

多クラスデータ分類のためのデータ選択と学習の加速法

原 一之 (東京都立工業高等専門学校)、
中山 謙二 (金沢大学)

第12回 回路とシステム(軽井沢)ワークショップ
1999年4月19日(月)・20日(火)
於 軽井沢プリンスホテル
西館 国際会議場

主催 電子情報通信学会 システムと信号処理サブサイエティ
回路とシステム研究専門委員会
VLSI 設計技術研究専門委員会
デジタル信号処理研究専門委員会
コンカレント工学研究専門委員会
非線形問題研究専門委員会

協賛 IEEE Circuits and Systems Society, Tokyo Chapter
IEEE Signal Processing Society, Tokyo Chapter
電気学会 電子回路研究専門委員会

多クラスデータ分類のためのデータの選択と学習の加速法

A Training Acceleration and Data Selection Method for Multi-Class Data Classification by Multilayer Neural Networks

原 一之

Kazuyuki Hara

東京都立工業高等専門学校 電子情報工学科
Tokyo Metropolitan College of Technology

中山 謙二

Kenji Nakayama

金沢大学 工学部 電気・情報工学科
Faculty of Engineering, Kanazawa University

E-mail: hara@tokyo-tmct.ac.jp, nakayama@t.kanazawa-u.ac.jp

1 はじめに

階層形ニューラルネットワーク(以下、階層形 NN)を分類問題に適用した場合、データ間の規則や特徴を学習過程で自動的に抽出できるため、分類のためのアルゴリズムが明確でない問題に多く適用されている。一方、階層形 NN の実用化に当たっては学習に用いなかったデータ、つまり未学習データに対しても高い分類能力を保證する学習法が必要があるが、そのような学習法はいまだ確立されていない。したがって、未学習データに対する分類能力(以下、分類能力)を高めるために多数の学習データを用いるが、データ数が多くなると学習が難しくなるとともに計算量が增大するため、実用化が難しい。したがって、少ないデータ数で高い分類能力を保證する学習法の開発が要請されている。

データ選択に関しては、D. Mackay[1]が Bayesian Learning の枠組みにおいて active data selection という方法を提案している。統計的なデータ選択法としては Cachin がネットワークの出力誤差に基づいた確率的なデータ提示法を提案している。また、Géczy は多数の学習データを用いた学習の高速化と Jacobean 行列の縮退を防ぐためのデータ選択法[3]を提案している。これらに対し、本文では少ないデータ数で分類能力を高めるためのデータの選択法とそのデータを用いた学習法について議論する。

分類問題ではクラスに属するデータは入力空間内で領域を形成し、特にクラス間でオーバーラップがない場合、クラスの領域はクラス間のデータを分離する境界、つまりクラス境界によって分離される。一方、階層形 NN は、入力空間を結合係数で形成される超平面によってクラス領域に適応した領域に分割することにより分類問題を解く。した

がって、階層形 NN の形成する境界(以下、ネットワーク境界)をクラスデータの分布に適応させるために必要不可欠なデータは、クラス間の最近傍データである。著者らはこのような発想に基づき、クラス間の最近傍データの選択法および選択データによる学習により高い確率で高い分類能力を保證できる学習法を提案した[4, 5, 6]。提案法は2クラスデータからクラス間の最近傍データ(以下境界データ)を抽出するためのデータ選択法および選択データを用いたオフライン、またはオンラインの学習法からなる。分類能力は全データを用いた学習法(以下、従来法)と同等であり、学習時の計算量は従来法の約10%である。

本文ではまず、データ領域に重なりがないこと、オフライン学習であることを前提に、全出の提案法を3クラス以上に適用できるよう拡張した境界データの選択法を提案する。つぎに、選択したデータを用いた学習法を提案する。提案法は前出の学習法より学習時の計算量を削減できる。提案法の有効性は、計算機シミュレーションにより確認した。

2 階層形ニューラルネットワーク

本文で扱うニューラルネットワークは入力層、一層の隠れ層、出力層からなる2層の階層形 NN である。学習アルゴリズムは誤差逆伝播法を用い、オフラインで学習を行う。

入力層の各ユニットに与えられた学習データ $\mathbf{x} = \{x_1, \dots, x_N\}$ は、入力層・隠れ層間の結合係数 w_{nj} で重み付けされ、 j 番目の隠れユニットの入力ポテンシャル net_j となる。

$$net_j = \sum_{n=0}^N w_{jn} x_n \quad (1)$$

ただし、 $w_{j0}, n=0$ はバイアスである。また、出力 y_j はシグモイド関数 $f_H(\cdot)$ により、

$$y_j = f_H(net_j) = \frac{1 - \exp(-net_j)}{1 + \exp(-net_j)} \quad (2)$$

と計算される。 o 番目の出力ユニットの入力ポテンシャル net_o および出力ユニット出力 y_o は隠れユニットの場合と同様に計算される。ただし、出力ユニットのシグモイド関数 $f_o(\cdot)$ は、

$$y_o = f_o(net_o) = \frac{1}{1 + \exp(-net_o)} \quad (3)$$

を用いる。学習が終了した場合、例えば入力データのクラスに対応する出力ユニットの出力値が最大となる。

3 境界データの選択法

m 個の N 次元ベクトル x からなるデータ集合 $X = \{x_1, \dots, x_m\}$ を考える。このデータが C 個のクラスにより構成されているとき、各クラスのデータ集合 $\{X_c, c = 1, \dots, C, X_c \in X\}$ は入力空間内でクラスごとに領域を形成する。集合 X の中から任意の2つのクラスのデータ集合 $\{X_{c1}, X_{c2}\}$ を取ったとき、これらの領域がオーバーラップしていないと仮定する。集合 X_{c1}, X_{c2} の任意のデータ x, y に対してある距離関数 $d(x, y)$ (たとえばユークリッド距離) を導入すると、クラスの境界 (以下クラス境界) $B(X_{c1}, X_{c2})$ はクラスの異なるデータ間の距離 $d(x_{c1}, x_{c2})$ がもっとも小さくなるデータの集合 X_{c1}^B と X_{c2}^B の間に引いた曲線となる。言い換えれば、 $B(X_{c1}, X_{c2})$ を挟むようにして境界データ X_{c1}^B, X_{c2}^B が存在する。

以下ではクラス境界付近のデータ (以下、境界データ) の選択法を提案する。本方法は、(1) データの分布が分かるような十分なデータがある、(2) クラス境界が境界データを滑らかに結ぶ曲線である (3) クラスデータの領域がオーバーラップしない、という3つの仮定が成り立つとき正しく境界を抽出することができる。

3.1 境界データの選択アルゴリズム

境界データを選択するアルゴリズムを次に示す。本文では、このアルゴリズムをダブルペアリング法と呼ぶ。

Step 1: 各クラスからランダムに N 個のデータを選択する。選択したデータを X^S とする。

Step 2: 全クラス C から1つのクラス $c \in C$ を選択する。

Step 3: X_c^S の中からランダムに1個のデータ x_c^S を選択する。

Step 4: x_c^S に対して最も距離の近い他クラスのデータ $x_d^p \in X_c^S$ を選択する。 X_c^S は選択したクラス c 以外のクラスの全データを表わす。

Step 5: x_d^p に最も近いデータ $x_c^{Sp'}$ を X_c^S から選ぶ。このとき、 x_d^p と $x_c^{Sp'}$ が境界付近の最近傍データのペアとなる。

Step 6: Step 3 で X_c^S のすべてのデータが選択されるまでステップ2から4を繰り返す。

Step 7: Step 2 ですべてのクラスが選ばれるまでStep 2 から6を繰り返す。

なお、提案法では同じデータを別のペアとして選択できる。

3.2 従来法との比較

従来のペアリング法 [5] では、Step 5 において、 x_d^p に最も近いデータをクラス d 以外のすべてのクラスから求めていた。この場合、図1のようなデータが与えられた場合、 x_{31} は選択されない。したがって、クラス3の境界データは選択されない。これに対し、提案法ではStep 3 で、クラス c から最近傍データを選択するため、 x_{31} を選択できる。

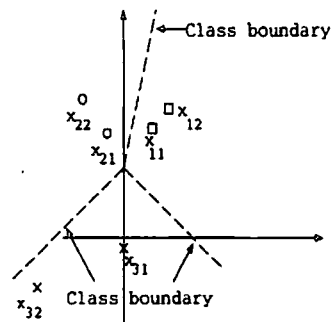


図 1: Data distribution of boundary data

4 境界データによる学習

学習後のネットワークは入力データのクラスに対応する1つの出力ユニットのみが最大値を出力するようになる。このネットワークにおいて、入力空間内から任意にサンプルした点に対して2個以上の出力ユニットが等しい最大値をとる場合、この点はネットワークの構成するクラス領域の境界上にある。本論文ではこのような点の集合からなる曲線をネットワーク境界と呼ぶ。

ネットワーク境界がクラス境界と同程度の滑らかさになるように隠れユニット数が決定されていると仮定したとき、境界データ X^B でネットワークを学習し、 X^B を正しく分類できた場合、境界データとして選択されなかったデータを高い確率で分類できる。これは選択されなかったデータとネットワーク境界との距離が選択されたデータとのよりも距離が大きい、つまり領域の中に位置するためである。

しかし、境界データを用いてネットワークを学習する場合、学習の初期においては境界データの組 $\{x_{c1l}^B, x_{c2l}^B\}, l = 1, \dots, L$ に対する結合係数の修正量が相殺されるため、学習が停滞する。この現象は特に学習初期の結合係数が小さな乱数(例えば ± 0.2)で決定されている場合に顕著である。この現象は境界データの組に対する距離が近かつクラスが異なるために起こる。つまり境界データの組に対する結合係数の修正量は絶対値がほぼ等しく極性が異なるためである [5]。そこで、本論文では境界データの組にランダムに選択したデータ(以下、ランダムデータ)を加え、修正量をアンバランスにすることにより学習を加速する学習法を提案する。

4.1 学習アルゴリズム

以下に学習アルゴリズムを示す。提案する学習法ではダブルペアリング法によって選択したデータを学習データとして用いる。

Step 1: 全データを用いてダブルペアリング法により境界データを選択する。

Step 2: 全データからランダムに少数のデータを選択する。

Step 3: Step 1, Step 2 で選択したデータを学習データとして学習を行う。ネットワークの出力の2乗平均誤差が ϵ 以下になったら学習を終了する。

なお、ランダムデータの数は4.2で検討する。

4.2 ランダムデータの学習加速効果の解析

本節ではランダムデータの学習加速効果を解析する。簡単のため、クラス数は2、境界データはクラス1と2から1個ずつ選択した1次元データ x_1^B, x_2^B とする。図2にデータの分布を示す。

ランダムデータの与え方として3つの場合を想定した。(a) ネットワークが正しく分類できないクラス1のデータ x_{11}^R を追加した場合、(b) ネットワークが正しく分類できるクラス1のデータ x_{12}^R を追加した場合、(c) ネットワークが正しく分類できるクラス2のデータ x_{21}^R を追加した場合である。以下、(a) から (c) の学習加速効果を定性的に解析する。

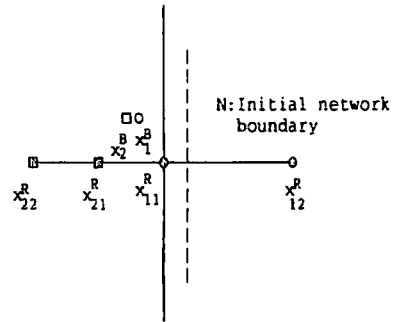


図2: Random data distribution

(a) の場合、ネットワークは追加したデータ x_{11}^R を正しく分類できないため、誤差が発生する。したがって、この3個のデータに対する結合係数の修正量は相殺されず、ネットワーク境界は境界データの方へ移動するように結合係数が修正される。(b) の場合、ネットワークは x_{12}^R を正しく分類できるが、初期のネットワークの結合係数が比較的小さな乱数で決定されている場合、出力ユニットの入力ポテンシャルはシグモイド関数の線形領域にあるため誤差が発生する。したがって、この場合にも3個のデータに対する結合係数の修正量は相殺されない。(c) の場合、ネットワーク境界を境界データに近づけようとする x_{21}^R に対する誤差が大きくなるため、そのような学習は行われぬ。しかし、シグモイド関数の傾きを大きくすることにより x_{21}^R に対する誤差を減らすことが出来る。したがって、 x_{21}^R を加えることによりシグモイド関数の傾きを大きくするような学習が行われるため、修正量が相殺されるのを防ぐことが出来る。

4.3 計算機シミュレーション

以上の解析を計算機シミュレーションで検証した。境界データを $x_1^R = (-0.45, 0.5)$, $x_2^R = (-0.5, 0.5)$ とする。これに対し、ランダムに選択したデータとして境界に近いデータとして $x_{11}^R = (0.0, 0.0)$, $x_{21}^R = (-1.0, 0.0)$, 境界から遠いデータとして $x_{12}^R = (5.0, 0.0)$, $x_{22}^R = (-5.0, 0.0)$ を使用する。初期のネットワークは ± 0.2 の乱数で決定され、 $(0.8, 0)$, $(1.0, 0.6)$ を通る直線である。データの分布は図2と同じである。

シミュレーションは (1) 境界データのみ (2) 境界データと x_{11}^R (3) 境界データと x_{21}^R (4) 境界データと x_{12}^R (5) 境界データと x_{22}^R (6) 境界データと x_{11}^R, x_{21}^R (7) 境界データと x_{12}^R, x_{22}^R の7通りのデータの組み合わせについて行った。その結果、学習の収束回数は (1) 9807回、(2) 3409回、(3) 3269回、(4) 3845回、(5) 3600回、(6) 2878回、(7) 3567回となった。

結果より、ランダムに選択したデータを加えることによって、結合係数の修正量が行われ、学習が加速されることがわかる。(6)、(7)の場合、(4)、(5)に対して学習が加速されているが、これは(2)と(3)の加速効果が協調しているためであると考えられる。さらに、初期の結合係数小さな乱数で決定されている場合、(4)、(5)、(7)のように境界から遠いデータを用いても学習を加速する効果がある。これは初期の結合係数が小さい場合、境界データから遠いデータに対しても入力ポテンシャルをシグモイド関数の線形部分に持ってくるためである。

この解析から境界データの組に対し、1個のランダムデータを加えることによって学習が加速できることがわかる。したがって、学習を加速するために必要なランダムに選択したデータ数 N_R は $N_R = (N_P/N_C)/2$ で表わされる。ここで N_P は境界データ、 N_C はクラス数である。

5 計算機シミュレーション

提案した境界データ選択法と学習法の有効性を検証するため、計算機シミュレーションを行った。問題は2次元データの3分類問題(以下、問題1)とIrisの分類問題(以下、問題2)である。また、シミュレーション結果を従来法、2段階学習法と比較し、その有効性を検証した。

5.1 問題1

図3(a)にクラスデータ分布の概念図を示す。クラス数は3、1クラス当たりのデータ数は1000である。クラス間には0.01のギャップがある。

(1) データ選択

図3(a)のデータにペアリング法を適用し、選択されたデータを図3(b)に示す。図より、境界が正しく抽出されていることがわかる。図では $N_P=315$ の境界データが選択された。

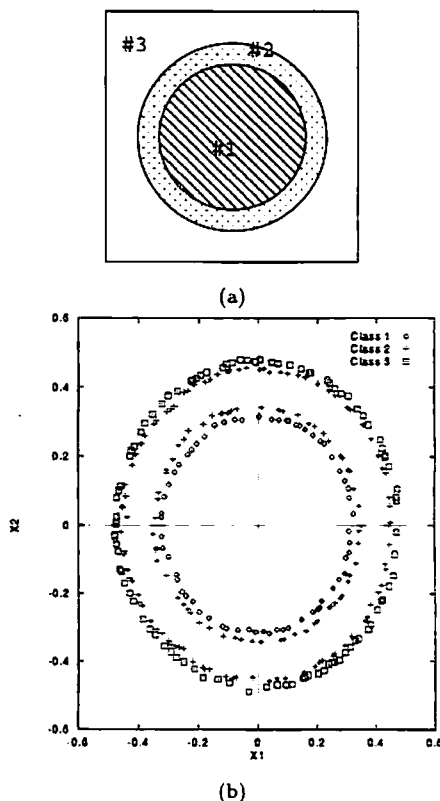


図3: (a)Concept of class data. (b)Boundary data selected by proposed method.

(2) 学習結果

学習データとして、境界データ数 $N_P=315$ および各クラスからランダムに選択したデータ数 $N_R = 150$ 、合計 465個のデータを用いた。ネットワーク構造は入力ユニット2、隠れユニット12、出力ユニット3とした。学習の収束条件はネットワーク出力の2乗平均誤差が0.01以下になったと

きとした。また、学習後のネットワークの性能は分類率によって判定した。ここで、分類率とは全データ数に対する正しく分類できたデータ数の比である。データの分類は、入力データのクラスに対応する出力ユニットの出力が最大となったとき、正しく分類できたとした。

計算機シミュレーションの結果、学習は 15097 回で収束し、分類率は 100%であった。

(3) 従来法と提案法の比較

すべてのデータを学習データとして用いる学習法を従来法とする。したがって、従来法では 3000 個の学習データを用いる。ネットワーク構造は提案法と同様に入力ユニット 2、隠れユニット 12、出力ユニット 3 とした。計算機シミュレーションの結果、従来法では学習が収束したときの分類率が 100%とならなかった。したがって、提案法の分類率が従来法が収束したときの分類率と同じになったときの計算量で比較した。1000 回の学習回数で両方法ともに 98.4%の分類率となった。したがって、計算量の比は (提案法/従来法) = $(465/3000) \approx 0.16$ となった。以上より、提案法は学習に要する計算量は大幅に削減できることがわかった。

(4) 2 段階学習法との比較

提案法は境界データとランダムデータを用いる学習法である。ランダムデータを加えることによる効果を以下で述べる 2 段階学習法と比較し、提案法の学習高速化について解析する。

2 段階学習法は、ランダムデータでネットワークの出力誤差が ϵ 以下になったところで学習を中断し、それ以降の学習では境界データのみを用いる学習法である。この学習法では、提案法と同様に境界データとランダムデータを用いるが、同時に用いてない点で提案法と異なる。この両者を比較することにより、ランダムデータの効果を確認できる。

学習には図 3 を用いた。第 1 段階の学習は各クラス 50 個のデータを、第 2 段階の学習では前節 (1) と同数の境界データを用いた。第 1 段階のデータ多数は、提案法で用いるランダムデータの 2 倍のデータを用いている。第 1 段階の学習はネットワークの出力誤差が 0.05 に達したとき中断し、分類率が 98.4%に達したとき、学習を終了した。学習の結果、第 1 段階の学習は 1875 回、第 2 段階の学習は 16690 回で収束した。したがって、この結果より提案法より多いランダムデータを用いたにもかかわらず、第 1 段階の学習で形成したネットワーク境界は、クラス境界の近傍には配置されなかったことが分かる。したがって、提案法はネットワーク境界は境界データに対して学習を行うことにより正しく

形成し、ランダムデータにより、その学習が加速されるという相乗効果があることがわかった。

5.2 問題 2

問題 2 は 3 種類の Iris を花卉、がくの幅、長さによって分類する問題である [7]。1 つのクラスは他の 2 クラスに対して線形分離可能であるが他の 2 クラスは線形分離できない。データ数は各クラス 50 個である。がくの幅、長さによる各クラスデータの分布を図 4 に示す。図より、クラス 2 と 3 のデータがオーバーラップしていることがわかる。

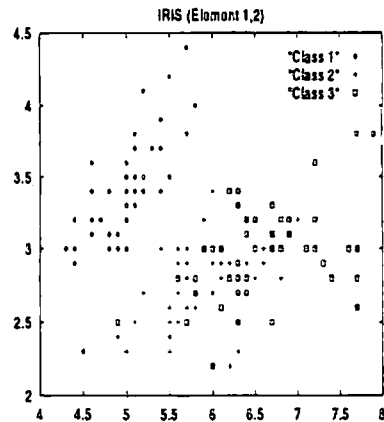


図 4: Data distribution of first and second element of iris data-base

(1) 学習結果

ダブルペアリング法で求めた境界データ 17 個、ランダムデータ 3 個の合計 20 個のデータで学習を行った。分類率は全データに対して計算した。ネットワーク構造は入力ユニット 4 個、隠れユニット 5 個、出力ユニット 3 個とした。クラス 2 と 3 のデータがオーバーラップしているため、分類率は 100%とはならない。そこで、分類率が 98%以上となったとき学習を終了した。その結果、学習は 46 回で収束した。また、分類率は 98.7%であった。

(2) 従来法と提案法の比較

問題 1 と同様に、従来法との比較を行った。従来法は問題 1 と同様である。学習の終了基準は提案法と同様とした。その結果、従来法は 94 回で学習が収束し、学習後の分類率は提案法と同じ 98.7%であった。したがって、計算量の比は (提案法/従来法) = $(46 \times 17)/(94 \times 150) \approx 0.06$ とな

り、提案法の計算量が大幅に小さくなった。

6 まとめ

本文ではオフライン学習で、データがクラス間でオーバーラップしないことを前提に、少ない学習データで高い分類能力を保證できるデータの選択法およびそのデータを用いた学習法を提案した。提案法は境界データとランダムデータを用いて学習を行うことにより、同じ学習回数で全データを用いた場合と同等の分類能力を獲得することができた。

参考文献

- [1] D. J. C. Mackay, *Information-based objective function for active data selection* Neural Computation, MIT vol. 4, No. 4, pp.589-603, 1992.
- [2] C. Cachin, *Pedagogical pattern selection strategies* Neural Networks, vol. 7, no. 1, pp. 175-181, 1994.
- [3] P. Géczy, S. Usui, *Dynamic Sample Selection: Theory* IEICE Trans., Fundamentals, vol. E81-A, no.9, pp.1931-1939 Sept. 1998.
- [4] K. Hara and K. Nakayama, *Selection of Minimum Training Data for Generalization and On-line Training by Multilayer Neural Networks*. Proc. ICNN'96, Washington D.C., 436-441, July 1996.
- [5] K. Hara, K. Nakayama, *Data Selection Method for Generalization of Multilayer Neural Networks*, IEICE Trans., Fundamentals, vol. E81-A, no.3, pp.371-381 March 1998.
- [6] K. Hara, K. Nakayama, *A training data selection in on-line training for multilayer neural networks*, Proc. IJCNN'98, pp. 2247-2252, Anchorage USA, May 1998.
- [7] Fisher R. A., *The use of multiple measurements in taxonomic problems*, Annual Eugenics, vol. 7, Part II, pp.179-188, 1936.