クラスター数決定法の比較

M2009MM020 志津綾香
指導教員：松田眞一

1　はじめに

多くの統計解析の調査研究では、類似度の評価を頼りに様々な分析を行う。その中でもクラスター分析は、類似度を評価するための基本的な方法であり、誰でも行うことができる数量的方法である。しかしクラスター分析は、解析者自身で適当なクラスター数を決定するため、その判断基準が曖昧である。そこで本研究では、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、比較、評価を与えることを本研究の目的とする。

2　クラスター分析

クラスター分析は、2 つ以上のデータがあるとき、類似度や距離 (非類似度) 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。まず、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。

階層的方法の分類方法には、距離を基準に使うサンプルクラスターと相関係数 (類似度) を基準に使う変数クラスターがある。階層的手法で主に用いられる方法に、最短距離法、最長距離法、群平均法、重心法、メジアン法、ウォード法とよばれるものがあり、これらは距離を基準に用い、その定義の仕方によってクラスター生成法が違う。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。階層的クラスタリング手法は、一般的にはユークリッド距離が使用される。つまり一般的に階層的方法は距離とクラスター生成法の組み合わせで方法が決定する。

また非階層的方法には、k-means 法、超体積法と呼ばれるものがある。非階層的手法の場合、ある評価関数を基準にクラスターを生成する。(菅 [5], 渡辺他 [8] 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 [4] 参照, 菅 [5], 渡辺他 [8])

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 [4], 菅 [5], 渡辺他 [8])

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。

4.2　超体積法 (Hypervolume method)

点集合を凸多面体の集まりとみなし、その体積を最少にすることで、最適な分割をみつけていく非階層的手法である。それ自体がクラスター数決定法としても使われる。

5　クラスター数自動決定法

ここでは、本研究でクラスター数自動決定法の精度を比較、評価するために、シミュレーションを行う方法を紹介する。

5.1　Jain and Dubes のクラスター数決定法

Jain and Dubes[3](Ngo et al.[7] 参照) が提案した以下の基準値を使う方法である。(以下では、JD 法と略す。)

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

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、$(2 \leq k \leq 1 + \log_2 n)$ である。

5.2 x-means 法

本研究では石岡 [2] によって改良された x-means 法を用いる。x-means 法は、k-means 法の拡張である。あらかじめクラスター数を決めておかなければならない k-means 法とは違い、最適なクラスター数を推測することができる。x-means 法の考え方は、k=2 で再帰的に k-means 法を実行していくというもので、クラスタの分割前と分割後で BIC 値 (ベイズ情報量基準) を比較し、値が改善しなくなるまで分割を続ける。つまり、分割前のベイズ情報量を BIC、分割後のベイズ情報量を BIC' とし、

- BIC>BIC′ ならば 2 分割する
- BIC≦BIC′ ならば 2 分割しない

を全ての場合で 2 分割出来なくなるまで繰り返すことにより最適なクラスター数を決定する。

詳しいアルゴリズムはここでは省略する。

5.3 Upper Tail 法

この方法は Mojena[6] によって提案され、階層的方法における重要なクラスター数決定法である。以下、Tail 法と略す。統計的な停止規則を用いてクラスター数を求める。その方法は、大きさ n の標本に対してクラスターを生成するための基準値 α が $n-1$ 個あるのを利用する。ここでは基準に距離のみを考えるのですべてが一つになる距離 α_1 から初めて降順に α_{n-1} までの基準値がある。この α の分布の平均と標準偏差を計算することによって、有意な α を導くことでクラスター数を決定する方法である。停止規則は、$j = 1$ から始めて条件

$$\alpha_j \leq \bar{\alpha} + k s_\alpha$$

を満たすまで j を増加させることである。停止した j が最適なクラスター数となる。$\bar{\alpha}$ と s_α はそれぞれ α の分布の平均と不偏分散の平方根をとったものである。この方法は最短距離法、最長距離法、群平均法、ウォード法で利用可能である。もし、この不等式を満たす α が無い場合、クラスターは 1 つとみなす。

k の値については、Mojena[6] では、2~4 の数を使っている。その時に用いたデータは、データの総数が 60~120 のものを、2~4 に分割したものである。本研究では、さらに 1 群のデータ数が多い場合など、k の値について掘り下げていきたい。

5.4 Upper Tail 法の改良

本研究では、Upper Tail 法の改良を提案する。Tail 法の停止規則を計算する時、従来の方法ではデータが正規分布に従うよう求めているが、カイ 2 乗分布に従うように計算する。一般的に多変量正規分布に従う 2 点間の距離の分布はカイ 2 乗分布に従う事が分かっているからである。

具体的な計算方法は、以下のように α を正規化し

$$\alpha' = \Phi^{-1}(F_p(\alpha / s_\alpha \cdot p))$$

それに前節の Upper Tail 法を用いている。ここで Φ^{-1} は正規分布の逆関数であり、F_p は自由度 p のカイ二乗分布の分布関数である。

5.5 その他

その他にも超体積法の判断基準を用い、凸包を拡張することで最適解を推測する凸集合の推測に基づく方法や、尤度比を用いたクラスター数決定法、移動平均品質管理規則、マリオットのテスト、ウォルフのテスト等がある。(Hardy[1] 参照)

6 プログラム

JD 法は上記の数式を基に、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。非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。

x-means 法については、石岡 [2] の、既に実装されているプログラムを用い、JD 法と同様の結果を表示させる。

Tail 法も JD 法と同様の結果を出すプログラムを作成した。Tail 法は階層的手法にのみ適用可能なので、最長距離法とウォード法の 2 方法のみを用いる。また今回は Tail 法を元に、停止規則を計算をする時のデータがカイ 2 乗分布に従う場合を提案したので Tail 法では従来の正規分布に従っているときの 2 方法の場合とカイ 2 乗分布に従っている場合の 2 方法、計 4 パターンで作成した。

7 シミュレーションとその結果

シミュレーションでは、図 1~4 の大きく分けて 4 パターンのデータ使用する。2 次元の乱数を用意し、各乱数の平均を正三角形、直線、十字型となるように置いたものを、データとして使用する。乱数は、正規分布、t 分布、ラプラス分布、対数正規分布に従うものを使用する。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、3 群の時の 1 群のデータ数は 100 個、5 群の時の 1 群のデータ数は 50 個とする。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。また、Tail 法についてのみ k の値を変化させた時の実験をし、Mojena[6] でも明記されていない k の値について、考察する。以後、主なシミュレーション結果を述べる。

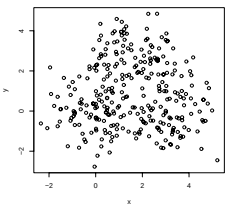

図 1　3 群正規分布: 三角形

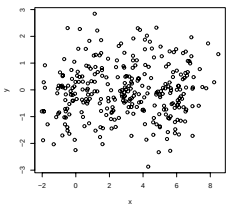

図 2　3 群正規分布: 直線上

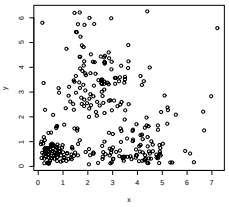

図 3　3 群対数正規分布

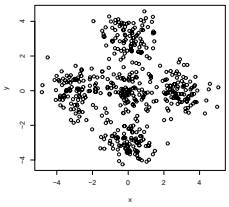

図 4　5 群正規分布

7.1　正規分布に従う三角形データの分散変化時

JD 法を用いた場合で一番良い結果となったのは、k-means 法である。Tail のカイ 2 乗分布を用いた時の最長距離法と ward 法が、分散が強くなった時にも耐えて良い結果を出している。

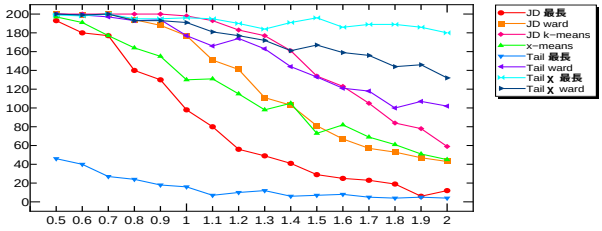

図 5　正規分布:分散を変化させた場合

グラフは省略するが、分散を 1 で固定し相関を 0 から 1 まで変化させた場合も、結果は似たものになった。k-means 法は極端なデータでなければ安定して良い結果が得られる。しかし相関が 0.7 を超え、データの形が極端になると、Tail のカイ 2 乗分布を使った最長距離法と ward 法が強くなる。

7.2　3 群の直線上に並ぶデータの相関を変化させた場合

分散 1 で固定した場合、データの重なりはそこまで大きくなくても、あまり離れていないため、Tail のカイ 2 乗最長距離法、カイ 2 乗 ward 法の 2 方法がほとんど正解するのに対し、それ以外は全く及ばないという極端な結果になった。一方、分散を 0.5 で固定した場合、相関をいくつに変化させても、常にくっきり 3 群に分かれているデータが得られる。その時の結果では、k-means 法、x-means 法、Tail の最長距離、カイ 2 乗の最長距離が 7,8 割の正解率を得られるようになった。これらの 4 方法は、明らかにデータが分かれているような場合でないと、判別するのが難しいことが分かる。また、JD 法を用いた場合の ward 法は、三角のデータなどではそれなりに良い結果を残しているが、データパターン 2 のように縦長に広がるデータの場合が、かなり苦手と言えるだろう。

7.3　3 群で他の分布に従う場合

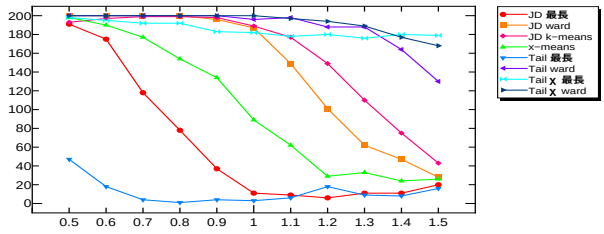

図 6　対数正規分布:標準偏差を変化させた時

対数正規分布の場合も、t 分布の時同様、他の方法が正解率を下げる標準偏差 1.2 以降においても、Tail のカイ 2 乗分布を用いた最長距離法と ward 法、Tail の ward 法が圧倒的な強さを見せた。

7.4　3 群それぞれデータ数の違う場合

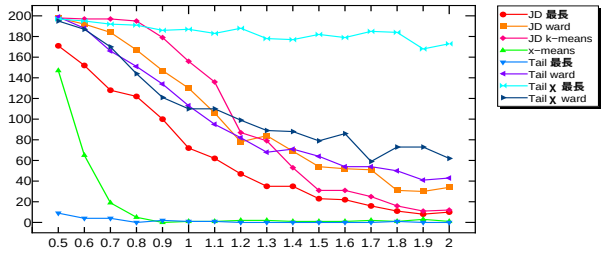

図 7　正規分布:分散を変化させた時

Tail のカイ 2 乗分布を用いた最長距離法が圧倒的な強さをみせた。しかしこれは k の値がかなりフィットしたからだろう。だからといって、ward の方は徐々に精度が下がっているので、k の値抜きにしてもカイ 2 乗分布を用いた最長距離法が、データ数が違う場合に強い事が分かる。

7.5　k の値

正規分布に従うデータにおいて、k の値を変化させた。t 分布、対数正規分布に従う場合においても k を変化させたが、ほとんど同じ結果になった。5 群の場合も同様である。よって k の値は分布よりも各群のデータ数に依存していると言える。

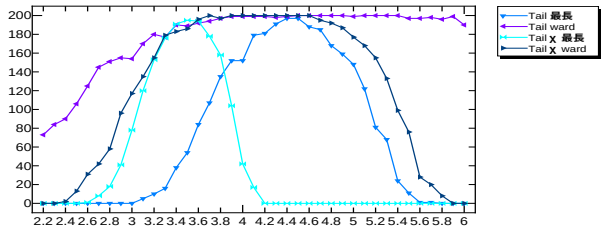

図 8　データ数 50 × 3 群の正規分布に従うデータ

実際の解析では、各群のデータ数は分からないので、データの総数と、解析者の予想、希望するクラスター数を想定し、k の値をある程度絞ってから使えば、Tail 法の精度が上がるだろう。今回私の行ったシミュレーションでは、一環して $k = 3$ と $k = 4.5$ を用いたが、最長距離法か ward 法かでも、最適な k の値が変わってくる。各群のデータ数が 30 から 50 前後の場合、$k = 3$ 程度がちょうどよく、各群のデータ数が 100 前後になると、$k = 4.5$ 程度がちょうどよくなってくる。データ数がもっと多くなるに従い、k の値も大きくする必要があることが分かった。全く結果が予測出来ないような場合は、正解率をとる範囲がかなり広い ward 法が有利である。実際、カイ 2 乗分布を用いた場合とほとんど同じ程度の精度が得られる。

7.6 それぞれの方法における結果

JD 法を用いた場合、最長距離法では、どの場合もすごく良い結果が出たわけではなかった。また、JD 法よりも Tail 法の方が合っているのかもしれない。

ward 法は、k-means 法には劣るが、たいていの場合で高い正解率を得られる。縦長なデータに弱く、まとまりのある場合であれば何群でも何分布でも、あまり影響を受けずに良い結果が得られる。

k-means 法は、ウォード法と同じく、かなり安定して高い正解率で最適なクラスターを求めることが出来る。JD 法で試した中では、1 番良い結果が出た。データの分布が偏っていたり、相関や分散が少し強めな場合は正解率を下げてしまうが、それでもデータが明確に分かれている場合などは、しっかりあてる事が出来る。

x-means 法は、全体としては、あまりよい成績は得られなかった。目で見て明確に判断できるようなデータは分けることができるが、そうでない場合は、信頼できないと感じた。

Tail 法を用いた場合、最長距離法は、ほとんど良い結果が得られなかった。唯一、他の方法が苦手とした直線上にデータが並ぶ時に良い結果を得た。k の高い正解率が得られる範囲もかなり少なく、k の値に結果が大きく左右され、クラスター数未知の場合に使うには不安要素が多い。

Tail 法を用いた ward 法は、今回実験した場合においては、ほとんどの場合で高い正解率が出た。

カイ 2 乗を用いた最長距離法は、各群の分散やデータ数が全て違う場合に、圧倒的強さを見せた。実際には、分布や 1 群の個数が分からないので、k の値を外さなければ、かなりよい精度で現実の場面で使えるかもしれない。また、カイ 2 乗分布を用いる前は、ほとんど機能しない状況だったので、十分改善されたと言える。

カイ 2 乗分布を用いた ward 法も正規分布使用時から少し改善することができ、全体としてかなり良い結果が得られた。他の方法でうまくいかなかった直線上に位置するデータの時にも、通用する。

8 考察

それぞれ得意、不得意があるが、クラスター数決定法として用いやすく実用可能と言える方法は、どんな場合にもある程度の対応が出来るものであると考える。そこで、実際の場面で有効だと思われる方法に、JD の k-means 法や Tail の ward 法、Tail のカイ 2 乗分布を用いた ward 法があげられる。また、うまく k の値がとれるようであれば、各群の分散やデータ数がばらばらでも、高い正解率が得られる Tail のカイ 2 乗分布を用いた最長距離法も実用的であると言える。k-means 法や ward 法などで、ある程度の各群のデータ数やクラスター数を予測してから、k の値を考えてもう一度 Tail のカイ 2 乗分布を用いた最長距離法で試すと、最適なクラスター数を求められる可能性が高くなるだろう。

9 おわりに

本研究では、クラスター数決定法が実際のクラスターの群数を当てることを善し悪しの基準にしてきたが、実際のデータに対する当てはまりに関してもっと言及出来ればよかった。しかし Tail 法の改良を提案し、良い成績を残すことができ、従来の方法を改善することができたと言える。ただそれらの方法も、最終的には、用いる k の値が正しければ良い結果を得る、という形になった。k の値に関しては、分けられたクラスター各群のデータ数に依存することが分かり、その動きも確認できたが、実際の分析でそれは分からないので、その辺りを更に工夫することで、クラスター数決定法がより実用的になるのではないだろうか。

参考文献

[1] Hardy, A.: On the number of clusters, *computational Statistics and Data Analysis*, **23**, pp.83-96, 1996.

[2] 石岡 恒憲: x-means 法改良の一提案 –k-means 法の逐次繰り返しとクラスターの再併合–, 『計算機統計学』, **18**(1), pp.3-13, 2006.

[3] Jain, A.K. and Dubes, R.C.: *Algorithms for clustering data*, Englewood Cliffs, NJ：Prentice-Hall, 1988.

[4] 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』, **18**(1), pp.59-65, 2003.

[5] 菅 民郎: 『多変量解析の実践（下）』, 現代数学社, 1993.

[6] Mojena, R.: Hierarchical grouping methods and stopping rules:an evaluation, *The Computer Journal*, **20**, pp.359-363, 1977.

[7] Ngo, C.W., Pong, T.C. and Zhang H.J.: On clustering and retrieval of video shots through temporal since analysis, *IEEE Trans. Mlt.*, **4**(4), pp.446-458, 2002.

[8] 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, 1988.