

G-004

自己組織化マップと情報量規準によるクラスタ数の推定法に関する基礎的研究

A Basic Study on Clusters Estimation by using Self-Organizing Map and Information Criterion

加藤 聡¹

Satoru Kato

堀内 匡²

Tadashi Horiuchi

1. はじめに

自己組織化マップ (SOM) を用いたクラスタリング [1] は、学習後のコードベクトル同士の距離の変化に着目してクラスタ境界を検出する手法であり、 k -means 法などと比較して、初期状態の違いによる結果のばらつきが少ないことや、任意形状のクラスタ抽出が可能であることが特長である。しかしながら、隣接セル間のコードベクトル間の距離に基づく「距離ベース」のクラスタ抽出法であるため、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しきい値の設定が困難になるという問題がある。

一方、個々のデータ同士のユークリッド距離の変動に注目せずに、データ群の局所的な「まとまりの良さ」を評価してクラスタを見出す手法を考えることもできる。これは、データ集合の分布の状態に注目してクラスタとしてのもっともらしさを評価することから、「分布ベース」のアプローチと考えることができ、Pelleg ら [2] は、 k -means 法にベイズ型情報量規準 (BIC) [3] を用いた再帰的なクラスタ分割を導入した手法である x -means 法を提案している。

情報量規準を用いた分布ベースのアプローチは、SOM によるクラスタリング手法にも比較的容易に導入できると考えられる。そこで本稿では、SOM を用いたクラスタリング手法において、情報量規準に基づいたクラスタ抽出法を提案し、クラスタリング実験によって提案手法の有効性について述べる。

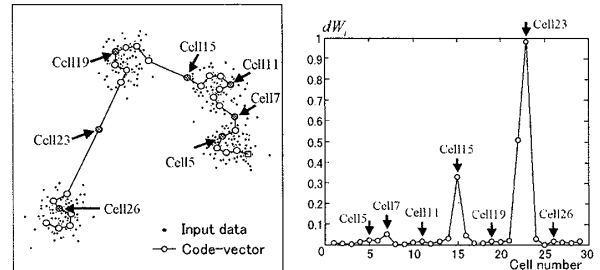
2. 提案手法

2.1 SOM を用いたクラスタ候補の抽出

Kohonen によって提案された SOM [4] の学習後に得られるマップでは、競合層上で隣接するセル間のコードベクトルが、データ空間上においても隣接しているという「位相保持写像」がなされており、さらに、入力データ空間でのデータの疎密が、学習後のコードベクトルの分布に反映されるという特徴がある (図 1(a) 参照)。ここで、セルが 1 次元的に並んだ 1 次元 SOM を学習に用いて、学習後にセル番号を横軸、隣接セル同士のコードベクトル間距離を縦軸とするような「データ密度ヒストグラム」を作成する。ヒストグラムにおける上向きのピークを機械的に検索していくと、クラスタをなすと思われるデータ集合を抽出することができる。

2.2 情報量基準を用いたクラスタ候補の併合

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。したがって、前述の方法で抽出されたデータ集合は、本来のクラスタの部分集合の



(a) 学習後のコードベクトルの分布

(b) データ密度ヒストグラム

図 1: 学習終了後の SOM のコードベクトル分布とデータ密度ヒストグラムの例 (4 クラスタデータ)

ようなものとなっている。提案手法では、クラスタ候補の併合の際に用いる評価基準として赤池情報量規準 (AIC) を採用し、以下に示す手順でクラスタリングを行う。

- 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する。
- 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生成し、個々のクラスタ候補に対して、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、AIC に基づいたクラスタ候補の選択的な併合を行う。

手順 C は、具体的には以下の手順によって行われる。

- C1. 2 つのクラスタ候補を仮に併合したとき、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 AIC_{single} と AIC_{twin} を算出し、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 ΔAIC を以下によって求める。

$$\Delta AIC = AIC_{single} - AIC_{twin} \quad (1)$$

- C2. 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 C1 を行った上で、 ΔAIC 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。

¹松江工業高等専門学校 情報工学科²松江工業高等専門学校 電子制御工学科

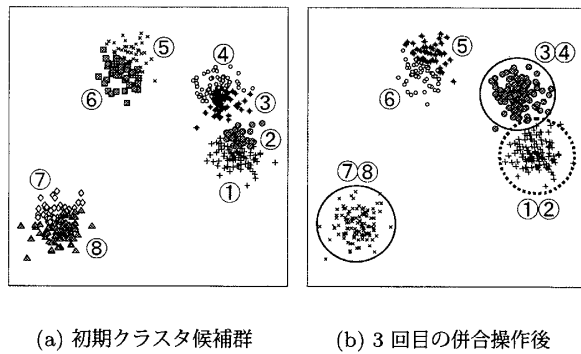


図 2: 情報量規準を用いたクラスタ候補の併合過程

C3. クラスタ数が指定の数に達するまで、C1～C2 の過程を繰り返す。

提案手法の実行例として、図 1(b) に示したデータ密度ヒストグラムから得られる初期のクラスタ候補群は図 2(a) のようになる。その後、手順 C を 3 回繰返した後のクラスタ候補の併合後の様子を図 2(b) に示す。

2.3 情報量規準を用いたクラスタ数の推定

クラスタ併合の判定に用いられる $\Delta AIC (= AIC_{\text{single}} - AIC_{\text{twin}})$ は、単一分布モデルおよび二分布モデルそれぞれにおける情報量規準の差分であり、2つのクラスタ候補を単一の分布とみなした方が良い場合に $AIC_{\text{single}} < AIC_{\text{twin}}$ となり、逆に 2つの分布とみなした方が良い場合には $AIC_{\text{single}} > AIC_{\text{twin}}$ となる。したがって、 ΔAIC の符号に着目すれば、クラスタ候補の併合処理を適切な段階で中断することができ、 x -means 法の場合と同様に、クラスタ数の自動的な推定が可能になると考えられる。

3. クラスタ数推定実験

実験では、図 1(a) に示した 4 クラスタからなる人工データと、Iris データセット [5] をクラスタリング対象データとして使用した。表 1 は、人工データに対してクラスタ併合を進めたときの、クラスタ併合のレベルとその併合レベルにおける各併合候補の ΔAIC の最小値 $\Delta AIC_{\min}$ を示したものである。このデータセットのクラスタ数は 4 であり、表 1 では、クラスタ数を 5 個から 4 個に併合するまでは $\Delta AIC_{\min}$ の値が負となり、4 個以下に併合する場合には $\Delta AIC_{\min}$ の値が正であることが分かる。

また、表 2 は、提案手法において SOM の初期状態を変化させて、上記のクラスタ数推定を 100 回行ったときの、推定されたクラスタ数とその頻度である。人工データに対しては、100 回中 99 回の試行において、正しいクラスタ数 (=4) が推定されている。なお、同様のクラスタ数推定を Iris データに対して行った場合、100 回中 81 回の試行において Iris のカテゴリ数に等しいクラスタ数 (=3) が推定された。ただし、クラスタリングの観点から、Iris データを 3 クラスとみなすことの妥当性については議論の余地がある [6]。

表 1: ΔAIC の最小値の変化

Number of clusters						
8→7	7→6	6→5	5→4	4→3	3→2	2→1
$\Delta AIC_{\min}$						
-113	-110	-102	-97.6	36.7	317	747

表 2: 推定クラスタ数の頻度

	推定されたクラスタ数の頻度			
	2 クラスタ	3 クラスタ	4 クラスタ	5 クラスタ
人工データ	0	1	99	0
Iris データ	3	81	2	14

以上のことから、 ΔAIC の符号に注目し、その値が負となるクラスタ候補のペアが存在しなくなった段階でクラスタ併合を止めることで、クラスタ数の推定と各クラスタの抽出を同時に実現できることが分かった。

4. まとめ

本稿では、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用することを検討し、SOM によって得られた初期クラスタ候補を、情報量規準 AIC に基づいて選択的に併合して行くクラスタリング手法を提案した。人工データおよび UCI の Iris データセットを用いたクラスタリング実験から、情報量規準を用いることによって、SOM を用いたクラスタリング手法においてクラスタ数の自動推定が可能であることが確認できた。

今後は、さまざまなデータセットに対して、さらに検証を重ねることが課題である。

参考文献

- [1] 寺島幹彦, 白谷文行, 山本公明, 自己組織化特徴マップ上のデータ密度ヒストグラムを用いた教師なしクラスタ分類法, 電子情報通信学会論文誌, Vol.J79-D-II, No.7, pp.1280-1290, 1996.
- [2] D. Pelleg, and A. Moore, X-means: Extending K-means with Efficient Estimation of the Number of Clusters, Proc. of the 17th International Conference on Machine Learning, pp.727-734, 2000.
- [3] 小西貞則, 北川源四郎, 情報量規準, 朝倉書店, 2004.
- [4] T. Kohonen: Self-Organizing Maps, Springer-Verlag, 1995.
- [5] UCI Machine Learning Repository, <http://www.ics.uci.edu/~mllearn/MLRepository.html>
- [6] D.-W. Kim, K.H. Lee and D. Lee, Fuzzy Clustering Validation Index based on Inter-cluster Proximity, Pattern Recognition Letters, Vol.24, pp.2561-2574, 2003.