっても、１枚のグラフで表示できるので便利です。

著者紹介

廣野　元久（ひろの　もとひさ）

1984年（株）リコー入社。以来、社内の品質マネジメント・信頼性管理の業務、SQCの啓蒙普及に従事、品質本部QM推進室長、NA事業部SF事業センター 所長を経て、現在、（株）リコー倫理審査委員会 委員。

東京理科大学工学部経営工学科 非常勤講師（1997～1998年）、慶應義塾大学総合政策学部　非常勤講師（2000～2004年）。（一財）日本科学技術連盟　多変量解析法運営委員会委員、講師。