2B3-NFC-02a-2

ニューラルネットワークを用いたTweetデータの分類に関する研究

A Study on Tweet Classification by Neural Networks

北島 良三 *1 Ryozo Kitajima　上村 龍太郎 *2 Ryotaro Kamimura　内田 理 *3 Osamu Uchida　鳥海 不二夫 *4 Fujio Toriumi

*1東海大学大学院 総合理工学研究科
Graduate School of Science and Technology, Tokai University

*2東海大学 情報教育センター・東海大学大学院 総合理工学研究科
IT Education Center and Graduate School of Science and Technology, Tokai University

*3東海大学 情報理工学部情報科学科
Dept. Human and Information Science, Tokai University

*4東京大学大学院 工学系研究科
Graduate School of Engineering, The University of Tokyo

In this paper, we try to classify tweet data into two groups. One is a group containing disaster information and the other one is a group not containing it. It is important for disaster victims to classify tweet data because, they have to quickly gather information such as disaster warning or shelter information and so on. Therefore, it is useful to create a tweet data classification model which classify tweet automatically. In this experiment, we try to create a classification model by neural networks, which is considered to be suited for analyzing complex data. Experimental data was composed of 600 tweets with 25 variables. The error rates for testing data by the neural network and the logistic regression analysis were 13.56 and 38.98 (%). Thus, much better generalization performance was obtained by the neural network model. The results certainly show a possibility of neural networks for tweet data classification.

1. はじめに

本研究では Twitter の Tweet データをニューラルネットワークを用いて，災害に関する情報を含む Tweet と，含まない Tweet に分類することを試みる．災害が発生した際，Twitter は非常に強力な情報メディアとなる．それは Twitter が従来のメディアとは異なり，現場にいる多数の人物から情報が発信されるからである．しかし，Twitter は基本的にはコミュニケーションツールであり，災害時の Tweet であっても災害情報が必ず含まれているとは限らない．その為，例えば被災者が災害関連情報を得ようとする場合，多数の Tweet の中から災害に関する Tweet のみを探し出し，その中から自分に必要な情報を見つけ出す必要がある．

Twitter 登場以前においては，情報を収集する主な手段はテレビやラジオ，Web サイトなどであった．テレビはニュース番組，特集番組，速報テロップなどを通じ避難情報や安否情報といった情報を提供してくれる．ラジオも同じく，ニュース番組や特集番組にて情報を提供してくれる．Web サイトの場合は情報を必要なタイミングで入手することが出来る．しかし，メディアそれぞれには特徴があり，例えばテレビは広範囲に情報を発信することが可能である一方，特定の地域の詳細な情報を入手することが難しい場合も多い．ラジオの場合は地域放送局を聴取すればテレビよりも狭い範囲の情報を入手することが出来るが，やはりある程度広範囲の情報となってしまう．また両者とも放送のタイミングといった問題もあり，必要な情報が必要な時に入手出来るとは限らない．一方，Web サイトは必要とする地域に関する情報を必要なタイミングで入手することが可能である．ただし，情報を発信しているサイトにたどり着くことが出来ないと情報を入手することが出来ない．また，発信している情報が，テレビやラジオよりも局所的な情報であっても限界がある．詳細な状況のリアルタイムな情報発信はその現場に居る人物にしか速報性の面で出来ないのである．

現場に居る人物がリアルタイムに情報を発信するメディアの一例として Twitter が挙げられる．Twitter は多数の現地の人間から，現場の状況が直接伝わってくるという，テレビ・ラジオ・Web サイトとは全く違う性質を持ったメディアであり，東日本大震災の際に活用されたことが報告されている [吉次 11]．しかし，Twitter は情報発信している人物を「フォロー」していないとその情報を受動的に入手することは出来ず，この場合は必要な情報を Twitter 内で検索する必要がある．

災害時の混沌とした状況で，情報収集だけに長時間を割り当てることは難しい．従って，テレビやラジオを長時間受信したり，目的の情報を発信している Web サイトや Tweet を見つける為に検索ワードを試行錯誤している余裕はなく，素早く容易に必要とする情報を入手出来る方法が望ましい．

表 1 に示す様に，各メディアは一長一短で理想的なものはない．だが，それぞれの特性を組み合わせれば理想的な情報源に近づくのではないだろうかと我々は考える．

例えば Web サイトと Twitter の組み合わせである．Twitter の「現場に居るからこそ発信できるリアルタイムで局所的な情報」をフォローなしに閲覧できる Web サイトがあれば，そのサイトにさえアクセスすれば検索の手間無く必要とする情報

表 1: 各メディアの特徴

メディア	情報取得方法	発信される情報の内容	情報取得に必要な事柄
テレビ	受動的	全体的	放送電波へのチューニング
ラジオ	受動的	やや全体的	放送電波へのチューニング
Web サイト	能動的	局所的	検索能力
Twitter	受動的	とても局所的	フォロー

連絡先: 北島 良三，東海大学大学院 総合理工学研究科，〒 259-1292 神奈川県平塚市北金目 4-1-1，3btad004@mail.tokai-u.jp

を取得することが可能となる．サイトで提供する情報は発信内容に応じて関連する情報毎に整理したものとし，必要とする情報を容易に取得出来る様にしておくものとする．

しかし，このようなサイトの構築には Tweet データの選別作業が必要となる．なぜなら，Tweet は基本的に個人が自由に発信しているものであり，災害発生時に被災地に居る人物であっても必ずしも避難情報の様な有益な情報を発信しているわけではないからである．

そこで本研究では Tweet データの分類を試みる．

2. 研究概要

2.1 研究範囲

先に述べたように Web サイトの構築には Tweet データをふるい分ける必要がある．本論文ではこのふるい分けに関する部分を今回の研究範囲とし，分類の手法や分類精度について論じていく．

2.2 研究の流れ

本研究は以下の流れで実施した．

1. Tweet データを形態素解析にて形態素に分解する．
2. 形態素データから解析用データを作成する．
3. 誤差逆伝播法を用いて Tweet 内容を「有益」・「無益」に分類する．
4. 分類結果を考察する．

以下，流れ 1 と流れ 2 までの概要をこの章にて記述し，分類とその結果に関しては次章にて記述していく．

2.2.1 Tweet データの形態素解析

本研究では東日本大震災時にやり取りされた Tweet から 1000 サンプルをランダムに抽出し，解析対象として使用した．解析の最初の処理として，Tweet データに形態素解析を適用する．

本研究では形態素解析器として「JUMAN」[黒橋・河原研究室 12] を利用した．JUMAN の辞書には「ドメイン」と「カテゴリ」という項目が用意されており，解析用データの変数はそのドメインとカテゴリの出現頻度を主に用いた．カテゴリとドメインとは言葉についての詳細事項のことである．例えば「コンピュータ」という言葉はカテゴリ「人工物-その他」，ドメイン「科学・技術」というものになる．

2.2.2 解析用データの作成

表 2 と表 3 は本研究で使用したカテゴリとドメインを示したものである．JUMAN に最初から用意されているカテゴリは 1 番から 22 番までの 22 種類，ドメインは 1 番から 13 番までの 13 種類であるが，今回はこれにカテゴリを 3 種類 (23 番から 25 番)，ドメインに 122 種類 (14 番から 135 番) を追加した．追加したものは，どのような Tweet データが情報として有益であるかということを考察した結果である．特に追加ドメインは場所についての事柄であり，情報を活用するには場所情報が必要であろうという仮説に基づいている．実際，場所情報は重要なものであり Tweet データより場所情報を抽出しようとする研究はいくつも行われている [原 13, 榊 14, 六瀬 15]．

解析用のデータはこれら「「カテゴリ」と「ドメイン」の出現頻度」，そして「ツイートの文字数」の大きく分けて 3 変数を用いて解析データとした．ただし，大きく分けて 3 変数であるが，詳細に個数を述べると 160 変数から成っている．具体的には表 2 に記載の 25 個のカテゴリと表 3 より番号 7 の「ドメイン無し」を除いた 134 個のドメイン，そして「Tweet の文字数」の合計 160 変数である．なお，表 3 の注釈 (※ 1) にあるように「XXX」は都道府県名を表しているが，番号 67 から 90 のドメインにおいては北海道から中部地方までと，近畿地方の一部の都道府県名までとなっており，全都道府県名ではないことをここで断っておく．

表 2: カテゴリ一覧

番号	カテゴリ名	番号	カテゴリ名
1	人	16	場所-機能
2	組織・団体	17	場所-その他
3	動物	18	抽象物
4	植物	19	形・模様
5	動物-部位	20	色
6	植物-部位	21	数量
7	人工物-食べ物	22	時間
8	人工物-衣類	23	ハッシュタグ
9	人工物-乗り物	24	リンク
10	人工物-金銭	25	地名:日本
11	人工物-その他		
12	自然物		
13	場所-施設		
14	場所-施設部位		
15	場所-自然		

表 3: ドメイン一覧

番号	ドメイン名	番号	ドメイン名
1	文化・芸術	16 - 60	交通-鉄道-XXX ※ 1
2	レクリエーション	61 - 65	交通-鉄道-XXX ※ 2
3	スポーツ	66	交通-鉄道線
4	健康・医学	67 - 90	交通-鉄道線-XXX ※ 1
5	家庭・暮らし	91 - 135	展示場-XXX ※ 1
6	料理・食事		
7	ドメイン無し		※ 1：都道府県名
8	交通		※ 2：地域名
9	教育・学習		
10	科学・技術		
11	ビジネス		
12	メディア		
13	政治		
14	交通-国道		
15	交通-高速道路		

3. データと分類結果

3.1 解析用データ

前章で述べた一連の作業により解析用のデータが作成されたので，このデータを入力データとする Tweet データの分類器作成を試みる．

本研究では分類手法として誤差逆伝播法を用いる．また，解析用データは，形態素解析を行った 1000Tweet から 600Tweet をランダムに抜き出し，600 サンプルのデータとした．これは，有益 Tweet と無益 Tweet の個数をほぼ同数にした為である．「有益」・「無益」の情報は災害に関する Tweet や Tweet 中に具体的な地名や建物名が含まれているなど，避難情報として活用できると判断したものを「有益」とし，目視にて手作業で割

表 4: 解析用データに採用した変数の一覧

番号	変数名	番号	変数名
1	人	16	文化・芸術
2	組織・団体	17	家庭・暮らし
3	人工物-乗り物	18	料理・食事
4	人工物-その他	19	交通
5	自然物	20	教育・学習
6	場所-施設	21	ビジネス
7	場所-自然	22	メディア
8	場所-機能	23	政治
9	場所-その他	24	交通-鉄道線-東京都
10	抽象物	25	文字数
11	形・模様		
12	数量		
13	時間		
14	地名:日本		
15	リンク		

り当てた．

ネットワークにデータを入力するにあたり，変数をそれぞれ考察した結果，全ツイートにおいて出現頻度が 0 の変数や出現頻度が極端に少ない変数が存在したのでこれを削除した．この作業により解析用データの変数は表 3.1 に示す 25 変数となった．

図 1 は今回使用したネットワークの概要図である．ネットワークは 3 層構造で，入力層のニューロンは 25 個，隠れニューロンは 33 個，出力層のニューロンは 2 個を使用した．今回ネットワークから得たい出力は Tweet が「有益」か「無益」を表すラベルであり本研究ではラベルを 2 値で示している為，出力ニューロンは 2 個である．なお，このラベルの数は 600 データ中，有益が 294 個，無益が 306 個である．

隠れ層のニューロン数は，ニューロン数を 1 個から 100 個まで変化させ，またそれぞれの隠れニューロン数で重みの初期値がランダムに異なる 10 個のネットワークを用いて学習させた結果を総合的に判断し 33 個と決定した．

データは有益・無益のデータからそれぞれ 7 割を学習用データ，2 割を検証用データ，1 割をテストデータとして抽出して使用した．各データの個数と有益・無益の内訳は，学習用データが合計 420 個 (有益 : 206 個，無益 : 214 個)，検証用データが合計 121 個 (有益 : 59 個，無益 : 62 個)，テストデータが合計 59 個 (有益 : 29 個，無益 : 30 個) である．検証用データは過学習を防ぐ目的で採用した早期停止法を行う為に利用した．なお，本研究では誤差逆伝播法による分析を行うに当たり，数値解析ソフトウェア「MATLAB」の「Neural Network Toolbox」を使用した．各層の伝達関数は隠れ層で双曲線正接関数，出力層で標準シグモイド関数を用い，ほかの設定は「Neural Network Toolbox」のデフォルトの状態を用いた．

3.2 分類結果

このデータを誤差逆伝播法にて分類した結果，誤判定率は学習データにおいて 22.38(%)，テストデータにおいて 13.56(%) で分類が出来た (図 3「誤差逆伝播法」)．図 2 は隠れニューロン数が 33 個で，重みの初期値がそれぞれ異なる 10 個のネットワークを作成して学習させた際の誤判定率を示している．なお，ここで学習させた 10 パターンの初期重みが異なるネットワークは，前記の隠れニューロン数を決定した際の 10 パターンのネットワークとは違い，33 個の隠れニューロン数での最良の結果を探索する目的で新たに作成したものである．本研究で

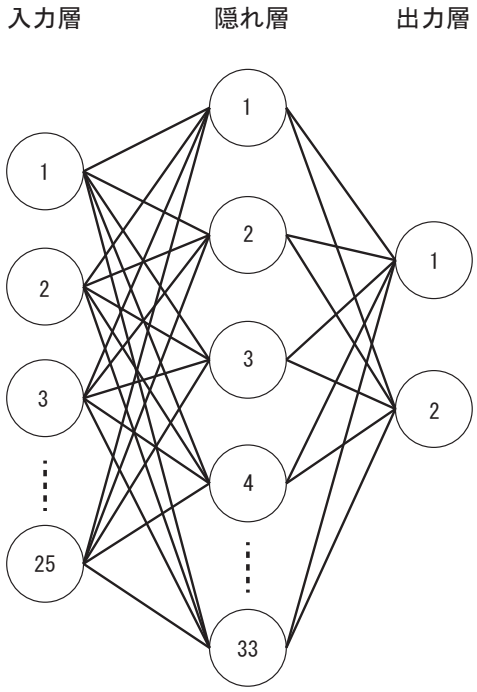

図 1: Tweet データ分類に使用したネットワーク

分類結果として示した誤判定率 (学習データ:22.38(%)，テストデータ:13.56(%)) は図 2 に示す，パターン 10 のネットワークを用いた際に得られたものである．これら 10 パターンの重みの初期値が異なるネットワークを用いた学習における平均誤判定率は，学習データで 27.12(%)，テストデータで 31.69(%) であった．また最大誤判定率と最小誤判定率は学習データで 50.95(%)，15.71(%)，テストデータで 52.54(%)，13.56(%) であった．

ここで誤差逆伝播法による分類精度を考察する為に，他の手法を用いてもこのデータを分類し，分類精度を比較することにする．今回はロジスティック回帰分析を用いた．分類精度は，学習データで 18.81(%)，テストデータで 38.98(%) であった (図 3「ロジスティック回帰分析 (変数番号 24 削除)」)．ただし，この分類結果を得る為にはデータを修正する必要があり，結果を単純に比較することは出来ない．修正した理由は誤差逆伝播法で使用したデータではロジスティック回帰分析にて処理出来なかった為である．回帰の係数を求める為に反復計算を行っており，当初その回数を 100 回として計算したが反復回数内で結果を得ることが出来なかった．この反復回数を 10000 回まで増加させてみたがそれでも収束することはなく，このデータを用いてロジスティック回帰分析にて分類することは出来ないと結論づけた．反復回数 100 回の時点での結果を見たところ，変数 24 のオッズが異常に大きい値であった．そこで変数 24 を削除してみたところ 100 回以内の反復で結果を得る事が出来たので，この結果をロジスティック回帰分析での解析結果とすることにした．

この様にして 2 つの手法により Tweet の分類が出来たわけであるが，これら得られた誤判定率より判断して，誤差逆伝播法による分類結果は高い精度であり Tweet データの分類に有効と判断した．

4. まとめ

Tweet データの分類に対して誤差逆伝播法とロジスティック回帰分析を比較したところ，テストデータの予測においてロジスティック回帰分析よりも誤判定率を 25.42(%) 低く予測することが出来た．これは約 15 個の判定の差であり無視できる差では無い．

くわえて精度以外にも，誤差逆伝播法が，用意した全変数を解析出来たのに対して，ロジスティック回帰分析では変数を削減しなければならなかった．削除した変数は公共交通機関の路線に関するものであり，今回の研究にとって重要な変数であるという仮説で作成した変数である．仮説に沿ったデータをそのまま分析出来るという点でも誤差逆伝播法は有効な分類手法であることが言える．

今回は 600 個のデータに対して分類実験を行ったが，今後はより大きいデータに対して誤差逆伝播法を適応し，最終的に作成を試みている情報提供サイトの構築を目指していく．また今回は分類精度にのみ着目し Tweet の分類に有効な手法について論じたが，今後は具体的な Tweet の分類結果も考察し，Tweet の分類に関する研究を更に進めていきたい．

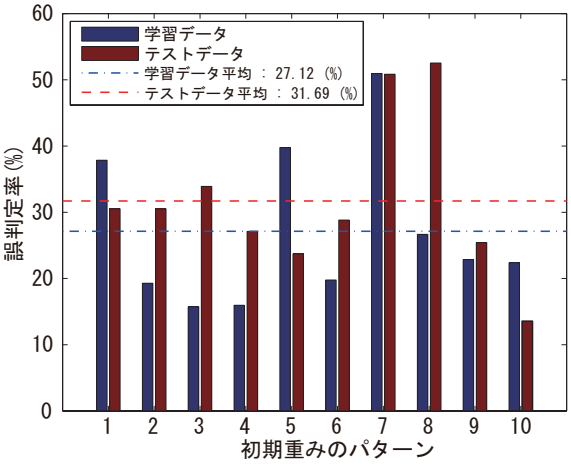

図 2: 隠れニューロン数 33 個のネットワークにおける，重みの初期値が異なる場合での各学習結果と平均値

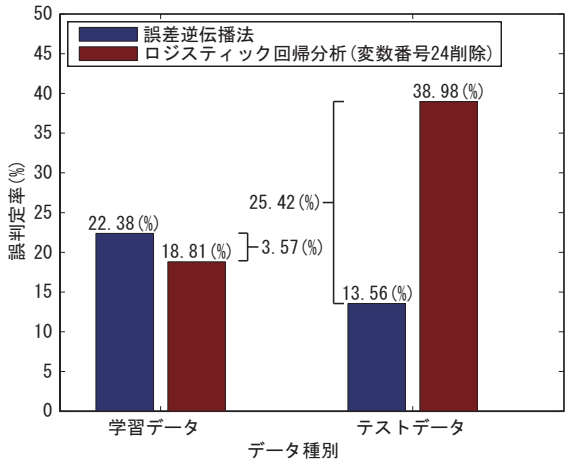

図 3: 分類結果

参考文献

[吉次 11] 吉次 由美 : 東日本大震災に見る大災害時のソーシャルメディアの役割　～ツイッターを中心に～, 放送研究と調査, Vol. 7, pp. 16–23 (2011)

[原 13] 原 久美子, 木野 泰伸, 鳥海 不二夫:字･町名をキーとした災害時 Twitter 情報の抽出と地図への展開　-「どこ」で「何」が起きているのかを知る-, in *The 27th Annual Conference of the Japanese Society for Artificial Intelligence* (2013)

[黒橋・河原研究室 12] 黒橋・河原研究室 京都大学大学院情報学研究科知能情報学専攻知能メディア講座言語メディア分野 : 日本語形態素解析システム JUMAN Ver.7.0, `http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN` (2012)

[榊 14] 榊 剛史, 原 久美子, 吉田 光男, 鳥海 不二夫, 篠田 孝祐, 栗原 聡, 風間 一洋, 野田 五十樹 : 災害情報基盤構築に向けたテキストデータからの地理情報抽出システム, in *The 28th Annual Conference of the Japanese Society for Artificial Intelligence* (2014)

[六瀬 15] 六瀬 聡宏, 内田 理, 富田 誠, 梶田 佳孝, 山本 義郎, 鳥海 不二夫 : 大規模災害時の情報提供を目的とした地名の曖昧性解消, 言語処理学会 第 21 回年次大会 発表論文集 (2015)