

ランダムフォレストにおけるノード数と木数の関係

On the trade-off between the number of nodes and the number of trees in Random Forest

熊野 颯^{*1}

So Kumano

阿久津 達也^{*2}

Tatsuya Akutsu

^{*1}京都大学大学院情報学研究科

Graduate School of Informatics, Kyoto University

^{*2}京都大学化学研究所

Institute for Chemical Research, Kyoto University

Expressibility of machine learning models has been extensively studied. For example, in a Neural Network, it is proved that the efficiency concerning the number of nodes is generated from the depth. On the other hand, it is not clear whether the efficiency exists in Random Forest. Therefore, in this research, we investigate whether the efficiency exists in Random Forest. We first show that Random Forest does not have the same kind of efficiency as Neural Network, and next we show that the efficiency concerning the number of nodes can be generated from the number of trees.

1. はじめに

封筒上の郵便番号の識別から店舗ごとの商品の需要予測、医療診断に到るまで幅広い分野において機械学習は活用されている。いずれの応用においても、まず学習のモデル (例えば、ニューラルネットワークなら層の数やニューロンの数、ランダムフォレストであれば木の木数や葉の数) を決定し、次に損失関数を最小化するパラメータを発見する。この学習モデルの決定、すなわち学習によって最適な関数の探索を行う関数族の決定は非常に重要である。関数族が不必要に大きい場合は過剰な計算コストや過学習の問題を引き起こし、逆に不十分な場合には、適切な関数を得ることはできない。課題ごとに適切な関数族を選択しなければならないのである。

こうした関数族の選択において有効な手がかりを得るべく、学習モデルの表現能力に関する研究が古くから行われてきた。例えば、深さ 2 のニューラルネットワークで任意のボレル可測関数を近似することができることを主張する universal approximation theorem が知られている。また、近年ではニューラルネットワークは層の数によって表現能力が大きく変化することが分かってきた。例えば、深さ 2 のニューラルネットワーク (SNN) は任意のボレル可測関数を近似することが可能であるが、必要となるノード数が膨大となることがある。しかし、深いネットワーク (DNN) を用いることで、SNN よりも少ないノード数で同じ関数を表現できる場合がある。こうした DNN が持つ、深さによって効率良く関数を表現することが可能となる性質は depth efficiency と呼ばれている。

ニューラルネットワークに対しては、表現能力の研究が盛んにおこなわれている一方で、ランダムフォレストにおいては表現能力に関する研究は少ない。そこで、本研究ではランダムフォレストがニューラルネットワークにおける depth efficiency と同様にノード数に関する効率性を持つかについて考察する。

2. 関連研究

ニューラルネットワークの表現能力は、その深さに対して指数的に増大すると考えられている。Montúfar らは区分線形関数を活性化関数として用いた際に、ネットワークが表現する区

分線形関数の線形領域がネットワークの幅に対しては多項式のオーダーでしか増加しないのに対し、深さに対しては指数的に増大することを示した [1]。また、Raghu らは、ネットワークの出力の軌道の長さの観点から、深さに対して表現能力が指数的に増大することを示した [2]。

DNN は SNN よりも効率的に関数を表すことが出来ることを実際に示した研究も存在する。Telgarsky は、SNN では指数オーダーのノードが必要だが、DNN では線形オーダーのノードで表現できる関数族の存在を示した [3]。Symanskiy らは、DNN において depth efficiency が生じる要因の一つとして入力の変動性に関与していることを明らかにした [4]。Bengio らは、Sum-product Network に関して DNN では SNN よりもノード数の観点で効率的となる関数が存在することを示した [5]。

ランダムフォレストの表現能力に関する研究においては、Mansour が入力の次元が d 、ノード数 N の二分決定木の VC 次元の下界が $\Omega(N)$ 、上界が $O(N \log d)$ であることを示した [6]。また、VC 次元が d の識別器を T 個使用するアンサンブル学習の VC 次元の上界は $O(dT \log(dT))$ である [7]。Oshiro らは、データの密度を定義し密度と最適な木の木数の関係を実験により確かめた [8]。これらの研究から、ランダムフォレストは木の木数に対して、表現能力が劇的に増加するといったことはないと思われる。しかし、これらの結果は depth efficiency と類似の性質の存在を否定する訳ではない。そこで本研究では、ランダムフォレストにおけるノード数に関する効率性について考察する。

3. ランダムフォレストの深さ

Bengio らは、ブースト木のニューラルネットワークにおける深さは 3 であると述べている。[9]。この結果から、ランダムフォレストは深さ 3 のニューラルネットワーク以下の表現能力しか持たず、depth efficiency と同種の効率性は存在しないと考えられる。

しかし、この結果は 選言標準形のアナロジーとして述べられたものである。一方、近年の DNN の depth efficiency に関する結果の多くは区分線形関数を活性化関数として用いたニューラルネットワークについて示されている。そこで、本研究ではまず、ランダムフォレストは区分線形関数を活性化関数として用いた深さ 3 のニューラルネットワーク以下の表現能力

連絡先: 熊野 颯, 京都大学大学院情報学研究科,
kumano@kuicr.kyoto-u.ac.jp

しか持たないことを示す。ただし、ニューラルネットワークの出力ノードは線形関数とする。以後、本研究ではランダムフォレストを構成する決定木として、各ノードごとに入力空間にある変数に対して垂直二分割（他の変数に対して軸並行に分割）する二分決定木を考え、各決定木の出力はラベルのみであるとする。つまり、ランダムフォレストが出力するラベルはそのラベルを出力する木の木数が最も多いラベルとなる。したがって2値分類問題におけるランダムフォレストはすべて奇数本の決定木から構成される。また、 c でクラス全体を表し、 H を

$$H(x) = \begin{cases} 1 & (x \geq 0) \\ 0 & (x < 0) \end{cases} \quad (1)$$

と定義する。また、ニューラルネットワークは出力ノードの値が最も大きいクラスに入力を分類するものとする。

補題 1. n ノードの決定木 $DT: R^d \rightarrow c$ は H を活性化関数に用いた $O(n)$ ノード、深さ3のNNで表現することができる。

Proof. NNの深さ1のノードとしてDTの各エッジが成立する場合に1、それ以外は0を出力するノードを作成する。すなわち、DTのエッジが $f_1 \geq a$ ならば $H(f_1 - a)$ を作成し、 $f_1 > a$ ならば $1 - H(a - x)$ を作成する。(1はこのノードの出力を受け取るノードのバイアスとして与えられる)

次に深さ2のノードとしてDTの各葉ノードを表現するノードを作成する。各ノードはDTにおける葉ノードから根までの枝に対応するノードの和から、DTにおける葉ノードの深さ-0.5を引いた値を活性化関数の入力として受け取る。最後に、深さ3の出力ノードとして各クラスに対応するノードを作成し、各クラスに属する葉ノードと対応するノードの出力を入力として受け取る。このNNがDTを表現することは明らかである。実際、DTにおいて入力が到達する葉ノードに対応する深さ2のノードのみが1を出力し、他の深さ2のノードは0を出力する（少なくとも1つの深さ1のノードの出力は0であるため）

□

例えば、下図のニューラルネットワークは左図の決定木を模倣する。

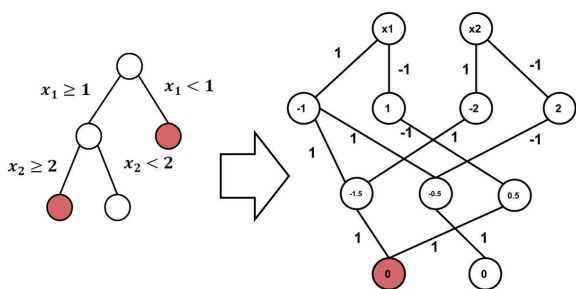


図 1: エッジの数字は重み、ノードの数字はバイアスを表す。例えば、深さ1の左端のノードは $H(x_1 - 1)$ を表す。

この補題を用いて、多くの区分線形関数に対し、ランダムフォレストはその関数を活性化関数として用いた深さ3のニューラルネットワーク以下の表現能力しか持たないことが示せる。以降、 $S \subset R^d$ は有限集合であるとする。

定理 2. n ノードの決定木 $DT: S \rightarrow c$ は、非有界な領域における傾きが異なる区分線形関数 $g: R \rightarrow R$ を活性化関数として用いた $O(n)$ ノードの深さ3のNNで表現することができる。

Proof. S は有限集合であるから、任意の S の要素が DT の識別境界上に存在しないと仮定してよい。 f は、 $a \neq b$ を満たすある定数 a, b に対して、 $\lim_{x \rightarrow \infty} f(x) = a$ 、 $\lim_{x \rightarrow -\infty} f(x) = b$ を満たすとする。この時、 f を用いて ($x \neq 0$ において) H を表現することができる。実際、

$$H(x) = \lim_{\epsilon \rightarrow 0} (f(\frac{x}{\epsilon}) - b) \frac{1}{a - b} = \frac{1}{a - b} \lim_{\epsilon \rightarrow 0} f(\frac{x}{\epsilon}) - \frac{b}{a - b} \quad (2)$$

である。さらに、 g を非有界な領域において $ax + b, a'x + b'$ と表される区分線形関数であるとする。この時、 $g(x+1) - g(x)$ は非有界な領域においてそれぞれ、 $a(x+1) + b - ax - b = a$ 、 a' となるため、 g を用いて、 H を活性化関数として用いたNNを表現することができる（図2）。したがって、補題1から非有界な領域における傾きが異なる区分線形関数を活性化関数として用いた $O(n)$ ノード、深さ3のNNでDTを表現することができる。 □

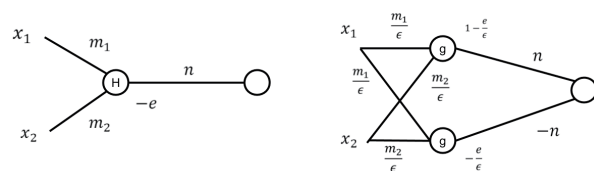


図 2: g として ReLU 関数を用いた場合、活性化関数として H を用いたネットワーク (左図) は右図のネットワークで表現される

定理 3. n ノードのランダムフォレスト $RF: S \rightarrow c$ は非有界な領域における傾きが異なる区分線形関数を活性化関数として用いた $O(n)$ ノードの深さ3のNNで表現することができる。

Proof. ランダムフォレストを構成するそれぞれの木に対して、対応するNNを作成する。深さ2の各ノードはそれぞれ対応する決定木の葉ノードに入力が到達したときに1、それ以外の場合には0を出力する。したがって、これらのNNの出力ノードを同一のノードとして見做すことで、RFに対応するNNを得ることができる。 □

これらの結果から、ランダムフォレストは深さ3のニューラルネットワーク以下の表現能力しか持たないこと、すなわち depth efficiency と全く同種のノードに関する効率性は存在しないことが分かる。

4. 木の本数と表現能力

4.1 ノード数の下界

前節から、ランダムフォレストにおいてはニューラルネットワークに対応する深さは固定であること、すなわち DNN と同様の効率性は存在しないことが分かった。次に、木の本数によって表現能力がどう変化するかの検証を行う。その為に、 n 本 $O(n)$ ノードのランダムフォレストを T 本の木からなるランダムフォレストで表現する際に必要となるノード数を求める。

入力全体を X で表す。 $x \in X$ とランダムフォレストを構成する木の各葉ノード $leaf$ に対して、 x のラベルを $c(x)$ 、 $leaf$ に割り当てられているラベルを $c(leaf)$ 、 x が $leaf$ に到達するか否かを $leaf(x)$ で表す。

補題 4. 次の性質を満たす集合 X を考える。この時、 $M = |X|$ とおくと X を T 本の木からなるランダムフォレストで表現するには $\Omega(M^{\frac{2}{T+1}})$ のノードが必要である。

X を識別する任意の T 本の木からなるランダムフォレストに対し、下記が成立する。

$$\forall x_1, x_2 \in X, c(x_1) = c(x_2) = c \rightarrow |\{leaf | c(leaf) = c, leaf(x_1), leaf(x_2)\}| < \frac{T+1}{2} \quad (3)$$

Proof. 上記の性質は任意の同一ラベルを持つ異なる 2 点は、 X を認識する任意のランダムフォレストにおける葉ノードのうち半数以上に同時に正しく認識されることはないということである。 $L = \{leaf | c(leaf) = 1\}$ 、 $l = |L|$ とする。つまり、 l はラベル 1 が割り当てられている葉の数である。同様に L' でラベル 0 が割り当てられている葉全体、 l' でラベル 0 が割り当てられている葉の数を表す。また、 M_1 で $|\{x \in X | c(x) = 1\}|$ 、 M_0 で $|\{x \in X | c(x) = 0\}|$ を表す。すなわち、 M_i はクラス i のデータの個数を表す。

$c(x) = 1$ を満たす x は少なくとも $\frac{T+1}{2}$ 以上の L の要素に対して $leaf(x)$ が成立する必要がある。一方、(3) より $leaf(x)$ が成立する L の要素のうち、どの $\frac{T+1}{2}$ 個の葉の組み合わせも他の $c(x) = 1$ を満たす x に対して同時に $leaf(x)$ を満たすことはない。したがって、各 $x \in M_1$ に対し他の $x' \in M_1$ では同時に 1 とならない $\frac{T+1}{2}$ 個の L に属する葉の組み合わせが存在するから、

$$\binom{l}{\frac{T+1}{2}} \geq M_1 \quad (4)$$

が成立する。また、ラベル 0 の場合も同様にして

$$\binom{l'}{\frac{T+1}{2}} \geq M_0 \quad (5)$$

が成立する。したがって、

$$l^{\frac{T+1}{2}} \geq M_1 \quad (6)$$

$$l'^{\frac{T+1}{2}} \geq M_0 \quad (7)$$

が成立する。したがって、このランダムフォレストの葉ノードの数は

$$l + l' \geq M_1^{\frac{2}{T+1}} + M_0^{\frac{2}{T+1}} \geq M^{\frac{2}{T+1}} \quad (8)$$

となる。したがって、必要なノード数の下界は

$$\Omega(M^{\frac{2}{T+1}}) \quad (9)$$

□

この補題を用いて、 n 本 $O(n)$ ノードのランダムフォレストを T 本の木からなるランダムフォレストで表現する際に必要なノード数の下界を求めることができる。下界は次のようになる。

定理 5. n 本 $O(n)$ ノードのランダムフォレスト $RF : \{0,1\}^n \rightarrow \{0,1\}$ を T 本の木からなるランダムフォレストで表現する際に必要なノード数の下界は $\Omega((\frac{2^n}{\sqrt{n}})^{\frac{2}{T+1}})$ である。

Proof. 任意の T ($T < n$) に対し、補題 4 の (3) の性質を満たし、 n 本 $O(n)$ ノードのランダムフォレストで識別できる X が存在することを示せばよい。各次元の要素の和が $\frac{n+1}{2}$ (ラベル 1 を持つ) もしくは $\frac{n+1}{2} - 1$ (ラベル 0 を持つ) となる n 次元の 0,1 ベクトル全体を X と置く。

X が任意の T ($T < n$) に対し、(3) の性質を満たすことを示す。 X を識別する T ($T < n$) 本の木からなるランダムフォレストが存在するとする。 X が (3) の性質を満たさないと仮定する。このとき、 $c(x_1) = c(x_2) = c$ を満たす、ある x_1, x_2 が存在して、これらに対し $leaf(x_1), leaf(x_2), c(leaf) = c$ を満たす $leaf$ が $\frac{T+1}{2}$ 個以上存在する。以降では一般性を失うことなく $c = 1$ と置く。

一方、 $x_1 \neq x_2$ であるから x_1 で 1、 x_2 で 0 となる要素が少なくとも 1 つは存在し、この要素は x_1 と x_2 の両方が同じ葉ノードに到達する木の出力に影響を与えない。したがって、 x_1 においてこの要素を反転させた x_1' はこのランダムフォレストにおいてラベル 1 と識別される。しかし、 x_1' は要素の和が $\frac{n+1}{2} - 1$ であるから、 X に属し、ラベル 0 を持つはずである。これは、 X が T 本の木からなるランダムフォレストによって識別されるということに矛盾する。したがって、 X は (3) の性質を満たす。

X が n 本 $O(n)$ ノードのランダムフォレストによって識別されるということは容易に示すことができる。それぞれの木において 1 つの特徴量を評価し、1 ならばラベル 1、0 ならばラベル 0 を出力すれば良い。また、

$$|X| = 2^{\binom{n}{\frac{n+1}{2}}} \geq \frac{2^n}{\sqrt{n}} \quad (10)$$

であるから、必要なノード数の下界は

$$\Omega((\frac{2^n}{\sqrt{n}})^{\frac{2}{T+1}}) \quad (11)$$

□

この定理から同じ関数を表現する際に、木の本数が n 本であるランダムフォレストと比較して T 本の木から構成されるランダムフォレストでは多量のノードを必要とすることがあること、すなわちランダムフォレストは木の本数に対してノード数に関する効率性を持つことが示された。

5. 結論と展望

本研究では、ランダムフォレストにおいて DNN の depth efficiency のようなノード数に関する効率性が存在するかということについて考察を行った。まず、多くの区分線形関数に対し、ランダムフォレストはその関数を活性化関数として用いた 3 層のニューラルネットワーク以下の表現能力しか持たないことを示した。このことから、ランダムフォレストにおいてはニューラルネットワークと全く同種の効率性は存在しないことが示された。次に、 n 本の木から構成されるランダムフォレストを T 本の木から構成されるランダムフォレストで表現する際には、 T が n よりも十分に小さい場合には多くのノードが必要となることを示した。このことから、木の本数に関してはノード数に関する効率性が存在することが示された。今後の課

題としては、定理 5 の下界の改善、 n 本の木から構成されるランダムフォレストを T 本の木から構成されるランダムフォレストで表現する際に必要となるノード数の上界の導出などが考えられる。

参考文献

- [1] Guido Montufar, F., et al. "On the number of linear regions of deep neural networks." Advances in neural information processing systems. 2014. p. 2924-2932.
- [2] Maithra Raghu, et al. "On the expressive power of deep neural networks." arXiv preprint arXiv:1606.05336 (2016).
- [3] Matus Telgarsky. "Representation benefits of deep feedforward networks." arXiv preprint arXiv:1509.08101 (2015).
- [4] Lech Szymanski, and Brendan McCane. "Deep networks are effective encoders of periodicity." IEEE Transactions on Neural Networks and Learning Systems 25.10 (2014): 1816-1827.
- [5] Olivier Delalleau, and Yoshua Bengio. "Shallow vs. deep sum-product networks." Advances in Neural Information Processing Systems. 2011. p. 666-674
- [6] Yishay Mansour. "Pessimistic decision tree pruning based on tree size." In Press of Proc. 14th International Conference on Machine Learning. 1977. p.195-201.
- [7] Shalev-Shwartz, Shai, and Shai Ben-David. Understanding machine learning: From theory to algorithms. Cambridge university press, 2014. p139
- [8] Mayumi Thais Oshiro, Pedro Santoro Perez, and José Augusto Baranauskas. "How many trees in a random forest?." International Workshop on Machine Learning and Data Mining in Pattern Recognition. Springer, Berlin, Heidelberg, 2012. p. 154-168
- [9] Yoshua Bengio, Olivier Delalleau, and Clarence Simard. "Decision trees do not generalize to new variations." Computational Intelligence 26.4 (2010): 449-467.