CF-008

# 深層学習モデルにおける特徴選択層の実装
# Feature Selection Layer for Deep Neural Network

若松 浩平 † 須鎗 弘樹 ‡ 森 康久仁 ‡
Kohei Wakamatsu Hiroki Suyari Yasukuni Mori

## 1. はじめに

近年, 画像の分類や超解像, ボードゲームのプレイなど多岐にわたる分野で成果を出しているディープラーニングという機械学習の一手法が存在する. ディープラーニングは他の機械学習の手法と異なり, モデルに対し人間が推論のために規則を与えたり, 特徴量を抽出する必要がない. 生のデータから推論に必要な特徴を抽出し, 非常に高い精度で推論を行うことができる. しかし, 抽出された特徴は多層のネットワークの重みに潜在しており, 推論の判断根拠は何かといった情報を人間が理解できる形で得ることは困難である. また, ディープラーニングは特徴の抽出は行うが, 推論に有効な特徴を選ぶ (特徴選択) 操作は行っていない.

特徴選択は機械学習の前処理としてしばしば用いられる. 学習に有用な特徴量のみを利用することでモデルの性能や学習速度を向上させる場合があり, 様々な研究が行われている [1]. この特徴選択は場合によっては特徴量について新たな情報を得るために利用できることがある. 例えば, あるモデルに入力する特徴量の数を減らしても精度が得られる場合, 使用されない特徴量はモデルの出力に貢献していない, つまり出力の理由を説明する力を持たないことがわかる. 一方で選択された特徴量は出力を説明する何らかの情報を有しているといえる.

ディープラーニングに特徴選択の機能を持たせることで, 出力に貢献する特徴量を知ることができる. どんな特徴量を利用したのか, 何を学習したのかを解明することによって, モデルの改善のヒントや, データに対するより深い知見が得られる可能性がある.

これまでディープラーニングの判断根拠を可視化, または理解しようとする試みは多数行われてきた. 一例としてネットワークの出力を最大化する入力の作成の研究 [2] や, 入力のどの部位にネットワークが反応し出力したかを可視化する研究 [3] などが行われているものの, 特徴量単位に焦点を絞った前例はない. そこで本研究では, 特徴量を選択する機能を持つ層を実装する. これにより, ディープラーニングで利用される特徴量を可視化し, その判断根拠の一端を理解することを目的とする.

## 2. 特徴選択層 (Feature Selection Layer)

本節では深層学習モデルにおいて特徴選択機能を有する特徴選択層について述べる. 特徴選択層は入力に対し 1 対 1 の関係で重みを乗算した値を出力する層である. 図 1 に特徴選択層のネットワークの構造を, 式

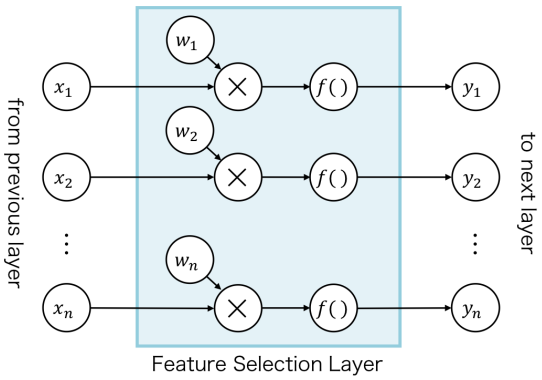



図 1: 特徴選択層のネットワーク構造

(1) に計算方法を示す.

$$\boldsymbol{y} = f(\boldsymbol{W} \otimes \boldsymbol{x}) \tag{1}$$

ここで, $\boldsymbol{x}$ は入力, $\boldsymbol{W}$ は重み, $\boldsymbol{y}$ は出力をそれぞれ表し, $\otimes$ は対応する位置の要素同士に関して乗算を行うことを表す. また, $f(\cdot)$ は活性化関数を表す. また, ここでの重みは学習の際に誤差逆伝播によって更新される.

重みに応じて後段のネットワークへの入力は変化することから, 特徴選択層は特徴量を擬似的に選択する機能を持つといえる. また, 重みは誤差逆伝播によって逐次更新される. 学習後, 重みの絶対値の大きい特徴量は後段のネットワークで出力を得るために有効な働きをしていると考えられる. 一方で絶対値の小さい重みとなった特徴量は出力に対しあまり貢献していないと考えられる. この結果からディープラーニングで用いられる特徴量を知ることができる.

## 3. 実験

提案した特徴選択層の機能を検証するため, 結果が既知である人工的なデータセットでの実験を行った. その後, 実際のデータに対する特徴選択層の有効性とその他の特徴選択機能を持つモデルとの違いを検証するため, その他のデータセットについて, 特徴に重み付けを行うことができる他のモデルと実験を行い, 精度と学習後の各特徴量に対する重みまたは重要度を比較する. これらの実験では結果を理解しやすくするために特徴選択層の活性化関数は用いずに線形で出力した.

### 3.1. 既知のデータセットによる機能検証

目的変数 $y$ を (2) 式のように決定した回帰問題を利用して特徴選択層の機能を検証した. データセットとして説明変数に値域が [0,1] である 10 次元のデータ $\boldsymbol{x} = (x_0, x_1, \ldots, x_9)^T$ をランダムに 1000 個生成した.

†千葉大学大学院 融合理工学府 Graduate School of Science and Engineering, Chiba University
‡千葉大学大学院 工学研究院 Graduate School of Engineering, Chiba University

(2) 式の通り, 目的変数に対して説明変数 $x_0, x_1$ が非線形な関係を, $x_2, x_3, x_4$ が線形な関係をもつ. 一方で $x_5, x_6, \ldots, x_9$ は目的変数の決定に寄与していないノイズである.

$$y = 5(x_0 - 0.5)^2 - 4(x_1 - 0.5)^2 + 3x_2 + 2x_3 + x_4 \quad (2)$$

これらのデータのうち 700 件を学習用データ, 300 件を評価用データとし, 学習用データに対し後述する 4 種の DNN モデルで学習させ, 評価用データに対する平均二乗誤差についての比較を行なった. 特徴選択層が挿入されたものについては学習後の重みを確認した.

#### 3.1.1. 実験条件

基本の構成は同じで, 全結合層の活性化関数の有無, 特徴選択層の有無が異なる計 4 種ネットワークを作成した. 表 1, 図 2 にこれらのネットワークの構成を示す. 図 2 内の網かけされた部分の層の有無が表 1 に対応している. これらのネットワークを表 2 に示す条件で学習させた.

表 1 に示すネットワークは全結合層の有無によってモデルの表現力が, 特徴選択層の有無によって特徴選択機能を有するかどうかが異なる. これら計 4 種類のネットワークの学習結果の比較を行うことで, 特徴選択層の有無による精度の変化, 特徴選択層の有効性, モデルの表現力の変化による特徴選択層の重みの変化について検証する.

表 1: 実験で利用するネットワーク

| モデル名 | 全結合層の活性化関数 | 特徴選択層 |
|---|---|---|
| network1 | 有 | 無 |
| FS network1 | 有 | 有 |
| network2 | 無 | 無 |
| FS network2 | 無 | 有 |

表 2: 学習の諸条件

| | |
|---|---|
| epoch 数 | 3000 |
| バッチサイズ | 100 |
| 学習アルゴリズム | Adam |
| 損失関数 | 交差エントロピー関数 |

#### 3.1.2. 実験結果・考察

表 3 に評価用データセットに対する平均二乗誤差 (MSE) を示す. 表 3 より, ネットワークの構造によらず, 特徴選択層が挿入されている場合の方が損失が小さいことがわかる. 特徴選択層は各特徴量に対して重みを持つため, 調整できるパラメータが増え結果としてモデルの表現力が向上したと考えられる.

また, 図 3 に FS network1, FS network2 それぞれの学習後の特徴選択層の重みのヒートマップを示す. 図 3 では, FS network1 の重みが (2) 式の係数の大きさと対応していることが確認できる. 一方で FS network2 では非線形の項に対する重みが小さな値を取っていることが確認できる. これらのネットワークは活性化関数の有無によってモデルの表現力に差があり, ある特徴量が目的変数に対して非線形な関係を持つ場合, それを表現できるか否かが異なる. FS network1 はネットワークの中で目的変数に対して非線形な関係を持つ特徴量を捉えることができ, その結果非線形項に対する重みが大きくなった. 一方 FS network2 では非線形な関係をネットワーク内で捕らえられていないために非線形項に対する重みが小さくなったことがわかる. 以上の結果から特徴選択層によって特徴量に付加される重みは, 実際の目的変数に対する特徴量の関係を表現するのではなく, ネットワークの中で目的の変数を表現するために利用されるかどうかを表現していると考えられる.

表 3: 各モデルの評価用データに対する平均二乗誤差

| model | MSE |
|---|---|
| network1 | 0.01025 |
| FS network1 | 0.01814 |
| network2 | 0.2343 |
| FS network2 | 0.2386 |

### 3.2.wine データセットによる特徴選択層の機能検証

次に, UCI Machine Learning Repository で提供されているデータセットである wine データセットを利用して特徴選択層の機能を検証した. wine データセットは 13 の特徴量を持つデータが 3 つのカテゴリに分類されており, 計 178 のレコードを持つ. その中から学習用データとして 7 割のデータを, 評価用のデータとして 3 割のデータをランダムに選択した. 学習用のデータについて特徴選択層を含む DNN と, 同様に出力に対する特徴量の寄与の程度を重みや重要度として表現するその他の機械学習手法で学習し, 評価用データで精度を算出した. また, それぞれのモデルの学習後の重み, または特徴の重要度を比較した.

#### 3.2.1. 学習条件

出力に対する特徴量の寄与の程度を重みや重要度として表現する学習モデルとして以下のモデルを使用した. また, DNN を除くモデルは scikit-learn および xgboost を利用して実験を行った.

- 特徴選択層を含む DNN
- Ridge
- Lasso
- ロジスティック回帰 (L1 正則化)
- ロジスティック回帰 (L2 正則化)
- サポートベクタマシン (線形カーネル)
- ランダムフォレスト

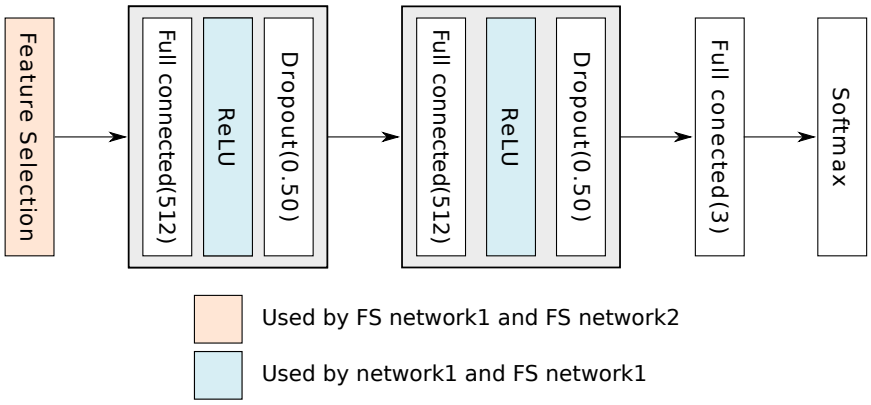
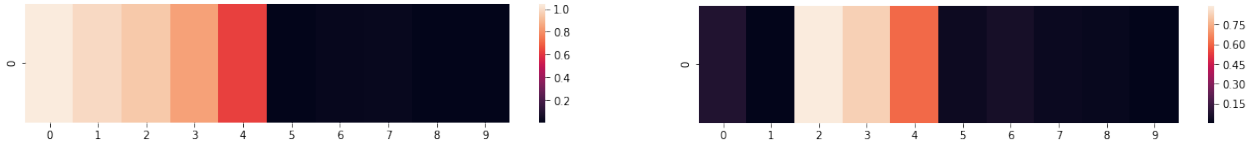



図 2: ネットワーク構造



(a) FS network1　　(b) FS network2

図 3: 学習後の各モデルの特徴選択層の重み

・勾配ブースティング木

図 4 に特徴選択層を含む DNN の構造を示す．また，表 4 に学習の各種条件を示す．本実験で用いた DNN は前節の実験で利用したネットワークに比べ全結合層数が増加している．また，Ridge および Lasso については正則化パラメータ (alpha) を 0.01 とした．また，ランダムフォレストおよび勾配ブースティング木については弱学習器の数を 1000 とした．その他，特に言及していないものについては scikit-learn および xgboost のデフォルトの値を利用した．

表 4: 学習の諸条件

| | |
|---|---|
| epoch 数 | 3000 |
| バッチサイズ | 32 |
| 学習アルゴリズム | Adam |
| 損失関数 | 交差エントロピー関数 |

### 3.2.2. 実験結果・考察

表 5 に各モデルの学習後の評価用データに対する精度を示す．表 5 から精度では ridge, lasso がおよそ 0.1 減少する結果となったが，それ以外の各手法では大きな差はないことが確認できる．

また，図 5 に学習後のそれぞれの特徴量に対する重みや重要度を示すヒートマップを示す．また，表 6 に図 5 中の各横軸の番号と wine データセットの属性の対応を示す．

図 5 について確認すると，特徴選択層込みの DNN を除く手法では Flavanoids もしくは Proline に対する重みまたは重要度が相対的に大きい値を取り，その他は小さな値をとっていることが確認できる．一方で特徴選択層込みの DNN では Flavanoids と Proline に対する重みがが相対的に大きな値を取っていることは他の手法と同じだが，他の手法に比べ全体的に重みが大きく，相対的に大きな差のある特徴量が少ないことが確認できる．また，ロジスティック回帰 (L2 正則化)，ridge を除く他の手法でも重みもしくは重要度の小さい Total phenols, Nonflavanoid phenols については特徴選択層も同様に重みが小さくなった．この結果から他の手法で利用される特徴量と似た傾向を持つことがわかる．

また，特徴選択層込みの DNN と同程度の精度を出している他のモデルと重みについて比較すると，特徴選択層込みの DNN は多くの特徴量をネットワークの推論のために利用していることがわかる．この結果から，DNN では高い精度を出すために他の手法で利用されていない特徴量を用いている可能性，または，特徴選択層では必要最小限の特徴に対して重みをつけていない可能性が推測される．

## 4. まとめ

本研究では DNN において学習可能な特徴選択層の実装を行なった．自作のデータセットによる実験では特徴選択層がネットワークで利用される特徴量に対して重み付けを正しく行えることが確認できた．wine データセットによる実験で特徴選択層は他の手法に比べ多くの特徴量を利用するよう重みづけをしていることが確認された．この結果より DNN が他の手法で利用されていない特徴量を利用している可能性，または特徴選択層が必要最低限の重み付けをしていない可能性が示唆された．よって，今後活性化関数の導入やモデルの損

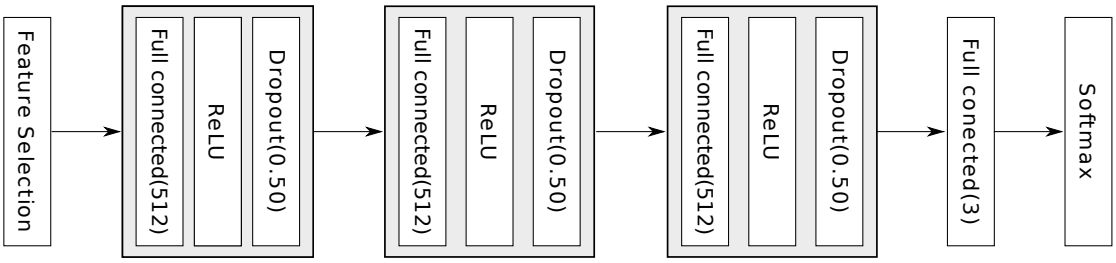



図 4: ネットワーク構造

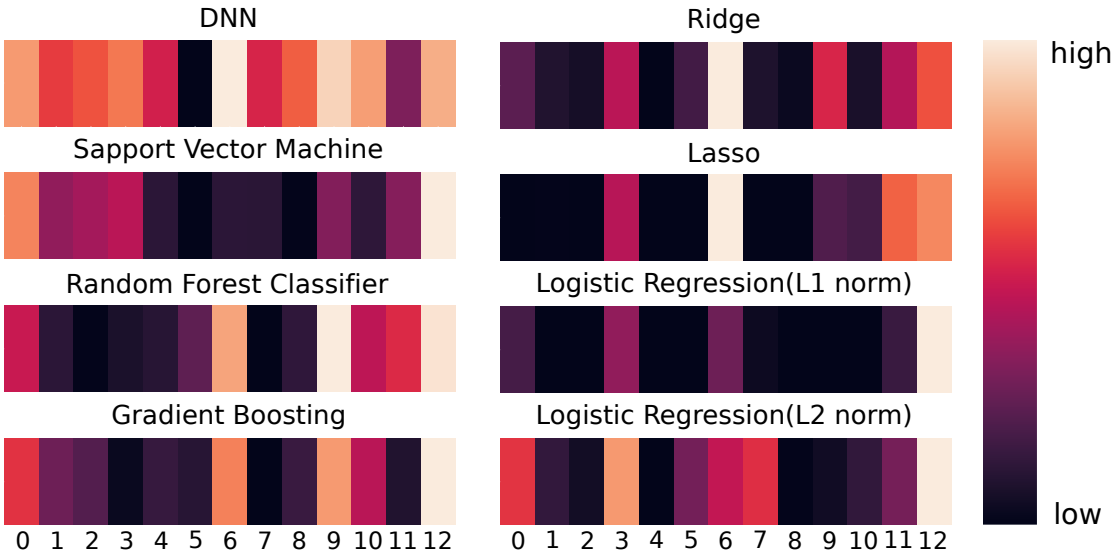



図 5: 学習後の各モデルの重みもしくは重要度

表 5: 各モデルの評価用データに対する精度

| model | accuracy |
|---|---|
| DNN with Feature Selection | 1.0000 |
| Ridge | 0.8898 |
| Lasso | 0.8683 |
| ロジスティック回帰 (L1 正則化) | 1.0000 |
| ロジスティック回帰 (L2 正則化) | 0.9815 |
| ランダムフォレスト | 1.0000 |
| 勾配ブースティング木 | 0.9815 |
| サポートベクタマシン | 0.9815 |

表 6: 数字と属性名の対応表

| 数字 | 属性名 |
|---|---|
| 0 | Alcohol |
| 1 | Malic acid |
| 2 | Ash |
| 3 | Alcalinity of ash |
| 4 | Magnesium |
| 5 | Total phenols |
| 6 | Flavanoids |
| 7 | Nonflavanoid phenols |
| 8 | Proanthocyanins |
| 9 | Color intensity |
| 10 | Hue |
| 11 | OD280/OD315 of diluted wines |
| 12 | Proline |

失関数の変更などで目的変数に対して関係のある特徴量をより正確に選択する機能を実装する必要がある.

## 参考文献

[1] Sklansky, J. : Comparison of Algorithms that Select Features for Pattern Classifiers, Pattern Recognition, Vol.33, No.1, pp.25-41 (2000)

[2] Aravindh Mahendran, Andrea Vedaldi : Understanding Deep Image Representations by Inverting Them, arXiv:1412.0035v1 [cs.CV] 26 Nov 2014

[3] Daniel Smilkov, Nikhil Thorat, *et al.* : SmoothGrad: removing noise by adding noise, Science 358, arXiv:1706.03825v1 [cs.LG] 12 Jun 2017