様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　２ 年　 ６ 月　 ５ 日現在

機関番号：１２５０１
研究種目：基盤研究(C)（一般）
研究期間：2017～2019
課題番号：１７Ｋ００２２７

研究課題名（和文）Big Data時代における超高次元特徴選択フレームワークに関する研究

研究課題名（英文）A Study on a High Dimensional Feature Selection Framework in the Big Data Era

研究代表者
森　康久仁（Mori, Yasukuni）

千葉大学・大学院工学研究院・助教

研究者番号：４０３６１４１４

交付決定額（研究期間全体）：（直接経費）　　2,200,000円

研究成果の概要（和文）：本研究では，様々な分野ですぐれた結果を出している深層学習を利用した特徴選択を行える新たな層モデルを提案した．提案した手法は，使用するネットワークモデルの入力層の次に，各特徴と１対１にユニットを配置した特徴選択層と呼ばれる層を追加し，訓練データを用いて学習を行う．これにより，対象としているタスクにおいて，そのタスクに有効に作用する特徴に対応するユニットの重みが大きくなり，不必要な特徴の重みが小さくなることが期待できる．したがって，この重みの値を利用することで，対象タスクに有効な特徴の選択をすることができる．

研究成果の学術的意義や社会的意義
本研究で提案した特徴選択手法を利用することで，従来の手法では非常に難しかった，超高次元のデータに対しても特徴の選択をすることが可能になった．これにより，例えば，数千を超える特徴集合の中から，注目している識別タスクに有効に作用する重要な特徴を選別することが可能になり，これまで以上に探索的データ解析における新たな知見が得られる可能性を見出した．

研究成果の概要（英文）：In this study, I propose a new layer model for feature selection using deep learning, which produces excellent results in various fields. The proposed method adds a layer called a feature selection layer in which units are arranged on a one-to-one basis with each feature to the input layer of the network model used, and performs learning using training data. As a result of learning, it is expected that the weight of the unit corresponding to the feature which effectively acts on the task increases and the weight of the unnecessary feature decreases in the target task. Therefore, by using the value of this weight, it is possible to select effective features for the target task.

研究分野：パターン認識

キーワード：特徴選択

様　式　C－１９、F－１９－１、Z－１９（共通）

１．研究開始当初の背景

計算機技術の発展に伴い，高速な中央演算処理装置や十分な容量の記憶装置を安価に利用できるようになってきた．さらに，ネットワーク技術の進歩から世界中の情報がオンデマンドで簡単に手に入るようになってきている．このような技術革新の結果，扱われるデータの規模が徐々に大きくなってきており，大規模なデータ(Big Data) の効果的な解析手法の確立が望まれている．

これまで，機械学習やデータマイニングと呼ばれる分野において，「識別系において如何に高精度な識別規則を生成するか」という点に焦点を当て，多くの識別アルゴリズムの報告がなされている．この観点からは，深層学習の発展に伴い，素晴らしい成果が報告されている．しかしながら，我々が日々蓄積していくデータの規模は爆発的に増大しており，計算機性能の向上スピード以上の早さで増えている．したがって，そのようなデータの量をなるべく少なくする技術が今後必要となると予想できる．そこで，本研究では識別系の全段階にある前処理部に焦点を当て，大量な特徴を持つデータから，識別に有効な特徴を選択・解析する問題について考える

２．研究の目的

これまでに提案されてきている代表的な特徴選択・解析のアルゴリズムでは，データの持つ特徴の数が増えるにつれて，計算量が膨大になり現実問題として最適な特徴の組を取り出すことが難しく，処理能力の高い計算機を何台も並列化させ何日も実行して始めて結果が出る場合もある．しかしながら，あらゆる状況でそのような計算機環境や人的コストを求めることは現実的には不可能であり，単体の計算機および，ごく僅かな人的パワーでデータを処理・解析できることも重要な点である．したがって，本研究の目的は，識別系に送られるデータを適切に処理可能な「大規模データに対しても有意に動作する特徴選択フレームワークの解析・開発」を行うことである．

３．研究の方法

図１のような入力に対し1 対1 の関係で重みを乗算した値を出力する層を提案する．各入力に乗算される重みは学習の際に誤差逆伝播によって更新されるものとする．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．

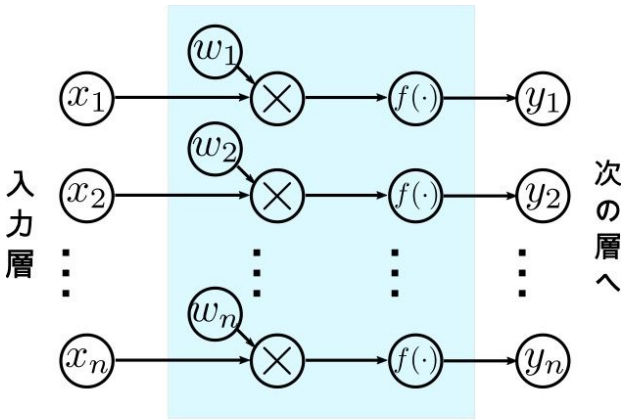

図 1 特徴選択層の構造

４．研究成果

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．目的変数 y を式(1)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 $\boldsymbol{x} = (x_0, x_1, \cdots, x_9)^T$ をランダムに1000 個生成した．目的変数は10個の変数のうち5個のみと関連しており，残りの5個は目的変数の決定に無関係な変数（ノイズ）である．

$$y = 5(x_0 - 0.5)^2 - 4(x_1 - 0.5)^2 + 3x_2 + 2x_3 + x_4 \qquad (1)$$

図２に学習後の特徴選択層の重みのヒートマップを示す．図が示している通り，目的変数に関連がある x_0 から x_4 までの特徴の重みが大きくなっており，それ以外の特徴の重みが小さくなっていることが確認できる．また，線型的な関係だけでなく，非線形な関係を持つ特徴量も捉えられていることがわかる．このことから，特徴選択層によって各特徴量に付加さ

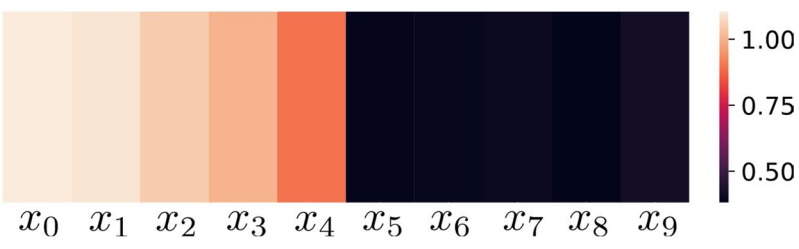

図 2 各特徴の重要度の可視化

れる重みは，後段のネットワークによって処理されるタスクにおける，各特徴量の重要度を表していると考えられる．

本研究で提案した特徴選択層はネットワークで利用される特徴量に対して，正しく重み付けを行えることが確認できた．本研究は，探索的に行う従来の特徴選択とは異なり，ネットワークの学習を取り入れた新たなアプローチである．したがって，非常に優れた識別性能を持つ深層学習において，特徴選択層を利用することで，これまでとは異なる特徴量を探索することができる可能性がある．また，従来の方法では，探索が不可能な超高次元データの場合でも，特徴の重要度の判定が行うことができ，医療データなどへの応用も期待できる．

５．主な発表論文等

〔雑誌論文〕　計2件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 森康久仁	４．巻 35
２．論文標題 機械学習の基礎と医療画像への応用	５．発行年 2018年
３．雑誌名 医用画像情報学会論文誌	６．最初と最後の頁 42--47
掲載論文のDOI（デジタルオブジェクト識別子） https://doi.org/10.11318/mii.35.42	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 鎌倉伊織，森康久仁，斎藤陽一，原田元，松葉育雄	４．巻 Vol. J100-A No.8
２．論文標題 VARモデルを用いたてんかん患者の発作脳波の伝播経路推定	５．発行年 2017年
３．雑誌名 電子情報通信学会論文誌	６．最初と最後の頁 309--312
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計10件（うち招待講演　2件／うち国際学会　0件）

１．発表者名 中村栞，田中健太，横田元，足立拓也，町田洋一，堀越琢郎，太田丞二，森康久仁，須鎗弘樹
２．発表標題 MRI画像による乳がんのサブタイプ 分類
３．学会等名 人工知能学会全国大会
４．発表年 2019年

１．発表者名 小名木佑来，橋本拓磨，太田丞二，高岡浩之，横田元，堀越琢郎，森康久仁，小林欣夫，須鎗弘樹
２．発表標題 Dilated Convolutionを用いたImage Post Processing Metal Artifact Reductionの提案
３．学会等名 医用画像情報学会
４．発表年 2019年

１．発表者名
小名木 佑来，橋本拓磨，黒澤隆那，村田泰輔，古山良延，太田丞二，髙 岡浩之，横田元，森康久仁，小林欣夫，須鎗弘樹

２．発表標題
深層学習を用いたメタリックアーチファクトの低減手法

３．学会等名
人工知能学会医用人工知能研究会

４．発表年
2018年

１．発表者名
森康久仁

２．発表標題
機械学習の基礎と医療画像への応用

３．学会等名
医用画像情報学会（招待講演）

４．発表年
2018年

１．発表者名
若松浩平，須鎗弘樹，森康久仁

２．発表標題
深層学習モデルにおける特徴選択層の実装

３．学会等名
第17回情報科学技術フォーラム

４．発表年
2018年

１．発表者名
森康久仁

２．発表標題
機械学習の基礎とその周辺技術

３．学会等名
日本放射線技術学会（招待講演）

４．発表年
2017年

１．発表者名
花待宏典，森康久仁，松葉育雄

２．発表標題
リカレンスプロットによるてんかん発作脳波の規則性解析

３．学会等名
電子情報通信学会ソサイエティ大会

４．発表年
2017年

１．発表者名
松永大，森康久仁，松葉育雄

２．発表標題
複数のテクニカル分析を考慮したRNNによる株価変動予測

３．学会等名
電子情報通信学会ソサイエティ大会

４．発表年
2017年

１．発表者名
関島優介，森康久仁，松葉育雄

２．発表標題
混合正規分布を用いた変分自己符号化器モデルの提案

３．学会等名
電子情報通信学会ソサイエティ大会

４．発表年
2017年

１．発表者名
関島優介，森康久仁

２．発表標題
混合正規分布を用いた変分自己符号化器による特徴抽出

３．学会等名
電子情報通信学会総合大会

４．発表年
2018年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考