F-042

計算負荷を軽減させた新しいニューラルネットワークの開発
Development of Neural Network that Reduces Calculation load

田浦 直樹† 堂園 浩‡
Naokj Taura Hiroshi Dozono

1. はじめに

近年，AI 技術の発展により，幅広い分野で NN が利用され，活躍している．しかし，モデルの規模が大きくなると，計算量や処理に要する時間の増加，エネルギー消費が高くなるといった課題がある．本研究では，計算負荷を軽減させた新しい NN の開発を試みる．コンピュータ内での実数演算は整数演算より計算負荷が高いため，NN での計算整数のみで行う必要がある．

先行研究では，実数演算ではなく整数演算のみを用いた推論に焦点を当てて研究が行われている．実数演算と比較すると，整数演算を用いた NN は推論時間が最大 50%短縮されたことが確認されている．[1]

2. 実験方法

本研究では，重み値を 0,±2^n (-7<n<7)の整数に限定した NN を提案する．重み値は，通常，浮動小数点数で表される．浮動小数点数は，図 1 のように，32bit(単精度)，または，64bit(倍精度)で表されることが多く，整数より多くのビット数を使う．浮動小数点数の演算では，仮数部の正規化，指数部の加算や減算，といった計算行程が整数演算よりも増加する．計算整数であっても乗算は加減算と比較すると多くの計算量を要する．しかし，2^n の計算はシフト演算で簡単に計算することが可能である．

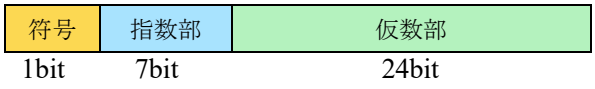

図 1　浮動小数点数の構成図

また，活性化関数に ReLU6 関数を用いる．これにより，ニューロンの値も整数で出力することが可能である．活性化関数が 1 つでは学習できない可能性があるため，複数の活性化関数の使用を検討する．学習アルゴリズムには，最適化手法の遺伝的アルゴリズム(GA)を用いる．重み値を 0,±2^n (-7<n<7)に限定しているため，整数値 n を離散的に調整する学習アルゴリズムでは学習できないと考えられる．また，活性化関数を複数使用する可能性があるため，最適化も可能である GA が有効である．

実験は，UCI Machine Learning Repository を用いて行った．[2]　これは，機械学習の研究において広く利用されているデータセットが集められており，UCI から提供されるデータセットを実験で使用する．

† 佐賀大学　Saga University
‡ 佐賀大学　Saga University

3. 実験結果と検討

作成した NN の正解率と学習時間 (最適解を導出するのに要した時間)，推論時間 (最適解を用いて未知のデータで推論するのに要した時間) の測定を行い，他の代表的な機械学習アルゴリズムである，ランダムフォレストと SVM，作成した NN と同じ構造を持ち，実数を用いて計算をする，多層ニューラルネットワークとの比較を行う．表中の GA が今回設計を行った NN である．

入力層から隠れ層への出力では，正規化を行い，0~256 の範囲の整数で出力を行う．隠れ層から出力層への出力では正規化を行い，0,1 で出力を行う．また，One-Hot エンコーディングを行い，クラス分類を行う．

3.1 XOR

最初に簡単なモデルである XOR を用いて，設計した NN の動作確認を行った．XOR での実行結果を表 1 に示す．

表 1 より，設計した NN は正常に動作することが確認された．また，XOR のような単純なモデルの場合，高精度での推測が可能なことが示された．

表 1　XOR での実行結果

回数	1	2	3	4	5	6	7	8	9	10
正解率	1	1	1	1	1	1	1	1	1	1

3.2 Iris

Iris というデータセットで実験を行った．この実験結果と他の機械学習アルゴリズムとの比較を表 2 に示す．

Iris データはアヤメ(Iris)という花の 4 つの特徴を記録したものであり，花の種類は 3 種類ある．4 つの特徴量から，どの種類のアヤメになるか推測する．データ量は 150 個あり，8 割を学習用データ，2 割をテスト用データとして実験を行う．

表 2 より，正解率はランダムフォレストや SVM より低いが，多層ニューラルネットワークより高く実用的な値が得られた．また学習時間は最も長いが，推論時間は最も短い値が得られた．

表 2　Iris での実験結果・比較

	正解率	学習時間 [s]	推論時間 [ms]
GA	0.803	62.889	0.295
SVM	0.973	0.005	3.725
ランダムフォレスト	0.930	0.418	32.341
多層ニューラルネットワーク	0.656	1.187	7.886

3.3 Wine

次に Wine というデータセットで実験を行った．この実験結果と他の機械学習アルゴリズムとの比較を表 3 に示す．

このデータセットは，ワインのアルコールや色相などの 13 個の特徴量から 3 種類のワインの分類を行う．データ数は 178 個あり，8 割を学習用データ，2 割をテスト用データとして実験を行う．

表 3 より，正解率はランダムフォレストに次いで高く，多層ニューラルネットワークと近い値が得られた．学習時間は最も長くなっているが，推論時間は最も短い値が得られた．

表 3　Wine での実験結果・比較

	正解率	学習時間 [s]	推論時間 [ms]
GA	0.869	136.344	0.592
SVM	0.442	0.018	6.441
ランダムフォレスト	0.986	0.462	31.959
多層ニューラルネットワーク	0.857	0.823	8.127

3.4 Breast Cancer Wisconsin (Diagnostic)

Breast Cancer Wisconsin (Diagnostic)というデータセットで実験を行った．この実験結果と他の機械学習アルゴリズムとの比較を表 4 に示す．

Breast Cancer Wisconsin (Diagnostic)は乳がん診断に関連するデータセットであり，細胞核の半径や周囲の長さなどの 30 個の特徴量から良性か悪性かの判別を行う．データ数は 569 個あり，8 割を学習用データ，2 割をテスト用データとして実験を行う．

表 4 より，正解率はランダムフォレストに次いで高く，多層ニューラルネットワークと近かい値が得られた．また，学習時間は最も長くなっているが，推論時間は最も短い値が得られた．

表 4　Breast Cancer Wisconsin (Diagnostic)の実験結果・比較

	正解率	学習時間 [s]	推論時間 [ms]
GA	0.921	552.589	0.774
SVM	0.636	0.021	10.161
ランダムフォレスト	0.962	0.230	10.955
多層ニューラルネットワーク	0.915	1.785	10.253

3.5 検討

推論時間はどの実験でも，最も短い値を得られたが，学習時間はどの実験でも，最も長くなった．これは，遺伝的アルゴリズムでは最適解を導出するのに時間を要すため，このような結果になったと考えられる．

どの実験でも，正解率は 0.8 を超えているため，実用的な値が得られたと考えられる．また，活性化関数が 1 つでも学習を行うことが可能なことが実証された．そのため，学習アルゴリズムに遺伝的アルゴリズム以外の学習アルゴリズムの使用が可能だと考えられる．今後は学習時間を短縮するため，遺伝的アルゴリズム以外の学習アルゴリズムの使用を検討する必要がある．また，多層ニューラルネットワークと比較して，高い値，もしくは近い値が得られたことから，重み値を 0,±2^n (-7<n<7)に限定した場合でも学習を行えることが実証された．

4. おわりに

本研究では，計算負荷を軽減させた NN の開発を目的に行った．その結果，重み値を 0,±2^n(-7<n<7)に限定し，整数のみで演算を行い，学習が可能であり，推論時間を短縮させられることが実証された．今後は，学習アルゴリズムに遺伝的アルゴリズム以外の使用を検討する必要があり，より複雑な問題である MNIST や CIFAR を学習できるアルゴリズムに改良していきたい．

謝辞

本研究を行うにあたって，常日頃から親切丁寧な御指導，御鞭撻を頂きました，堂園　浩　准教授，並びに諸先生方に深く感謝申し上げます．また，本研究において御助言，御協力を頂きました同研究室の大学院生の皆様に深く感謝申し上げます．

参考文献

[1] Jacob B., Kligys S., Chen B., Zhu M., Tang M., Howard A., Adam H., Kalenichenko D.AUTHOR Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference(2018) Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, art. no. 8578384, pp. 2704 - 2713,

[2] Home - UCI Machine Learning Repository