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　４ 年　 ６ 月 ２２ 日現在

機関番号：７７１０３
研究種目：基盤研究(C)（一般）
研究期間：2018～2021
課題番号：１８Ｋ１１４８７

研究課題名（和文）スパースモデリングの導入による人が理解できる深層学習

研究課題名（英文）Comprehensible deep learning with the aid of sparse modeling

研究代表者
石川　眞澄（Ishikawa, Masumi）

一般財団法人ファジィシステム研究所・研究部・特別研究員

研究者番号：６０２２２９７３

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：　学習結果の理解が困難という深層学習の問題点解消に向け、情報圧縮のための積層自己符号化器を対象とし、スパース化のため６種類の正則化項を深層学習に適用した。自乗誤差評価値と結合数からなるパレート最適曲線で構成される閉領域の面積を正則化項の有効性評価とすることを提案し、標準的な結合重みの(a)L1ノルムよりも(b)選択的L1ノルムが有効であり、隠れ層出力の(c)選択的L1ノルム、(d)選択的L2ノルム、(e)KL-情報量、(f)非対角共分散和の(b)への追加が更に有効であることを実証した。得られた疎構造モデルから恒等写像と疑似恒等写像の存在及び情報損失との関係を明らかにし、人の理解向上に貢献した。

研究成果の学術的意義や社会的意義
　スパースモデリングに用いられる正則化項がいろいろ提案されてきたが、知る限りではこれらの有効性を評価する方法が無かった。本研究はデータ適合度及びスパース度（結合数）からなるパレート最適性の概念を利用して正則化項の有効性指標を提案し、大規模実データを用いてその有効性を実証した点に研究成果の学術的意義がある。
　正則化項の有効性指標を用いてその有効性を評価し、少なくとも積層自己符号化器において最も有効な正則化項群による学習結果を人が理解することを可能とし、またクラス分類課題においても同様のアプローチで疎構造を得ることができ、人に理解できる人工知能への第一歩を示せた点に研究成果の社会的意義がある。

研究成果の概要（英文）：Deep learning has a serious drawback in that the resulting models tend to be a black box. A sparse modeling approach applied to stacked autoencoders for information compression is expected to ameliorate the drawback. I propose to use the concept of Pareto optimality composed of data fitting and the sparseness of models for judging the effectiveness of regularization terms. Based on it, I demonstrate that compared to (a)the popular L1-norm of connection weights, (b)the selective L1-norm of connection weights is more effective, and that the addition of (c)the selective L1 norm, (d)the selective L2 norm, (e)KL-divergence, or (f)off-diagonal squared covariances of hidden outputs are still more effective. The resulting sparse structure of stacked autoencoders enables the clarification of information compression mechanism composed of identity mappings and pseudo-identity mappings. It further clarifies the relation between mappings and information loss, and contributes to human understanding.

研究分野：人工知能

キーワード：深層学習　積層自己符号化器　スパースモデリング　正則化　疎構造　情報圧縮　ブラックボックス

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

2012 年以降、パターン認識コンテストにおいて深層学習が従来手法の性能を凌駕し大きな注目を浴びてきた。また 2016 年に AlphaGo が囲碁のトップ棋士に打ち勝ったことも大きな衝撃を与えた。深層学習はデータに基づいて特徴量を生成できる点が学習におけるブレークスルーとの主張がなされた。ただいずれの場合も深層学習による学習結果がブラックボックスとなり、明確な特徴量が生成されたわけではなく、深層学習の判断を人が理解することが困難である。

製造現場や製品に深層学習を導入しても、時間的に変動しまた不完全情報に溢れた実世界で無謬ではあり得ない。その際、何故誤りが発生したかを究明し、解決策を講じ、これらを社会に公表することが企業の社会的責任である。ブラックボックス的な深層学習では、誤りの解明が困難であることが企業の深層学習活用を妨げており、その技術的解決が急務である。深層学習の急速な進展及び近未来に予想される産業や家庭への普及を念頭に置き、本研究は人が理解できる深層学習の実現に向けた第一歩という位置付けを有する。

２．研究の目的

筆者はニューラルネットワークモデルに対する正則化項として、世界に先駆け結合重みの L1 ノルム（ラプラス正則化、結合重みの絶対値和）を提案した(1989)。ただ、これだけでは分散表現が生じることから、隠れ層出力の明確化学習及び選択的 L1 ノルムを併せ提案し、提案手法の有効性を実証した。なお本研究の申請時点で、深層学習に対し、結合重みの L2 ノルム、結合重みの L1 ノルム、KL-divergence、隠れ層出力の非対角共分散和などの正則化項が提案されていた。

本研究は､筆者が提案した正則化項にこの過程で培ったさまざまなアイデア･ノウハウも加え、申請時点以降他で提案された種々の正則化項も念頭に置き、これらを組み合わせ駆使して人が理解できる深層学習を探求することを目的とした。

３．研究の方法

深層学習は一般に超多層ネットワークであり、これに直接スパースモデリングを適用しても、計算量が膨大であり、学習結果を人が理解するのは困難であろうと考えた。そこで、まず取り扱い易い部分構造に着目し、これに対して適切なスパースモデリング手法を適用し、これを段階的にネットワーク全体に拡張するアプローチが現実的と考えられる。積層自己符号化器は貪欲学習を用いて分割学習できるので、本研究では深層学習の典型例としてこれを取り上げた。

スパースモデリングのための正則化項がさまざま提案されているが、最も良く用いられる(a)結合重みの L1 ノルムはその原理が単純明快で理論的解明もある程度可能なので、これを基準と考える。申請者が提案した(b)結合重みの選択的 L1 ノルムは自乗誤差評価が大きくなりがちな L1 ノルムの欠点を解消できるのでもう一つの基準とする。これらに対して後述の４種類の正則化項のいずれかを追加し併せて６種類の正則化項の有効性を実証する。

正則化項の有効性評価を行うための実データとして、本研究では、(A)赤ワインデータ（12 次元、1599 個）、(B)胎児心拍数陣痛計データ（21 次元,2126 個）、(C)米国国勢調査データ（55 次元、45155 個）を取り上げた。(A)(B)は連続値データであり、(C)は連続値とカテゴリー値が混在している。

まず３層の自己符号化器を対象とし、(a)結合重みの L1 ノルム、(b)結合重みの選択的 L1 ノルムに加え、(c)隠れ層出力の選択的 L1 ノルム、(d)隠れ層出力の選択的 L2 ノルム、(e)隠れ層出力の KL-divergence、あるいは(f)隠れ層出力の非対角共分散和を(b)に追加する正則化項群を用いて学習を行う。なお以下では (a)を L1 ノルム、(b)を選択的 L1 ノルム、(c)を選択的 h ノルム、(d)を選択的 h2 ノルムと略称することがある。次に、(A)(B)は貪欲学習をもう１回実行し５層モデルとし、(C)は貪欲学習をもう２回実行し９層モデルとし、(a)～(f)の正則化項を用いて学習を行う。正則化項の有効性評価を用いることにより最も有効な正則化項群を決定でき、人が理解できる深層学習に近づくことが可能となる。

当初の研究計画は取り扱い易いと考えられる積層自己符号化器のみを対象としたが、研究の進展とともにこのアプローチがより広範な対象、すなわちクラス分類課題や主成分分析（PCA）に対しても同様に適用できるのではないかと考え、時間の許す限りこれらに対しても人が理解できる深層学習への接近を試みた。

４．研究成果

(1) 米国国勢調査データを用いた積層自己符号化器

米国国勢調査データは連続値とカテゴリー値から構成され、後者はダミー変数として扱う。モデル構造は下層から順に{61，14，10，8，6，8，10，14，61}個の素子を持つ９層自己符号化器である。図１に３種類の正則化項群に対応する５個の L1 ノルム重み（マーカー表示）からなるパレート最適曲線群を示す。(a)結合重みの L1 ノルムよりも(b)結合重みの選択的 L1 ノルムの

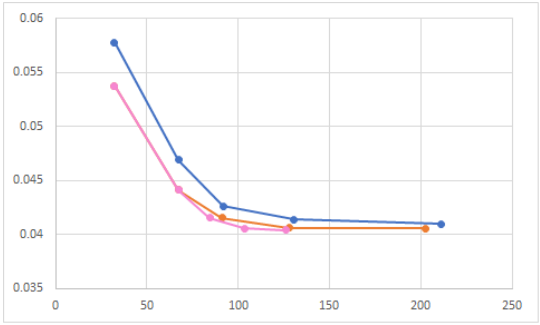

図１. 米国国勢調査データから得られる積層自己符号化器で(a)L1 ノルム(青)、(b)選択的 L1 ノルム(橙)、(b)選択的 L1 ノルム＋(c)選択的 h ノルム(ピンク)のパレート最適曲線。縦軸は自乗誤差評価値、横軸は結合数を表す。学習率 0.05、選択的 L1 ノルムの閾値 0.1、h ノルム重み 0.00001、選択的 h ノルムの閾値 0.1、折線上のマーカーは L1 ノルム重みに対応し、右から順に{0.00005, 0.0001, 0.0002, 0.0005, 0.001}である。なお、各マーカーは５個の初期乱数で学習した結果の平均値を示す。

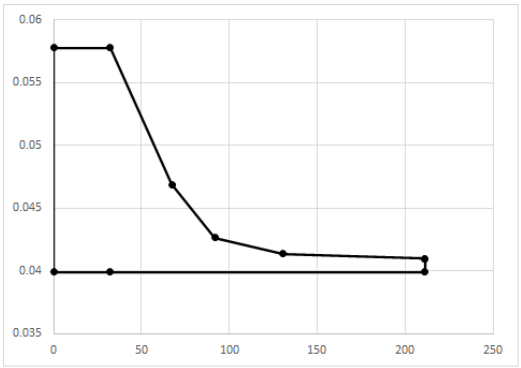

図２. 米国国勢調査データの場合の L1 ノルムの有効性指標としての閉領域面積。下方の水平な線分は、正則化項無しで５個の初期乱数で十分な回数学習しその最小の自乗誤差評価値を表す。

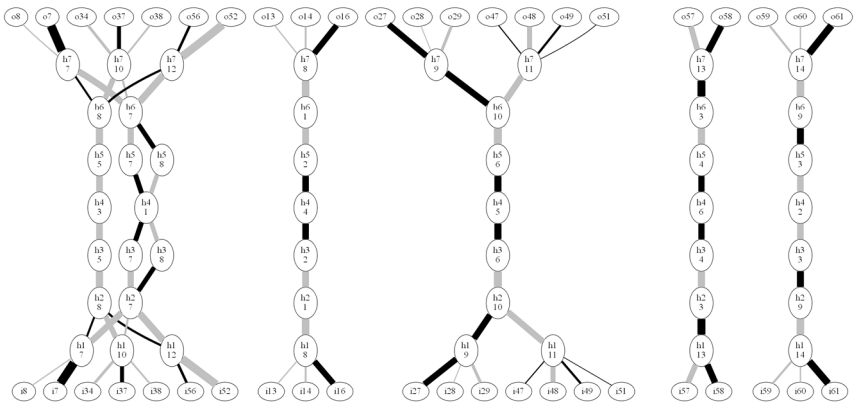

図３. 米国国勢調査データで(a)L1 ノルムによる構造図。学習率 0.05、L1 ノルム重み 0.0002。灰色の結合重みは正の、黒色の結合重みは負の重みを表す。なお５個の初期乱数での学習のうち乱数#1 の結果を示す。図４, 図５、図７も同様である。

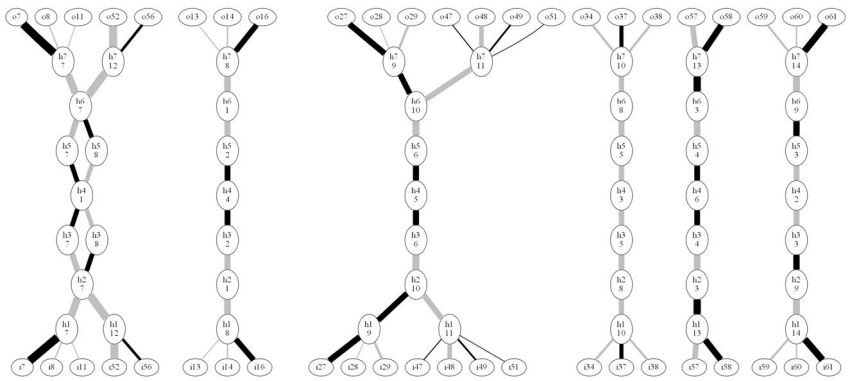

図４. 米国国勢調査データで(b)選択的 L1 ノルムによる構造図。L1 ノルム重み 0.0002、選択的 L1 ノルム閾値 0.1。

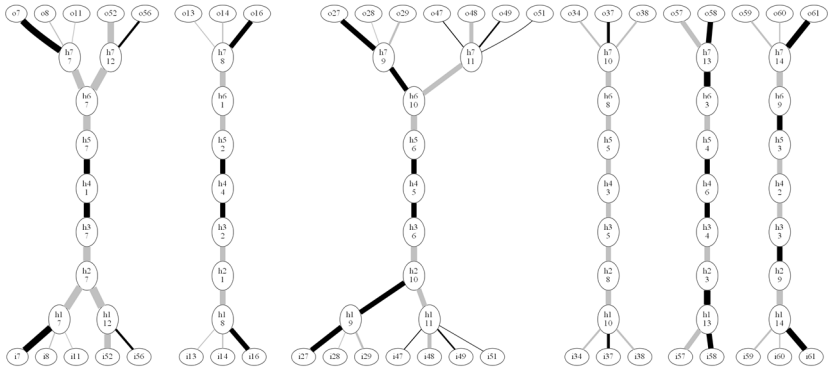

図５. 米国国勢調査データで(b)選択的 L1 ノルム＋(c)選択的 h ノルムによる構造図。L1 ノルム重み 0.0002、選択的 L1 ノルム閾値 0.1、h ノルム重み 0.00001、選択的 h ノルムの閾値 0.1。

方が原点に近く、(b)に(c)選択的 h ノルムを追加すると更に原点に近くなる。正則化項の有効性としては(a)L1 ノルムよりも(b)選択的 L1 ノルムが優れており、(b)に(c)選択的 h ノルムを追加すると更に性能が向上することが分かる。図２は正則化項の有効性指標として用いる閉領域の定義を示す。この面積が小さいほどパレート最適曲線が原点に近く、対応する正則化項の有効性が高いことを示している。ただ、(c)選択的 h ノルム追加による性能向上はわずかである。これは 61 次元データを６次元まで圧縮させるというかなり無理な課題を遂行しているためと考えられる。実際、下記に述べる(2)のクラス分類課題での対応する図６では、(d)選択的 h2 ノルムの追加により大きな性能向上が得られている。

図３～５は米国国勢調査データに(a)L1 ノルム、(b)選択的 L1 ノルム、(b)に加えて(c)選択的 h ノルムを適用して得られた構造図を示す。図３、図４、図５の順に得られた構造が単純化され、(b)選択的 L1 ノルムおよび(c)選択的 h ノルムの同時使用の有効性を示している。61 入力を 6 素子にまで圧縮するので、恒等写像は無理であるが、複数個の入力が上位層の素子につながる疑似恒等写像が６個生成される。なお、積層自己符号化器において圧縮率が大きいとすべての入力を取入れるのは無理なので、分散の小さい入力素子が無視されることになる。

表１は各種 DB とタスクに対して６通りの正則化項群を適用した場合の有効性指標の一覧表である。同表から(a)L1 ノルムを基本とすると、(b)選択的 L1 ノルムは常に性能が向上していることが分かる。これに(c)選択的 h ノルム、(d)選択的 h2 ノルム、(e)KL-divergence、(f)非対角共分散和のいずれかを追加すると更なる性能向上が見られる。

表１ DB×タスクに対する正則化項の有効性指標。有効性指標は図２で示される面積に相当する。SAE は積層自己符号化器、class はクラス分類課題、PCA は主成分分析、L1 は(a)結合重みの L1 ノルム、sL1 は(b)結合重みの選択的 L1 ノルム、sh は(c)隠れ層出力の L1 ノルム、sh2 は(d)隠れ層出力の L2 ノルム、KL は(e)隠れ層出力の KL-divergence、decs は(f)隠れ層出力の非対角共分散和である。各列で最も有効な正則化項組み合わせを太字で示す。

DB×task 正則化項	米国国勢調査 SAE	胎児心拍数陣痛計 SAE	赤ワイン SAE	米国国勢調査 class	胎児心拍数陣痛計 class	胎児心拍数陣痛計 PCA
L1	1.311015	0.127590	0.130191	1.080737	7.326976	0.244130
sL1	0.924626	0.104179	0.085135	0.869520	4.768420	0.180121
sL1 + sh	**0.845926**	**0.072876**	**0.053682**	0.664465	**2.742830**	**0.128059**
sL1 + sh2	0.928940	0.073267	0.053955	**0.657461**	3.531300	0.129032
sl1 + KL	0.856266	0.079314	0.059532	0.766004	3.313636	0.143750
sL1 + decs	0.855108	0.077304	0.111627	0.718478	2.927744	0.136336

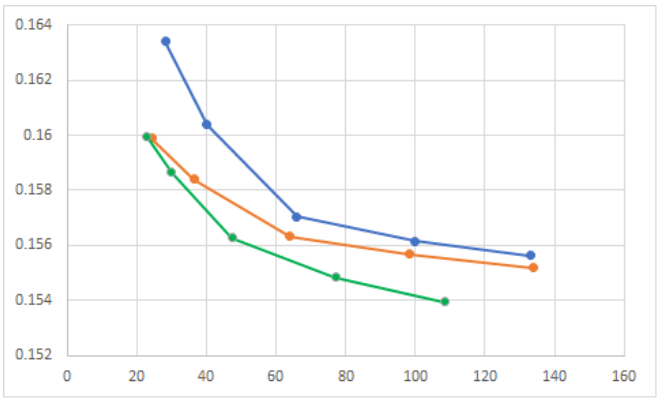

図６．米国国勢調査データのクラス分類用モデルのパレート最適曲線。(a)L1 ノルム(青)、(b)選択的 L1 ノルム(橙)、(b)選択的 L1 ノルム＋(d)選択的 h2 ノルム(緑)。縦軸は自乗誤差評価値、横軸は結合数を表す。学習率 0.05、選択的 L1 ノルムの閾値 0.1、h2 ノルム重み 0.0001、選択的 h2 ノルムの閾値 0.01、折線上のマーカーは L1 ノルム重みに対応し、右から順に {0.0005, 0.0001, 0.0002, 0.0005, 0.001}である。

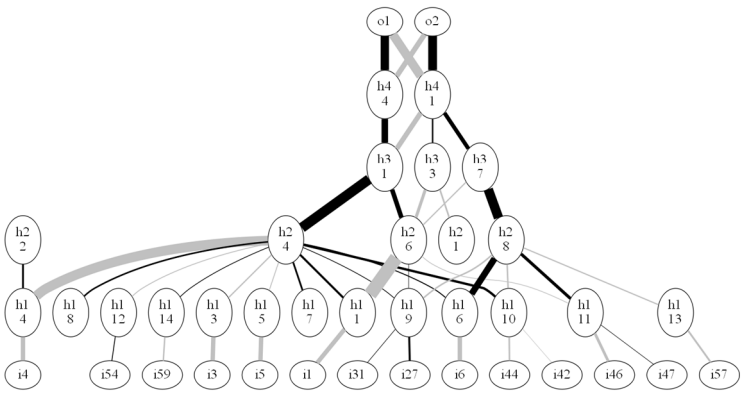

図７．米国国勢調査データのクラス分類課題で(b)選択的 L1 ノルムに(d)選択的 h2 ノルムを追加した場合の構造図。L1 ノルムの重み 0.0002、h2 ノルム重み 0.0001、選択的 h2 ノルムの閾値 0.01、乱数#1 を用いた場合のモデル構造。

(2) 米国国勢調査データを用いたクラス分類課題

米国国勢調査データを年収５万ドル以下/５万ドル超というクラス分類で学習させる。図６は図１に対応するパレート最適曲線を示す。(b)選択的 L1 ノルムに(d)選択的 h2 ノルムを追加した場合の性能向上が、図１で(c)選択的 h ノルムを追加した場合よりもずっと大きいことが読み取れる。図７はクラス分類課題における典型的なケースのモデル構造を表す。本タスクの分類正答率は調べた限りでは高々85%程度なので、得られたモデル構造に明快な意味づけを与えるのは困難である。同じデータで積層自己符号化器を学習させた場合には何とかモデル構造に意味づけを与えることができたが、この場合も圧縮率をさらに上げると意味づけが困難になる。

(3) スパースモデリングによる胎児心拍数陣痛計データの主成分分析

主成分分析（PCA）はデータの特性を解明するのに有効で広く用いられている手法である。各主成分には殆どすべての変数が含まれるため、各主成分を特徴量として把握することが困難という問題点がある。主成分に含まれる変数の数を減らすため L1 ノルムを利用するスパース PCA が提案された。これに対し本研究では多様な正則化項を用いて主成分に含まれる変数をさらに減らし、人が理解できるスパース PCA を指向する。

図８は胎児心拍数陣痛計データに対する PCA およびさまざまな正則化項を用いるスパース PCA の累積寄与率及び絶対値が 0.01 より大きな変数の平均個数を示す。PCA および BP 学習を用いる PCA と比較して、(a)L1 ノルムを用いるスパース PCA の平均個数が半分程度になっているが、(b)選択的 L1 ノルムではさらに小さく、これに(c)選択的 h ノルムを追加すると一層平均個数が減少することが分かる。これにより人が理解し易い主成分群に近づいた。

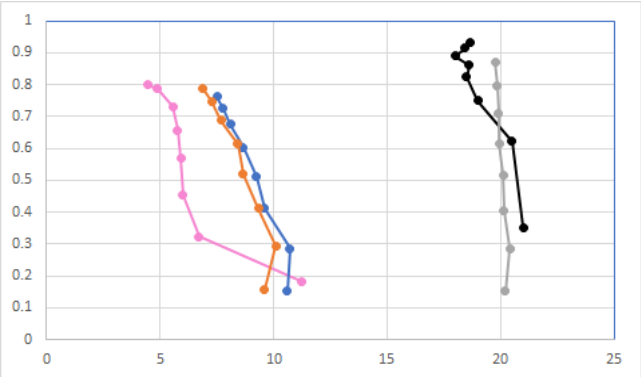

図８. 胎児心拍数陣痛計データの PCA およびスパース PCA。縦軸は累積寄与率、横軸は上位 n 番目までの主成分のうち絶対値が 0.01 より大の平均個数である。各折線のマーカーは下から順に第一主成分、第２主成分などに対応する。PCA（黒色）、BP（灰色）、(a)L1 ノルム（青）、(b)選択的 L1 ノルム（橙）、(b)選択的 L1 ノルム＋(c)選択的 h ノルム（ピンク）。L1 ノルム重みは右から順に{0.00005, 0.00007, 0.0001}である。

(4) 得られた成果の位置づけとインパクトおよび今後の展望

深層学習の学習結果がブラックボックスであるという困難を解消するための有力なアプローチの一つがスパースモデリングである。スパースモデリングに際して用いる正則化項が種々提案されてきたが、知る限りではこれらの有効性を評価する方法が無かった。本研究では、データ適合度及びスパース度（結合数）からなるパレート最適性の概念に基づく閉領域面積により正則化項の有効性を定量的に評価することを可能とした。さらに実データでその有効性を実証した。以上が得られた成果の位置づけである。

単なるパターン認識に留まらず、それに基づいて車の自動運転を行うとか、プラントの自動操業を行うなどパターン認識の結果が重大な結果をもたらす現場では、ブラックボックス的な深層学習では何かが起きた場合のリスクが大き過ぎるので、ブラックボックスからの脱皮が不可欠である。本研究はこのための第一歩としてデータ適合度及びスパース度（結合数）の観点から見て有効な正則化項群を見つけ出す方法を呈示し、今後の技術開発の方向性を示すというインパクトを与えるものである。

もちろんこのためのアプローチとして有望と考えられるものはスパースモデリングだけではない。たとえば DARPA は“explainable AI”を提唱し、関連論文数は 2018 年以降急増している。著者の理解では、explainable AI は与えられたモデルの入力群から出力群への影響を解明するというさまざまなアプローチを提案している。これとは違い、スパースモデリングは学習によりモデル構造自体をスパースにするというアプローチを採っている。両者は相補関係にあるように思われ、両者の何等かの形での融合が今後必要となり、またより良い理解にとって有効ではないかと考えている。今後このような方向での研究が進展することを期待したい。

５．主な発表論文等

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

１．著者名 石川眞澄	４．巻 30
２．論文標題 巻頭言「自己組織化マップの俯瞰と今後への期待」	５．発行年 2018年
３．雑誌名 知能と情報	６．最初と最後の頁 63
掲載論文のDOI（デジタルオブジェクト識別子） 10.3156/jsoft.30.2_63	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計7件（うち招待講演　0件／うち国際学会　0件）

１．発表者名 石川眞澄
２．発表標題 積層自己符号化器における人が理解できる深層学習を目指して
３．学会等名 電子情報通信学会ニューロコンピューティング研究会
４．発表年 2019年

１．発表者名 石川眞澄
２．発表標題 積層自己符号化器における冗長表現およびブラックボックスの抑制
３．学会等名 電子情報通信学会ニューロコンピューティング研究会
４．発表年 2019年

１．発表者名 石川眞澄
２．発表標題 層毎貪欲学習および各種正則化項によるクラス分類深層ネットワークのスパース化
３．学会等名 電子情報通信学会ニューロコンピューティング研究会
４．発表年 2020年

１．発表者名
石川眞澄

２．発表標題
自己符号化器とスパースPCAの性能比較

３．学会等名
電子情報通信学会ニューロコンピューティング研究会

４．発表年
2020年

１．発表者名
石川眞澄

２．発表標題
スパースモデリングによる積層自己符号化器の情報圧縮機構の明確化

３．学会等名
第19回情報科学技術フォーラム（FIT2020）

４．発表年
2020年

１．発表者名
石川眞澄

２．発表標題
連続値とカテゴリー値データが混在する深層学習における種々のスパース化とその有効性評価

３．学会等名
電子情報通信学会ニューロコンピューティング研究会

４．発表年
2022年

１．発表者名
石川眞澄

２．発表標題
積層自己符号化器学習における種々のスパース化の適用と有効性評価および情報圧縮機構の解明

３．学会等名
電子情報通信学会ニューロコンピューティング研究会

４．発表年
2022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

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関