The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

深層学習における学習ネットワークからの分類パターンの抽出

Extraction of Classification Patterns from Deep Learning Networks

安藤 雅行 *1 Masayuki Ando
河原 吉伸 *2*4 Yoshinobu Kawahara
砂山 渡 *3 Wataru Sunayama
畑中 裕司 *3 Yuji Hatanaka

*1滋賀県立大学大学院工学研究科
Graduate School of Engineering, The University of Shiga Prefecture

*2九州大学 マス・フォア・インダストリ研究所
Institute of Mathematics for Industry, Kyushu University

*3滋賀県立大学工学部
School of Engineering, The University of Shiga Prefecture

*4理化学研究所 革新知能統合研究センター
RIKEN Center for Advanced Intelligence Project

In deep learning, there is a problem that concrete classification patterns for deriving reasons for classification are often incomprehensible. In this paper, we propose a classification patterns extraction system from deep learning networks and verified the effectiveness of the system. The proposed system takes out learning networks from the learning result of deep learning and extracts classification patterns from the learning networks. Then the system displays the extracted classification patterns so that users of the system can interpret the learning networks. In verification experiments, the significance of the extracted classification patterns was estimated by chi-square test. The results showed that users of the system can extract classification patterns effective for interpretations of the learning networks by using the proposed system.

1. はじめに

インターネットの普及などによって世の中で増大する文章データへの対処方法として，深層学習を用いたテキストマイニングシステムが注目されている [1][2]．深層学習は，一般に多層から構成されるニューラルネットワークを用いた学習を指す．

その一方で，深層学習は，その出力を導いた根拠についての解釈が困難であることも知られている．テキスト分野においても，学習ネットワークの解釈を行うことで深層学習の分類基準をより深く理解できれば，例えば医療分野において新人とベテランの書いた電子カルテの違いから，良い電子カルテを書く方法を容易に理解したり，企業においても良い報告書や企画書を書く方法を短時間で習得できるなど，深層学習の新しい活用が期待できる．

深層学習の分類基準を示す研究としては，アテンションと呼ばれる手法を用いた研究 [3][4] が注目されている．しかし，アテンションは内部でどのような学習が行われているかは考慮していない．一方で画像に関してだが，深層学習の学習ネットワークの解釈に注目したものがある [5]．また，ニューラルネットワークを用いた研究では，中間層の役割と重みの持つ意味について考察したもの [6] や中間層が学習から，入出力の様々な関係を得ることを目指したもの [7] があり，中間層に注目すれば，出力に関して意味のある情報が得られることを示している．また，重みに注目した研究として，重みを用いて入力データ間の相関関係を抽出する研究 [8] があるが，分類基準を提示する目的で重みに注目したものは存在しない．

そこで本研究では，RNN（Recurrent Neural Network）を使用し，テキスト集合の学習によってネットワークの層に付けられた重みの値を取り出し，学習ネットワークに対する解釈を行うために，それぞれの出力特有の特徴となる分類パターンの抽出を行うシステムを提案する．

連絡先: 安藤 雅行，滋賀県立大学大学院工学研究科先端工学専攻，〒 522-8533 滋賀県彦根市八坂町 2500, oh23mandou@ec.usp.ac.jp

2. 深層学習の重みを用いたテキストの分類パターン抽出システム

本章では，本研究で開発した深層学習の重みを用いたテキストの分類パターン抽出システムについて，システムの構成とその詳細について述べる．

2.1 分類パターン抽出システムの構成

分類パターン抽出システムでは，まず，図 1 に示すように，各分類先ごとにラベル付けしたテキスト集合を RNN にて分類し，その分類先を導いた学習ネットワークと，学習ネットワーク上の重みから，提案システムの分類パターンの抽出・可視化処理によって各出力 (分類先) を導くネットワーク上のパスと，パス上の各ノードの情報（そのノードで学習された単語）の決定，表示を行う．最後に，システムの利用者は，システムによって得られた学習ネットワークの表示を自分が見やすいように調整し，分類パターンを抽出する．そして分類パターンの意味を理解しやすくするための機能を利用できる．

2.2 深層学習による学習ネットワークの形成

2.2.1 文中の単語のベクトル化

深層学習で学習を行う前に，テキストデータは文中の単語を抽出したあと，単語は One hot 法 [9] と呼ばれる手法に従い単語ベクトルの羅列に直される．そして，文中の各単語を単語ベクトルに置き換え，深層学習への入力データとする．なお，抽出される単語は名詞する．これは，文章の特徴とその順序関係をより学習・抽出しやすくするためである．

2.2.2 学習によるネットワークの重み付け

One hot 法によって単語ベクトルの羅列に変換され，さらにラベル付けされたテキストデータは，RNN の学習にて，そ

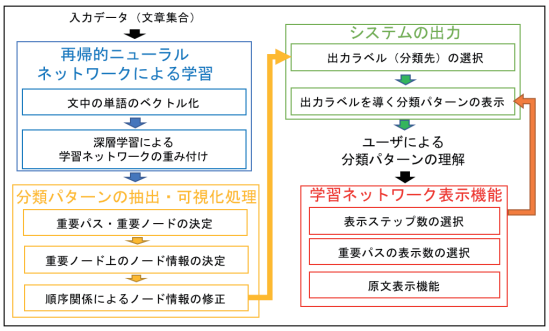

図 1: 分類パターン抽出システムの構成

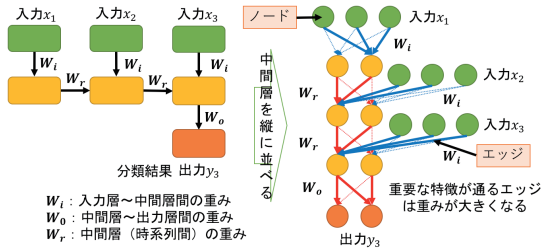

図 2: RNN の学習ネットワークと学習の様子

れぞれの出力ラベル（分類先）を導くネットワークへの重み付けがされていく．その様子を図 2 に示す．入力文章は各単語がベクトル化され，タイムステップごとに単語ベクトルが順番に入力されていく．また，RNN での分類時は，最後の単語が入力されたタイミングで，出力層から出力される．

2.3 学習ネットワークからの分類パターンの抽出・可視化処理

本提案システムの分類パターンの抽出･可視化処理では，RNN によって得られた学習ネットワークから，各出力を導く最も関係が強いパス（ネットワーク上のエッジの繋がりの線）を決定する処理を行う．この出力と繋がりが強いパス（重要パスと呼ぶ）の決定について，図**??**に示した 4 層全結合型ネットワークモデルを例として，具体的な手順を述べる．

まず，ある分類先（図**??**の例では層 D のノード b）に到達するパスについて，パス上のエッジについた重みの積で定義される，重要度と呼ばれる値を算出する．図**??**の入力層のノード i からの太矢印のパス上のエッジに付いた重みを $W1_{a,i}$，$W2_{i,j}$，$W3_{j,b}$ とすると，このパスの重要度 $B_{a,i,j,b}$ は以下の式 (1) で導かれる．

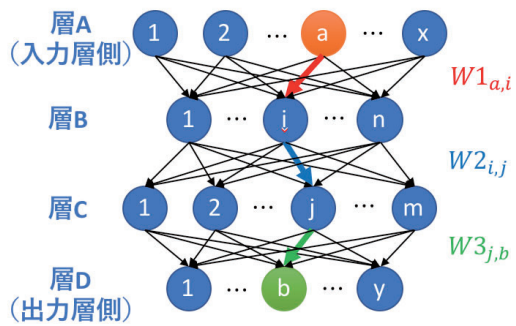

図 3: 4 層全結合型ネットワーク

$$B_{a,i,j,b} = W1_{a,i} \times W2_{i,j} \times W3_{j,b} \tag{1}$$

そして，式 1 で出力 b に到達する全てのパスの重要度を計算して比較し，最も値の大きいパスを，重要パスと決定する．出力 b の重要パスの重要度を MB_b として，その計算式 (2) を以下に示す．図 2 の右図のように，出力層と中間層間の重み $\boldsymbol{W_o}$ のあと，中間層間の重み $\boldsymbol{W_r}$ を繰り返し辿ることで，過去の中間層を遡ることになる．

$$MB_s = \max_{i,j,b} B_{a,i,j,b} \tag{2}$$

次に，重要パス上のノードについて，ノードの情報を決定の処理を行う．ノード情報の決定方法は，出力ごとの重要パスを決定した時と同様に，入力層からノード情報を決定したい中間層ノード間の重み $\boldsymbol{W_i}$ の大きさから，ノード間の結びつきの強さを求める．ただし，図 2 に示す様に，各中間層にはそれぞれ個別の入力層が対応している．ここで，RNN の入力には One hot 法による単語ベクトルを用いているため，入力層ノードにはノードごとに 1 種類の単語が対応していることになる．そこで，重要パス上の入力層ノードの単語を参照し，その単語をノードの情報と決める．最後に，ノード情報の単語について，前後の中間層ノード情報の単語と，原文中でその順番で表示されているかどうかを確認し，されていないなら単語の重要度を下げる．

こうして各出力を導く重要パスとパス上のノード情報を表示した学習ネットワークから，過去の中間層の重要ノードから現在の中間層の重要ノードの単語より，時系列を考慮した特徴の並びとしての分類パターンを抽出することができる．

2.3.1 出力ラベルを導く分類パターンの表示

本研究で開発した分類パターンの抽出システムでは，分類先に強く結びつく，重要パスの集合としての学習ネットワーク上に重要ノードの情報が表示される．例として，5 種類のお菓子の作り方に関するテキストの分類を行った場合の，システムのメイン画面を図 4 に示す．表示分類先は「マカロン」とする．このネットワークは，RNN が学習した情報を，過去から順番に中間層上に表示したものである．過去から出力層直前までの中間層の重要ノード情報（単語）の並びは，分類パターンを示し，重要パス 1 本につき 1 つの分類パターンが表示されていることになる

図 4 では，入力層を除いて選択したステップ数（図 4 では 2）だけ中間層と，出力層が表示され，出力層には分類された分類先を示す出力ノードが表示されている．また，分類先を示す出力ノードには，そこから選択したステップ数の長さで，選択した重要パス数だけ重要パスが表示されている．そして，重要パス上の各中間層ノードには，ノード情報として単語が表示されている．

2.4 システム上での学習ネットワーク表示機能

システムには，利用者が分類パターンの抽出・理解を行いやすいように，その表示内容を変更できる機能がある．その主なものを表 2 に示す．

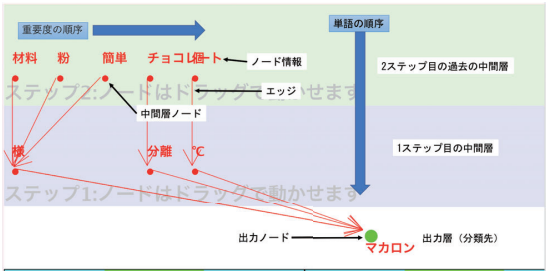

図 4: システムの画面

表 1: 学習ネットワーク表示機能

機能名	効果
表示するステップ数の選択	RNN において，何ステップ過去までの中間層を表示するか決定する
重要パスの表示数の選択	分類先ごとに，何本の重要パスを表示するか決定する
原文表示機能	分類パターン（順序を考慮した単語の組み合わせ）が，原文中でどのように出現しているかを表示する

3. 分類パターン抽出システムの有効性の検証実験

本章では，深層学習の重みを用いた分類パターンの抽出システムにより，学習ネットワークから抽出できる分類パターンが，テキストデータを理解する（学習ネットワークの解釈を行う）ために有効かを検証した実験について述べる．なお，本実験では，分類パターンを「2 種類の異なる単語の，順序を考慮した組み合わせ」と定義する．単語を 2 種類にした理由は，最も基本的な分類パターンであるためである．

3.1 実験準備

3.1.1 使用テキストデータと学習モデル

分類パターン抽出の対象とするテキスト「童話 5 種類」について，詳細を表 2 に示す．また，計算上の分類パターンの総数（抽出単語数 × 抽出単語数 − 抽出単語数）とその中で「ある分類パターンが，ある分類先に特有のものである」という仮説を有意水準 5%で検定したカイ二乗検定で，有意性があると推定された分類パターン数を表 3 に示す．

続いて，深層学習として使用した RNN モデルの概要を表 4 に示す．システム上での表示設定は，学習ネットワーク表示機能により，表示ステップ数 2，重要パス数 20 とする．また，テキスト「童話 5 種類」に対する RNN の分類精度は 100%であった．

3.2 実験手順

実験の対象者は著者 1 名とした．対象者はテキスト「童話 5 種類」について，本研究で開発したシステムを用いて，全ての分類先で分類パターン抽出システムによって抽出された分類パターン 20 個について，カイ二乗検定で有意性があるかどうか，ある場合，カイ二乗検定の順位で何位にあたるのかも調べ考察を行った．なお，カイ二乗検定の順位とは，カイ二乗検定

*1 Google：https://www.google.co.jp で「童話名　あらすじ」と検索し，表示された上位 50 位のサイトの本文をテキストとして利用

表 2: テキストデータの詳細

データ名	内容
童話 5 種類	日本の童話「かぐや姫」「鶴の恩返し」「さるかに合戦」「桃太郎」「浦島太郎」の，それぞれの概要やあらすじなどについて書かれたテキストをネット上 *1 から 1 種類あたり 50 テキストずつ用意した．

表 3: 有意性のある分類パターンの総数

分類先名	カイ二乗検定での分類パターン総数	分類パターン総数
かぐや姫	93,211	15,527,540
鶴の恩返し	9,208	15,527,540
さるかに合戦	24,839	15,527,540
桃太郎	46,749	15,527,540
浦島太郎	53,334	15,527,540
平均	45,468	15,527,540

で有意性のあると推定された分類パターンの集合を，カイ二乗値が高い（より強く有意性があるとされる）順に並べた時の順位である．

3.3 結果と考察

抽出された分類パターンで，有意性がないとされた分類パターンについての考察を行うため，有意性がある・ないとされた分類パターンの個数を図 5 に示す．

図 5 より，まず有意性があるとされた分類パターンの個数と，有意性がないとされた分類パターンの個数の比率は，4 対 6 ほどであることがわかる．その中で，有意性があるとされた分類パターンについて見ると，どちらのテキストも，ほとんどがカイ二乗検定での分類パターンの上位 50%以上に含まれることがわかる．これは，カイ二乗値の高い，より有意性があるとされた分類パターンと，有意性が全くないとされた分類パターンが混ざって抽出されていると言える．

その中で，有意性がないとされた分類パターンが 60%以上と最も多い「浦島太郎」を例に考察する．「浦島太郎」の抽出された分類パターン上位 10 個を表 5 に示す．有意性がないと判断される要因は，他のテキストでも頻度が高い，もしくはそのテキスト内での頻度が低いことが原因である．しかし，それでもシステム上で重要度の高い分類パターンとして抽出されたということは，有意性があるとされた分類パターンと合わさることで，よりテキストへの理解を深めることができる補助的

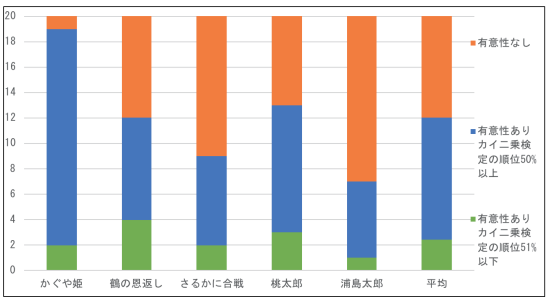

図 5: 抽出された分類パターン

表 4: 学習モデルの詳細

データ名	抽出した単語集合	入力層ノード数	中間層ノード数	出力層ノード数
童話 5 種類	名詞 3,941 種の単語	3,941	50	5

表 5: 抽出された分類パターン（「浦島太郎」）

	抽出された分類パターン			
順位	分類パターン	「浦島太郎」内の頻度	「浦島太郎」以外での頻度	カイ二乗検定での順位
1 位	世界→気	12	4	34,300
2 位	日→展開	6	3	0
3 位	風土記→日本	12	0	2,118
4 位	物語→日本	18	11	0
5 位	日→日本	16	10	0
6 位	長期間→ストーリ	3	0	0
7 位	集→嶋	3	0	0
8 位	漁師→共通	3	0	0
9 位	年→日本	19	11	0
10 位	亀→女性	16	0	964

な役割を持つと考えれる．なお，表 5 で有意性がない分類パターンのカイ二乗検定での順位は 0 と表記している．

表 5 をみると，まず単語「日本」が，4 つの分類パターンで共通しており，そのうち 3 つの分類パターンに有意性がなく，1 つが有意性があるとわかる．また，表 3 より，有意性のある分類パターンのカイ二乗検定での順位は，53,334 中 2,118，つまりカイ二乗検定での順位の上位 4%以内と高い順位であり，ここから，強い有意性を持つ 1 つの分類パターンを基準に，3 つの有意性のない分類パターンがテキストへの理解の補助を行なっていると仮定する．例として，順位 3 位の「風土記→日本」（有意性あり）から原文表示機能を用いて得られるテキストへの理解「丹後国の**風土記**など，**日本**各地で逸文されている」に，順位 4 位の「物語→日本」から得られる「浦島太郎**物語**は，**日本**書紀にも記述が見られる」と順位 9「年→日本」（有意性なし）から得られる「明治 29 **年**に巌谷小波氏が書いた物語をきっかけに**日本**中に広まった」を合わせることで，「浦島太郎物語は丹後国の風土記や日本書紀など，日本各地で逸文されており，明治 29 **年**に書かかれた物語をきっかけに，**日本**中に広まった」という理解が得られる．そして，この理解は「浦島太郎」特有のものである．よって，システムでは，有意性の高い分類パターンと，有意性がないが，それらを補助する役目を持つ分類パターンが得られていると言える．

以上より，実験結果から，本研究で開発したシステムでは，深層学習の学習ネットワークから，学習ネットワークの解釈に役立つ，分類先に対して特有の特徴を表す，有効性のある分類パターンの抽出が可能であると言える．そして，カイ二乗検定での分類パターンと比較することで，カイ二乗検定では有意性がないとされた分類パターンでも，有意性のある分類パターンの補助として役立っていることが考察によって判明し，カイ二乗検定では得られない隠れた有効性のある分類パターンを抽出することができると結論つけた．

4. おわりに

本研究では，複数のテキストデータの分類を深層学習である RNN で行い，学習ネットワークの解釈を行うための，分類パターンの抽出システムの開発を目的とした．本研究の特徴として，重みを辿ることで，RNN の学習ネットワーク内の情報の伝達を過去に向かって探索している点が挙げられる．システムの有用性を確かめる検証実験では，があると推定された分類パターンと，本システムで抽出した分類パターンをのカイ二乗検定で有意性を検定することで，本研究で開発したシステムにより分類パターンの抽出が容易になり，テキストの特有の特徴などを理解しやすくなったと結論づけた．今後の研究では，理解した内容からより容易にテキスト自体への解釈を行えるよう，分類パターンの抽出・可視化アルゴリズムや，システム上での表示を検討していきたい．

参考文献

[1] ボレガラ ダヌシカ，“自然言語処理のための深層学習”，人工知能学会誌，Vol.29, No.2, pp.195–201, 2014

[2] Ebru Arisoy, Tare N. Sainath, Brian Kingsbury, Bhuvaba Ramabhadran, “Deep Neural Network Language Models”, In Proceedings of the NAACLHLT Workshop, Will We Ever Really Replace the N-gram Model?, pp.20-28, 2012

[3] M Daniluk, T Rocktaschel, J Welbl, S Riedel, “Frustratingly Short Attention Spans in Neural Language”, ICLR, 2017

[4] A.Vaswani, N.Shazeer, N.Parmar, J.Uszkoreit, L.Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin, “Attention Is All You Need”, CoRR, vol. abs/1706.03762, 2017

[5] 西銘 大喜，“ディープニューラルネットワークによる画像からの表情表現の学習”，人工知能学会全国大会論文集 3L4-3，pp.1-4, 2015

[6] 立石 雅彦，山崎 晴明，“手書き数字認識における改装ニューラルネットワークの中間層に関する考察”，情報処理学会論文誌，Vol.30，No.10，pp.1281-1288, 1989

[7] 松倉健太郎，村田 昇，“ニューラルネットワークの中間層における独立な特徴量の抽出”，電子情報通信学会技術研究報告，Vol.106，No.102，pp.63-67, 2006

[8] Michael Tsang, Dehua Cheng, Yan Liu, “DETECTING STATISTICAL INTERACTIONS FROM NEURAL NETWORK WEIGHTS”, ICLR, 2018

[9] M. Feurer, A. Klein, K. Eggensperger, J. Springenberg, M. Blum, and F. Hutter. “Efficient and robust automated machine learning”, In Neural Information Processing Systems (NIPS), 2015