クラスター数決定法についての研究

2005MM066 志津綾香
指導教員：松田眞一

1　はじめに

私はこれまで、いくつかの解析法を学んできたが、特にクラスター分析は、クラスター数を決定させるための判断基準が曖昧であるため、理論的に決定付ける方法はないものかと思い、この研究を行うに至った。本研究ではクラスター分析の種類、各方法について学んでいくと同時に、クラスター数を自動的に決定させる方法を実際にプログラミングし、それらを使ってシミュレーションを行い、クラスター数自動決定法の有用性や、各計算方法での特徴を調べていく。

2　クラスター分析

クラスター分析とは、2 つ以上のデータがあるとき、類似度を手がかりにデータをいくつかのグループに分類させる方法である。その分類方法には、距離を基準に使うサンプルクラスターと、相関係数を基準にする変数クラスターがある。距離の指標の種類には、ユークリッド距離やマハラノビス距離等がある。また、クラスター分析には階層的方法と非階層的方法の 2 つの計算方法がある。さらに階層的方法には、距離の定義の仕方によって、主に最短距離法、最長距離法、群平均法、重心法、メジアン法、ウォード法等があり、非階層的方法には、K-means 法がある。（菅 [3], 渡辺他 [5] 参照）

3　階層型クラスタリング手法

階層型手法は、最初に各サンプルを 1 つのクラスターとして、最も近いサンプルから順に合併させ、新たなクラスターを形成していく方法である。結果は樹形図で表示させることができ、似ているものから順に並べられている。そしてクラスター数は、その樹形図を見て決定させる。クラスター分析において、クラスター間の距離の定義が重要になってくる。（神嶌 [2] 参照, 菅 [3], 渡辺他 [5]）

3.1　最短距離法

最短距離法とは、各クラスター間の距離における最短距離を、クラスター間の距離とする方法である。この方法は、小さいクラスターを徐々に集めていくもので、データをいくつかのクラスター数に分けるというよりは、主流となるクラスターを発見するのに役立つ方法である。また、異常値の発見も可能である。

3.2　最長距離法

最長距離法とは、各クラスター間の距離における最長距離を、クラスター間の距離とする方法である。この方法は、1 つのクラスターが極端に大きくなるのを抑えられ、大きさのそろったクラスターを得ることができる。

3.3　ウォード法

ウォード法は、クラスターとしてサンプルをまとめるときに生じる、各サンプルの情報の損失量の増加分をクラスターの距離とする方法である。すべてのクラスター内の偏差平方和の和を出来るだけ小さくするように組み合わせていくので、比較的まとまりのあるクラスターがいくつか得られる。

4　非階層型クラスタリング手法

非階層的手法は、あらかじめいくつのクラスター数にするかを決めておき、その数に従ってサンプルを振り分けていく方法である。出来るだけクラスター間の距離は大きく、各クラスターのサンプル間の距離は小さくなるようにサンプルを振り分けていく方法であるので、サンプル間に包含関係がないことが多い。非階層的手法は、計算量が膨大な為、処理時間が長くなるのが欠点である。（神嶌 [2], 菅 [3], 渡辺他 [5]）

4.1　k-means 法

K-means 法は、あらかじめクラスター数を決めておき、各サンプルを分けていく方法である。クラスターに含まれる各サンプルとそのクラスターの重心の距離が、他のどのクラスターの重心よりも小さくなるように求める。

5　クラスター数自動決定法

クラスター数自動決定法は、Jain[1](Ngo et al.[4] 参照) で紹介されている以下の数式を用いた。

$$p(k) = \frac{1}{k}\sum_{i=1}^{k} \max_{1\le j\le k}\left\{\frac{\eta_i + \eta_j}{\xi_{ij}}\right\}$$

ここで

$$\eta_j = \frac{1}{n_j}\sum_{i=1}^{n_j} D(F_i^{(j)}, \mu_j)$$

$$\xi_{ij} = D(\mu_i, \mu_j)$$

である。

このとき、μ_i はクラスタ i の平均ベクトルで、ξ_{ij} はクラスタ i とクラスタ j の平均距離、$F_i^{(j)}$ はクラスタ j 内の i 番目のベクトルで、n_j はクラスタ j 内のベクトルの個数を表している。$p(k)$ の値を、ある範囲内で一番小さくなるようにとる k が、最も最適なクラスター数となる。ある範囲内とは、この場合、スタージェスの公式を用いることにした。つまり、$(2 \le k \le 1 + \log_2 n)$ である。

6　プログラム

上記の数式を基に、階層的手法の場合と非階層的手法の場合のプログラムを作成した。階層的手法の場合は最長距離法、ウォード法を用いる。階層的手法の場合は k-means

法を用いる。データを指定し、階層的手法の場合は距離の定義をする。出力結果はデータ、k の範囲内における $p(k)$ の値、最適なクラスター数 k の値、それに対応する最適な $p(k)$ の値、求められたクラスター数 k の場合の各変数の群分け、を表示させるようにした。

7 シミュレーション

シミュレーションでは、2 次元の正規乱数を 3 つ用意し、各データの平均を 1 つの正三角形となるように置いたものを、データとして使用する。また、繰り返し数 200 回程で 1000 回繰り返した時とほぼ変わりない結果が得られることが分かったので、以後の繰り返しは 200 回とすることにした。また、1 つの群のデータ数は 100 個とする。つまり、1 回の実験に使われるデータ数は 300 個である。以上のような条件で、分散や相関係数を変化させたときの、各方法における最適なクラスター数の変化の動きを調べていく。以下の図がその時の結果である。

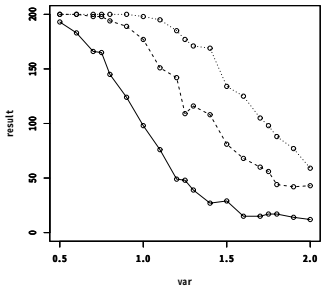

図 1 分散を変化

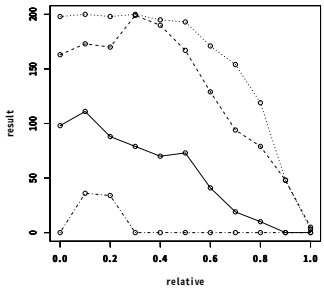

図 2 相関係数を変化

分散を変化させた時 この時の結果は、グラフからもわかるように、優れている方から k-means 法、ウォード法、最長距離法となった。

相関係数を変化させた時 相関係数を変化させると、データの重なりが少なくなっていくので、最短距離法で試すと何かわかるかもしれないと思い、調べてみることにした。しかし、相関係数を変化させても、分散を変化させたときと同様、全体的に k-means 法が一番性能が良く、その次にウォード法、最長距離法となり、その 3 方法と比べると最短距離法は劣る結果となってしまった。

8 考察

分散や相関係数を変化させ、クラスター数自動決定法がどのように変化していくのかを見てきた結果、どちらの場合も k-means 法が一番優れた結果になった。

しかし、相関係数を 1 まで変化させた時、データでは、はっきりと 3 群に分かれているにも関わらず、どの方法もほとんどクラスター数 3 になることがなくなってしまったのが気になった。そこで、クラスター数自動決定法に従った場合、データはどのような群分けになり、どの程度当てはまっているかを見てみることにした。これらの結果から、以下のようなことが分かった。

クラスター数自動決定法の結果では、k-means 方がダントツに良いという結果になっている。しかし中身をみてみると、クラスター数自動決定法の結果には反映されないが、それぞれに得意な条件や特徴があることが分かる。

まず、ウォード法、k-means 法は、データにかなりの散らばりがあっても、データに沿って、正確に分けることが出来る。しかし、その 2 方法の長所である樹形図のまとまりの良さから、クラスター数自動決定法では、3 群それぞれをさらに 2 群ずつくらい細かく分けた結果が最適となってしまった。

次に、後から付け加えた最短距離法だが、クラスター数自動決定法のみの結果では全く健闘できなかった。しかし予想通り相関係数 1 の時だけ、ほぼ正確に当てることができた。クラスター分析において最短距離法は、偏りのない時は適さないが、他の 3 方法が苦手な偏りの強い場合適していることが分かった。

また、最長距離法は、ウォード法や k-means 法と比べると、随時劣る結果になってしまったが、相関係数が 1 の時は最短距離法には及ばないものの、その他の 2 方法と比べるとかなり正確にデータを分けられていた。

9 おわりに

クラスター数自動決定法について研究したが、クラスター分析における様々な方法の特徴を改めて感じた。シミュレーション結果からもわかる通り、k-means 法は、クラスター数自動決定法において幅広くカバーしているが、あまりにも散らばりや相関が強いときは、クラスター数自動決定法は使えないことが分かった。実際のデータで、クラスター数自動決定法のみで判断させるのは、やはり難しいと感じた。やはり統計学では、どの解析方法においても言えることであるが、結局は、データの特徴や、解析者の求めている結果にあわせて方法を選び、解析する必要がある。とは言っても、人工の何らかの基準のあるデータであれば、k-means 法の場合のクラスター数自動決定法の正解率はかなり良いものだった。これくらいの精度の、階層的手法の場合のクラスター数自動決定法についても研究したかった。

参考文献

[1] A.K.Jain,Algorithms for clustering Data. Englewood Cliffs, NJ：Prentice-Hall, 1988.

[2] 神嶌 敏弘, データマイニング分野のクラスタリング手法 (1) -クラスタリングを使ってみよう！-, 人工知能学会誌, vol.18, no.1, pp.59-65 2003.

[3] 菅 民郎:『多変量解析の実践（下)』現代数学社, 京都, 1993.

[4] C.W.Ngo,T.C. Pong ,H.J. Zhang : On clustering and retrieval of video shots through temporal since analysis,IEEE Trans. MIt., 4-4,446/458, 2002.

[5] 渡辺 洋, 南風原 朝和, 大塚 雄作, 石塚 智一, 山田 文康, 藤森 進, 前川 眞一：『心理・教育のための多変量解析法入門-基礎編』福村出版, 東京,1988.