

自己組織化マップと情報量規準を用いた 高次元データによるクラスタリング

日大生産工 ○佐久間 潤 日大生産工 山内 ゆかり

1 まえがき

自己組織化マップ(SOM)を用いたクラスタリングでは、個々のクラスタのサイズやデータ密度が大きく異なる場合に、クラスタ境界の判定に伴う閾値の設定が困難になる。そこで、情報量規準を用いることで、クラスタ分割を適切に行うことができる。加藤らによる「自己組織化マップと情報量基準を用いたクラスタリング手法の提案」では、AIC、BIC での情報量規準を用いる研究を試みている。

2 実験方法

本実験は加藤による自己組織化マップと情報量基準を用いたクラスタリング手法に基づき、異なる情報量規準によるクラスタリング性能の違いを検証する。

SOM の基本学習アルゴリズムでは、勝利セルとその近傍に位置するセル群に対して、コードベクトルの更新が行われる。これを利用して 1 次元 SOM を作成する。

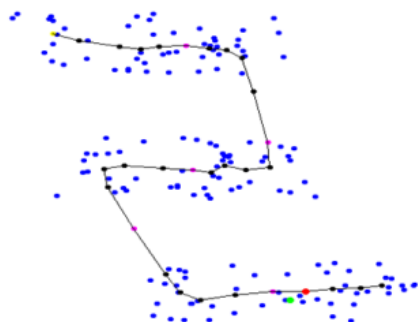


図1 学習後の 1 次元 SOM

SOM を用いたクラスタリング手法では、コードベクトル間の距離が大きく異なる部分をクラスタ境界とみなしてクラスタ抽出を行っている。

クラスタを抽出する資本として、セル番号を横軸とするグラフにコードベクトル間の距離をプロットした「データ密度ヒストグラム」を作成している。

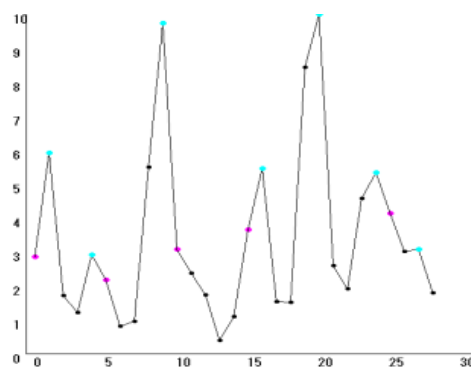


図2 データ密度ヒストグラム

「データ密度ヒストグラム」では、クラスタ境界部分以外にも複数のピークが現れる。これらのピークを機械的に検出することで、対象データからクラスタ候補を抽出することができる。ここで、何らかの評価基準に基づいて、これらのクラスタ候補を適切に併合していけば、対象データに含まれる本来のクラスタ群が抽出可能であると考えられる。

以上のことから、クラスタ候補の併合の際に用いる評価基準に情報量規準を採用し、以下の手順でクラスタリングを行っている。

- A. クラスタリング対象データを 1 次元 SOM に学習させ、学習結果からデータ密度ヒストグラムを作成。
- B. データ密度ヒストグラムから初期のクラスタ候補集合を作成し、個々のクラスタに対して、競合層のセルの並びに準じた通し番号を付ける。
- C. 通し番号が連続するクラスタ候補の任意のペアに注目し、クラスタ候補の選択的な併合を行う。このとき、併合すべきクラスタ候補のペアを情報量規準によって判断する。

情報量規準を用いたクラスタ候補の選択的な併合は、以下の手順によって行われる。

C1. 情報量規準を用い、番号が隣接するクラスタ候補を仮に併合した時の、 IC_{single} 、 IC_{twin} を算出する。それぞれ、仮併合後のクラスタに対して、単一分布モデルあるいは二分布モデルを当てはめた場合の情報量規準の値である。

C2. 当てはめた分布モデルの違いによる情報量規準の値の変化量 ΔIC を求める。

$$\Delta IC = IC_{single} - IC_{twin}$$

C3. 番号が隣接するクラスタ候補のすべての組み合わせについて手順1及び手順2を行ったうえで、 ΔIC が最も小さいクラスタ候補のペアを併合して1つの新たなクラスタ候補とする。その後、クラスタ候補全体の通し番号を再度付け直す。

単一分布モデルへの当てはめが、二分布モデルへの場合よりもっともらしい場合、 $IC_{single} < IC_{twin}$ となる。したがって、 $\Delta IC = IC_{single} - IC_{twin}$ の値の大小によって、クラスタ候補の併合の可否の度合いを測ることができる。

従来手法では、ベイズ型情報量規準(BIC)と赤沼情報量規準(AIC)を使用し、比較している。その結果、SOM単体やk-means法を用いた場合よりも、より本来のクラスタ分割に近いクラスタリングを安定的に行うことができ、エラーが少ない結果が出ている。

しかし、この2つの情報量規準の場合、データの次元数が多くなるにつれ、k-means法よりも劣ることが判明している。また、使用された2つの情報量規準を比較した場合、BICよりもAICのほうがより良い結果を示していた。これは、BICがクラスタのサンプル数を参照するモデルであるため、併合を続けていくとサンプル数がクラスタ候補ごとに大きく異なっていくためである。

本実験では、上記以外の情報量規準を用い、次元数が多い場合などに本来のクラスタ分割に近いクラスタリングを行えるか実験、検証を行った。

また、どの情報量規準を用いるかを考えるうえでAICのようにクラスタ併合を行った場合に、大きく変動するデータを用いない情報量規準を採用することで、より高い結果が出る可能性が高い。

3 実験結果および検討

情報量規準として、竹内情報量規準(TIC)のクラスタリングの検証を行った。

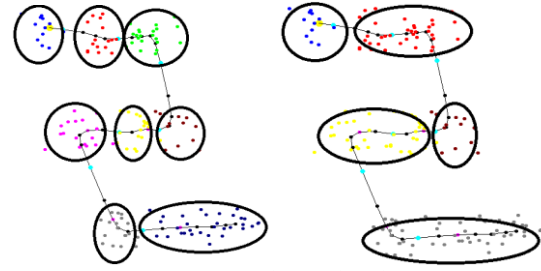


図3 TICでの併合前、併合後

図2の左図が併合前、右図が併合3回目である。図2では、色ごとにクラスタ候補が存在しているが、併合を繰り返すことで、クラスタ候補の数が少なくなっている。これを繰り返すことで、本来のクラスタ数まで、併合を行うことができた。

2次元SOMの場合、AIC、BICとTICでの結果の差は、ほとんど見られなかった。

4 まとめ

本研究では、SOMを用いたクラスタリング手法に対して情報量規準を適応し、検証を行った。

2次元SOMの場合、AIC、BICとTICの差はほとんど見られなかった。これは、クラスタ併合を行った場合のデータの変化が少ないためと考えられる。

しかし、クラスタ併合を行っていくと、クラスタ候補内のデータが多くなっていく。特に次元数が多い場合、この現象が顕著になっていく。

今後の課題として、ペナルティ項に、クラスタ併合するごとにクラスタ候補ごとのデータが大きく変動してしまうデータを使用しない情報量規準を用い検証を行っていく。

「参考文献」

- 1) T.Kohonen, Self-Organizing Map, Springer-Verlag, 1995.
- 2) 加藤 聡, 堀内 匡, 伊藤 良生, 自己組織化マップと情報量規準を用いたクラスタリング手法の提案, 知識と情報(日本知能情報フアジィ学会誌), Vol.21, No4, pp.452-460, (2009)
- 3) 小西貞則, 北沢原四郎, 情報量規準, 朝倉書店, 2004.
- 4) 赤池弦次, 甘利浚一, 北川原四郎, 樺島洋介, 下平英寿, 赤池情報量規準AIC, 共立出版, 2007.