G-012

自己組織化ニューラル木立を用いた追加学習に関する研究

Self-Organizing Neural Grove and Its Incremental Learning Performance

梅本 雄大† 井上 浩孝‡
Yudai Umemoto Hirotaka Inoue

1．緒　　言

近年，パターン認識の技術は文字認識・画像認識など様々な分野で使われている．パターン認識における代表的な手法の一つにニューラルネットワークがあげられる．自己生成ニューラルネットワーク(SGNN)は，競合学習により自己生成ニューラル木(SGNT)を自動的に生成するため，高速な学習特性を有し，ネットワーク設計が容易である[1]．ニューラルネットワークを用いたパターン認識の方法も多種多様であるが，本研究では T.Kohonen によって提案された自己組織化マップ（SOM）[2]を応用し，SGNN を複数用いてアンサンブル学習によって認識精度を高めた自己組織化ニューラル木立(SONG)[3] を扱うことにする．SGNN は与えられた訓練データセットより，自動的に SGNT を構築することで，訓練データの特徴空間を木構造内に写像し，高速な学習特性を持つ識別器である．本研究では，SONG を用いた追加学習によるパターン認識の認識率の向上について検討する．

2．自己組織化ニューラル木立

SGNT は根，節点，葉からなる木構造のモデルである．根は SGNT 内に 1 つだけ存在し，入力データが提示される部分である．葉は未学習データに対する出力候補となる部分であり，競合学習により選択される．根と葉を結ぶ部分が節点であり，SGNT の生成過程で自動的に生成される．生成過程において競合学習により p 次元訓練データ e_{jk} に対する SGNT 内の勝者ニューロン n_{win} を決定する．根から n_{win} に至る節点に存在するニューロン n_j の重み w_{jk} は次式を用いて更新する．

$$\omega_{jk} \leftarrow \omega_{jk} + \frac{1}{c_j} \cdot (e_{jk} - \omega_{jk}), \quad 1 \le k \le m. \quad (1)$$

ここで，k=1，2，...，p であり，c_j は n_j に含まれる葉の数である．テスト過程において，未学習テストデータセットを SGNT の根に入力する．訓練と同様に根より各階層ごとに再帰的に競合学習を行い，到達した葉の持つ出力を SGNT の出力とする．ここで，訓練時は子とその親のニューロンを候補として根から葉へ再帰的に競合学習を行うのに対して，テスト時は SGNT の出力が存在する葉まで到達させるため，親を除いた子のニューロンを候補として競合学習を行う．図 1 に例として，1 次元データを e_{01}=1，e_{11}=2，e_{21}=3，e_{31}=4 の順番で学習した SGNT を示す．図 1 において，丸がニューロンを表し，丸の中の数はニューロン番号を表し，ニューロンの左下の数は重みを表し，ニューロンの右下の数はその木に含まれる葉の数を表している．

擬似 C 言語による SGNT 生成アルゴリズムは図 2 のように表される．

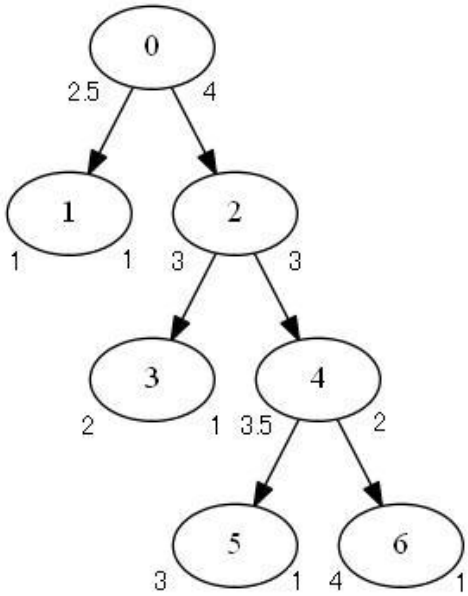

図 1 SGNT の生成例

```
Input:
  A set of training examples E = {e_i},
 i = 1, ... , N.
  A distance measure d(e_i,w_j).
Program Code:
  copy(n_1,e_1);
  for (i = 2, j = 2; i <= N; i++) {
    n_win = choose(e_i, n_1);
    if (leaf(n_win)) {
      copy(n_j, w_win);
      connect(n_j, n_win);
      j++;
    }
    copy(n_j, e_i);
    connect(n_j, n_win);
    j++;
    prune(n_win);
  }
Output:
  Constructed SGNT by E.
```

図 2 SGNT 生成アルゴリズム

† 呉工業高等専門学校 専攻科 機械電気工学専攻
‡ 呉工業高等専門学校 電気情報工学科

3. 実験結果

今回の実験では UCI 機械学習リポジトリ [4]の iris データ（データ数 150，4 次元，3 クラス）と letter データ（データ数 20000,16 次元，26 クラス）を用い，SGNT を生成し認識率を測定する．データを非復元抽出を用いてランダムにファイルを 10 分割し，そのうちの 9 個のファイルは SGNT の生成，残りの 1 個のファイルはテストデータとし，認識率を測定するのに用いた．9 個のファイルのうちまず 1 個のファイルから復元抽出を用いてランダムにデータを取り出し，SGNT を生成した．生成した SGNT にテストデータを入力し認識率を測定した．さらに木の数を 1，3，5，7，…，25 個と増やしていき，それぞれの認識率の平均を測定した．次に，生成した SGNT に 1 個のファイルを追加学習させ，SGNT を生成し，同様に認識率を測定した．この作業を 9 個のファイルで SGNT を生成するまで繰り返した．

図 3 に iris データによる実験のデータ数と木の数の変化による認識率を示す．SGNT を生成するデータ数を増やすほど，また，生成した木の数が多いほど認識率は高くなっていることが分かる．また，木の数を 25 個以上に増やしても認識率に大きな変化はなかった．以下がその理由と考えられる．平均 2 乗誤差はバイアス誤差と分散の和で表される．アンサンブル学習法により，木の数，データ数を増やした場合バイアス誤差に大きな変化はなく，分散は木の数，データ数を増やしていくと次第に減少し，限りなく 0 に近づくことが証明されている．従って木の数が 25 個付近で分散の値がほぼ 0 となり，認識率に大きな変化が見られなくなったと考えられる．また，今回用いた iris データは平均認識率は高いが，データ数が少なく，テストデータも 15 個のみである．従って，1 回認識を誤ると約 6.667%の誤差が生じてしまうことになり，中にはデータ数や木の数に比例しない結果もあった．

図 4 に letter データによる実験のデータ数と木の数の変化による認識率を示す．データ数，木の数が増加するにつれ認識率が高くなっていることが分かる．また iris データと比べると認識率がデータ数，木の数にほぼ比例しており，非常に安定した変化をしていることが読み取れる．データ数が多いことで 1 回の認識ミスによる認識率の低下の影響が少なくなり，追加学習による認識率の変化を高い精度で測定することができたと考えられる．しかし letter データのような膨大な数のデータは計算コストが大きくなってしまうので枝刈り法を用い，効率の良いパターン認識を行う必要がある．

4. 結　　言

本研究では，iris データ及び letter データを用いて追加学習によるパターン認識を行い SGNT の認識率の変化を検討した．実験結果より SGNT を生成するデータ数を増やすほど，また，生成した木の数が多いほど平均認識率は高くなっていることが確認できた．また，データ数，クラス数が増大することでの追加学習による認識率の安定化を確認できた．

今後は letter データを用いクラスごとに木を生成し，追加学習した場合の認識率の変化を測定する予定である．

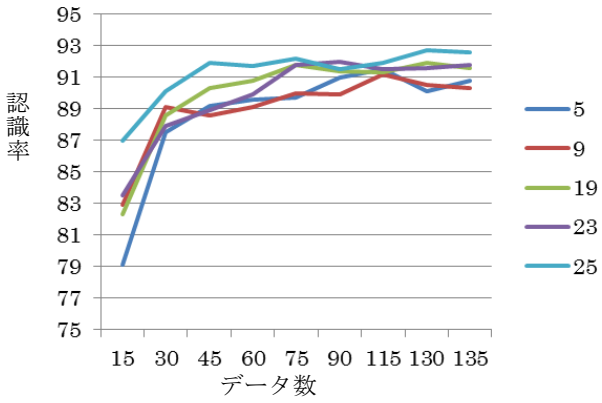

図 3 iris データの認識率

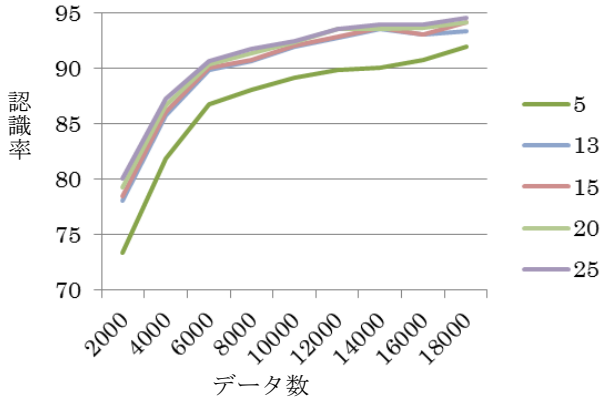

図 4 letter データの認識率

参考文献

1)W.X.Wen，A.jennings，H.Liu Learning a Neural Tree，Proc.International Joint Conf on Neural Networlks，pp.751-756，1992.
2)大北 正昭，徳高 平蔵，藤村 喜久郎，権田 英功. 自己組織化マップとそのツール，Springer，2008.
3)H. Inoue. Self-Organizing Neural Grove: Efficient Multiple Classifier System with Pruned Self-Gernerating Neural Networks，pp.281-291, Springer, 2009.
4)A.Frank，A.Asuncion. UCI machine learning repository，2010. http://archive.ics.uci.edu/ml