ISSN 2186-5647

2-20

局所分布学習に基づくSOINNのスパースデータクラスタリング

日大生産工　○石川 正春　　日大生産工　山内　ゆかり

1. まえがき

近年の研究では教師なし学習は、事前にクラス数を決定する必要があり初期値に依存している。更に、実世界のデータは分布の形状が変化するなど非定常な場合があり、そうした分布の非定常性に対処するためには，追加学習の機能が必要となる。つまり、実世界のデータを適切に学習・認識するには、既存の学習結果に対して、新しいクラスからの入力や分布の形状変化などに対応し、動的にネットワークの構造を変化させることができる学習手法でなければならない。そこで長谷川らは、Self-Organizing Map(SOM)とGrowing Neural Gas(GNG)に着想を得て構築した教師なし学習手法の一種である自己増殖型ニューラルネットワーク(以下SOINN)参考文献1)を提案した。SOINNは、加学習可能なオンライン教師なし学習手法であり、事前にネットワークの構造を決定する必要がなく、高いノイズ耐性を有し、計算処理が軽いなどの特長がある。またSOINNの発展としてローカル行列学習とSOINNの利点を組み合わせたLocal-SOINNやILDNを提案されている。
Youlu Xingらは、行列学習と増分学習の利点を組み合せLocal Distribution Self-Organizing Incremental NeuralNetwork(以下LD-SOINN)参考文献2)は、ILDNを拡張したものであり、Hebb則に基づくデータ空間のトポロジーと、それを利用したクラスタリング戦略を導入している。しかしながら、LD-SOINNは距離計算に、マハラノビスを用いているため、疎なデータに対して、従来のSOINNよりも性能が劣ってしまう。そこで、疎なデータに対しての距離計算を従来のSOINNで使用していたものを用いることにより性能の向上を図る。

2. 従来手法

2-1.SOINN

SOINNとは、1 層構造をなし、2 つのパラメータを有する。SOINNは、各入力に対して、「ノードの挿入」、「エッジの挿入・削除」、「ノードの位置ベクトルの更新」、「ノードの削除」という処理を適宜組み合わせて実行し、適切な分布の近似を行なう。以下にSOINNのステップを示す。

Step1　初回の学習である場合学習データからランダムに2つのノードを挿入を行う
Step2　学習データから入力されたパターンに対する勝者ノードを第二勝者まで探索し勝者ノードに対して閾値計算を行う
Step3　閾値の範囲外の場合新規ノードとして挿入し、範囲内なら勝者ノード間にエッジを引くまた引いていた場合はエッジの年齢を初期化する。
Step4　第一勝者につながる全エッジの年齢＋1をし、勝者ノード位置ベクトルを更新する
Step5　エッジの年齢が一定以上のものを削除し削除されたノードの中で連結ノードがないものを削除する。またλ周期毎に連結ノードが一つ以下のノードを削除する。

ノードの挿入は、分布の近似が十分でない領域に対して実行される。分布の近似が十分であるかどうかの判定は、既存のネットワーク構造から類似度閾値T(1)(2)参照を算出し、これに基づいて行なわれる。

$$\mathrm{N} \neq \phi \quad \mathrm{i} \qquad \mathrm{Ti} \leftarrow \frac{max}{c \in Ni} \ ||\mathrm{Wi} - \mathrm{Wc}|| \tag{1}$$

$$\mathrm{N} = \phi \quad \mathrm{i} \qquad \mathrm{Ti} \leftarrow \frac{min}{c \in A \backslash \{i\}} \min \ ||\mathrm{Wi} - \mathrm{Wc}|| \tag{2}$$

また、第一勝者、第二勝者いずれの閾値を超える場合ノードの分布が、不十分として、新規ノードを挿入する。超えない場合は、第一勝者と第二勝者間のノード間の重みを、状態を更新しノード間のエッジを０にする。そして、第一勝者につながる全エッジの年齢を＋１する。この時に勝者ノード間に、エッジが存在しない場合エッジを挿入し年齢を更新する。そして、以下の式に従い勝者ノードとその近傍ノードの重み更新を行う。

$$\Delta W_{s1} \leftarrow \frac{1}{t_i}(\xi - \Delta W_{s1}) \tag{3}$$

$$\Delta W_i \leftarrow \frac{1}{100 t_i}(\xi - \Delta W_i) \tag{4}$$

ただしt_iは、ノードが勝者ノードに選択された回数。これらを入力のたび繰り返しエッジの年齢が事前パラメータのage_{max}を超えた際に、エッジを削除し、エッジに連結されているノードでエッジを持たなくなったノー

The Characteristics of Trial Production Equipment

Masaharu ISHIKAWA and Yukari YAMAUCHI

ドを削除する。事前パラメーターのλ毎に連結ノードが１つ以下の全てのノードを削除する。この時にノイズが削除される。これによりノイズ耐性が得られる。

2-2.LD-SOINN

ここから着想を得て、作られたLD-SOINNは、「ノードの挿入」、「エッジの挿入・削除」、「ノードの位置ベクトルの更新」、「ノードの削除」に加えて「ノードのマージ」という処理を行う。このマージの処理は勝者ノードとその近傍ノードが結合の以下条件を満たすことによってStep5で行われます。

i)二つのノードがエッジにより接続されていること

ii)結合されたノードの体積が二つのノードの体積より小さいすなわち

$$V\,olume(m) < V\,olume(i) + V\,olume(j) \quad (5)$$

さらに初期のSOINNと違い距離関数にマハラノビス距離を用いている。マハラノビス距離は以下の式で求めることができる。

$$Di(x) = \sqrt{(x - ci)\,TM - 1\,i\,(x - ci)} \quad (6)$$
$$i = 1, 2, \ldots, |N|$$

マハラノビス距離とは、データの各方向への散らばり具合まで考慮した「データ群からの距離」を考えることができる。それによりデータの異常性を感知できるようになりよりノイズに対しての耐性が上がる。

従来のSOINNとは違い主成分が似ているノードをマージするためデータ表現が緩和される。

以上のメリットに加えLD-SOINNは基本的にほかの従来のSOINNよりも良い正規化相互情報量 (以下NMI)の結果を得ている。

3. 提案手法

第2節で説明したLD-SOINNは距離計算にマハラノビス距離を用いている。このためマハラノビス距離の特性上、疎なデータ(充分な)データでないとき性能を損なってしまう。この場合に対して距離関数を、従来のSOINNで使用されている距離関数を用いて性能の損失の回避が可能であるかを検討する。

4. 実験方法および測定方法

ISOLET、Iris、Glass、Letter、Segment、Soybean、Wine、COIL、USPS、YaleBこれらのデータセットを用いて従来手法、提案手法に従って学習、クラスタリングさせNMIの数値を比較することで性能の向上を期待する。またこの時にパラメータσの値に注意をしたい。

5. まとめ

LD-SOINNがスパースデータに対してクラスタリングの性能が損なってしまう問題に対して局所分布学習に基づくSOINNのスパースデータクラスタリングを提案し性能を損なわないように改善を試みた。

参考文献

1) 山崎和博, 巻 渕 有 哉, 申 富 饒, 長 谷 川 修「自己増殖型ニューラルネットワークSOINN とその実践」日本神経
回路学会誌（2010）Vol. 17, No. 4 187–196
(2) Youlu Xinga, Xiaofeng Shia , Furao Shena,, Ke Zhouc , Jinxi Zhaoa “A Self-Organizing Incremental Neural Network Based on Local Distribution Learning”(2016).