

複数の競合層を持つ自己組織マップによるクラスタリング

日大生産工 ○森田由希 日大生産工 山内 ゆかり

1 まえがき

自己組織化マップ (SOM:Self-organizing maps) は T.Kohonenによって1981年ごろに発表された、教師なし学習のニューラルネットワークであり、与えられた入力情報の類似度をマップ上の距離で表現するモデルである。[1]

松下春奈、西尾芳文らの「二種類の SOM を用いたクラスタリングに関する研究」では、入力データが集中している部分を局所的に自己組織化する SOMl と、入力空間全体に広がる入力データを自己組織化する SOMg の特徴の異なる二種類の SOM を同時に使用し、偏りのある入力データに対する二種類の SOM の競合作用を調査した上で k-means 法と比較することで有効性を確認していた。[2]k-means 法では入力データにノイズが混ざっている場合、クラスタに所属するデータのみを正確に抽出することが困難であるとされている。そのため、二種類の SOM を用いるとノイズを多く含んだデータ、偏りのあるデータにたいしてノイズを除去することができるかとわかっている。しかし、完全にノイズを除去できているわけではなく、クラスタリング精度が 90%程度にとどまっている。そこで本研究では、近傍関数やしきい値、クラスタリング後の分類方法を変更することで分類精度の向上を図る。

2 提案手法

(Step1) 入力データを1600点、0~1の範囲でランダムに与える。

(Step2) SOMlのすべての重みベクトルの初期値を0~1の範囲で、SOMgのすべての重みベクトルの初期値を入力データの中心付近でランダムに、それぞれ与える。

(Step3) 入力ベクトルを並列的にSOMlとSOMgの各ニューロンに同時に入力する。

(Step4) 入力ベクトルとSOMl、SOMgとの距離を計算し、勝者ノードを求める。

(Step5) SOMl、SOMgの勝者ノードから、より入力ベクトルにより近いものを勝者ニュー

ロンとし、勝者ニューロンのあるSOMのすべてのニューロン重みベクトルを更新する。

(Step6) すべての入力データに対してStep1~Step5を繰り返す。

(Step7) ここでSOMgのみの学習を行う。

(Step8) Step1 ~Step4までと同様に入力、計算を行う。

(Step9)勝者ニューロンがSOMlであり、しきい値よりも距離が大きい場合、SOMgのニューロン重みベクトルを更新する。

(Step10)すべての入力データに対してStep8, Step9を繰り返す。

(Step11) 再度、入力ノードと各SOMとの距離を調べ勝者ニューロンを求める。

(Step12) 勝者ニューロンがSOMlの場合、計算した距離が任意の距離Rよりも小さければSOMlに対応するクラスタに分類する。

3 実験

図1に本件の実験に使う入力データを示す。

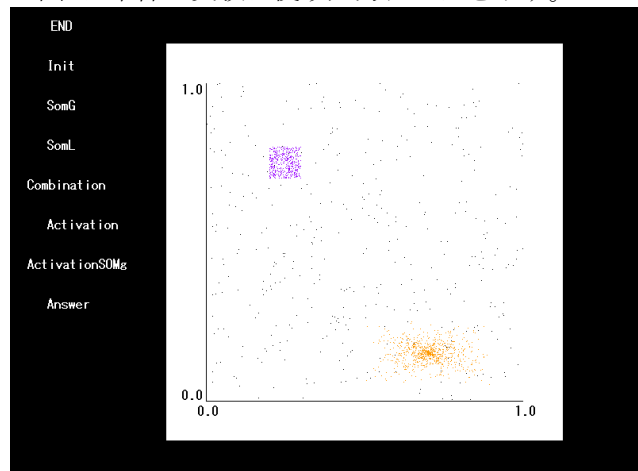


図1 入力データ

入力データ1600点、うち400点を横軸0.2~0.3、縦軸0.7~0.8の範囲で分布させこれをクラスタC1とする。800点を横軸0.2~0.6、縦軸0.2~0.3の範囲で分布させクラスタC2とする。残り400点を0~1の範囲でランダムに分布させノイズとする。

図2には入力空間全体に広がる入力データを自己組織化するSOMgと入力データが集中してい

The Clustering by Self-Organizing Map with Multiple Competitive Layers

Yuki MORITA and Yukari YAMAUCHI

る部分を局所的に自己組織化するSOMIの初期状態をしめす。SOMgは10×10で1つ、SOMIは10×10のものを2つの計300個のニューロンよりなる。

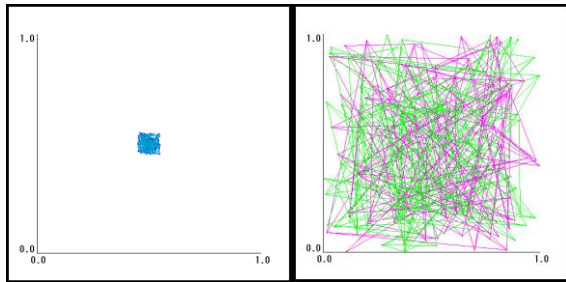


図2 SOMg、SOMIの初期状態

図3はそれぞれの入力データにたいして4回、計6400回のシミュレーションを行った後、SOMgのみのシミュレーションを6400回行った結果を示す。この結果からSOMgが入力空間全体をSOMIがデータの集中している部分をそれぞれクラスタリングしていることが分かる。

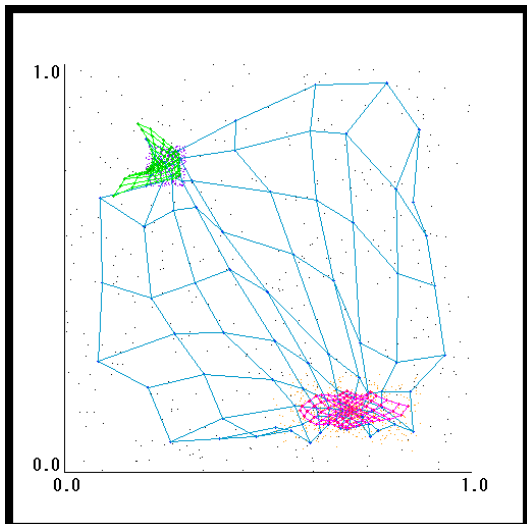


図3 シミュレーション後

クラスタリングが終了した時点で、それぞれの入力データを各SOMに分類する。SOMIに関しては、入力データと勝者ノードとの距離（R）が0.05以上のものは近傍距離の短いものであってもSOMIに分類せず、SOMgに分類するものとする。

表1 正答率

	C1	C2
従来手法	86.8	91.28
提案手法	96.5	91.3

表1の正答率から、どちらのクラスタに関しても従来手法より特徴抽出ができていことが分かる。

4 実験結果および検討

2つのクラスタに対して、それぞれのSOMIがデータの偏りのあるところを集中的にクラスタリングできていることが分かった。また、データが密集しているC1に関しては90%~98%のクラスタリング精度が保てていることが分かった。しかし、C2に関してはデータが広域に分散しているためか、正答率は90%前後にとどまってしまった。クラスタの分類方法はこれからさらに改善していく必要がある。SOMgに関しては、入力空間全体を覆うようにクラスタリングできているが、一部に偏りがみられる。

5 まとめ

それぞれのSOMが入力データの特徴に沿ってクラスタリングできていることが確認できているが、データが広域の場合に正答率が低いことが分かる。これは、SOMgが入力データ全体をクラスタリングするものであるのに、偏りのある部分に寄ってしまい、全体のバランスが悪いからであると考えられる。そのため、SOMgそのもののクラスタリング方法を見直し、さらにSOMIの特徴抽出の方法に関しても見直す必要があると考えられる。結果が向上した場合は2次元マップだけでなく3次元でも同様にクラスタ抽出ができるかどうかを考えていきたい。

「参考文献」

- [1] T.コホネン著,徳高平蔵,岸田悟,藤村喜久郎訳:自己組織化マップ,シュプリンガー・フェアラーク東京株式会社,1996
- [2] 松下春奈,西尾芳文,「二種類のSOMを用いたクラスタリングに関する研究」電子情報通信学会技術研究報告. NLP, 非線形問題 105(125), 51-56, 2005-06-16