機械学習を用いた連絡通知の自動分類に関する研究

黒木 亮人[†1] 高山 拓斗[†1] 成 凱[†1] 安部 恵介[†1]

概要：大学生は在学中に毎日様々な連絡通知を受け取っている。連絡通知には全学生対象、留学生向け、就活生向け、資格取得者向けなど対象が異なる連絡が混在しているため、重要な連絡通知の確認がおろそかになることがある。送信者側からみると、対象者を細かく分類し、ターゲットを絞って送信する方法もあるが、広く周知したくて直接関係のない者にも送信してしまうケースが多々ある。本研究では、機械学習を用いて連絡通知を自動分類し、重要な連絡通知を見逃すことを防ぐことを目的とする。まず、大学から受信したメールデータ 200 件の本文をテキストデータとして抽出し、手動でラベルをつけることで、機械学習の教師データとして用意する。前処理作業としてテキストデータの形態素解析を行い、Bag-of-Words モデルを適用し、単語の出現頻度による重み TF-IDF を計算する。最後に機械学習のアルゴリズムを訓練し、分類器を作成し、テストを行う。今回は特に就職活動中の学生を想定し、連絡通知を「就活関連」と「就活以外」に分類する。作成した分類器は、テストデータによる分類の精度とトレーニングデータを使用した k 分割交差検証によって評価する。その結果、テストデータによる精度は最高 97.8%であって、有効な学習であるといえる結果を示した。

キーワード：機械学習、自動分類、連絡通知、メール

Study on Machine Learning Based College Message Classification

Akito KUROGI[†1] Takuto TAKAYAMA[†1] Kai CHENG[†1] Kesuke ABE[†1]

Abstract: In this paper, we study the issue of text data classification using machine learning for college students overwhelmed by so many messages / notices. To do this, we collect 200 messages and label the plain texts with two categories, namely job-hunting and others. Using this dataset, we train two machine learning algorithms, RandomForest and LogisticRegression. By experiments, we found RandomForest is more suitable for the message classification and when parameter α=0.25, that is when 3/4 of dataset are used as fit data (or training data) and 1/4 are used for test, the accuracy of classification is highest.

1. はじめに

大学生は在学中に毎日様々な連絡通知を受け取っている。これらの連絡通知は在学生・教職員・保護者専用 WEB サイト、メール等の方法にて受け取る。連絡通知には全学生対象、留学生向け、就活生向け、資格取得者向けなど対象が異なる連絡が混在しているため、重要な連絡通知の確認がおろそかになることがある。送信者側からしてみると、配信対象を細かく分類し、ターゲットを絞って送信する方法もあるが、広く周知したくて直接関係のない者にも送信してしまうケースが多々ある。

本研究では、機械学習を用いて連絡通知を自動分類し、重要な連絡通知を見逃すことを防ぐことを目的とする。まず、大学から受信したメールデータ 200 件の本文をテキストデータとして抽出し、手動でラベルをつけることで、機械学習の教師データとして用意する。前処理作業としてテキストデータの形態素解析を行い、Bag-of-Words モデルを適用し、単語の出現頻度による重み TF-IDF を計算する。最後に機械学習のアルゴリズムを訓練し、分類器を作成し、テストを行う。今回は特に就職活動中の学生を想定し、連絡通知を「就活関連」と「就活以外」に分類する。作成した分類器は、テストデータによる分類の精度とトレーニングデータを使用した k 分割交差検証によって評価する。

2. テキストデータの前処理

本研究の対象となる連絡通知はメール本文のテキストのみとする。また、分類結果となるカテゴリは「就活関連」と「就活以外」とする。以下は連絡通知の具体例である。

連絡通知（例）　　【カテゴリ】就活関連

〇〇様へ：弊社の説明選考会は 10 名程度の座談会形式で開催します。他の学生の就職活動の軸は大切にしたい価値観をお互いの自己紹介を交えて、社員とともに語り合う時間を多く用意しています。不動産業界をこれまで見てこなかった人も、是非、この機会にお越しください。

連絡通知（例）　　【カテゴリ】就活以外

学生の皆さんへ：喫煙者の中に、喫煙場所の囲いの外にはみ出て、喫煙区域外で喫煙する姿が散見されます。喫煙する場合は「喫煙場所に指定されている区域内」で喫煙し、吸殻は必ず吸殻入れに始末してください。

また、喫煙場所が混んでいる場合は、喫煙が終わり次第、速やかに待っている方に代わってください。

コンピュータに自然言語（日本語などの）を理解するこ

†1 九州産業大学
Kyushu Sangyo University

とはできないため、自然言語で書かれたテキストデータはそのまま機械学習に使用できない。テキストデータを機械学習に適応させるために行う一連の処理を行う必要がある。これらの処理を総じて前処理という。テキストデータの前処理には行う工程が主に 3 つある。

1) 形態素解析
2) 単語の重み付け
3) データのクレンジング

2.1 形態素解析

形態素解析とは、私たちが普段生活の中で一般的に使っている自然言語を、言葉が意味を持つまとまりの単語の最小単位、いわゆる形態素にまで分割する技術のことである。例えば、「私は大学で経済学を学んでいる。」という文を形態素解析する。「私(代名詞)/は(副助詞)/大学(名詞)/で(副助詞)/経済学(名詞)/を(副助詞)/学んで(動詞)/いる(助動詞)」というように言葉を分割し、辞書などの情報と照らしあわせ、それらの品詞の種類、活用形の種類などを割り出していく。

形態素解析は自然言語処理に欠かせない技術である。例えば、Google などの検索エンジンに入力されたキーワードはそのまま処理されるのではなく、形態素解析によって最小単位にまで分割される。この分割によって検索に必要のない単語を省くことができ、余分なデータ処理をしないですむ。

形態素解析は言語によって難易度が異なる。英語で書かれた文書の場合は、単語が元々分かれているため、単語の分割は非常に容易である。しかし、日本語の場合、単語同士が助詞によってつながっており形態素解析が非常に複雑なものとなる。そのため形態素解析ソフトを使用することが必須である。以下に、有名な形態素解析ソフトについて紹介する。

MeCab はオープンソースの日本語形態素解析エンジンである。言語や辞書、またデータベース化された言語資料であるコーパスに依存しない、汎用的な設計が MeCab の特徴である。名前の由来は「和布蕪(めかぶ)」からきている.

JUMAN は京都大学大学院の黒橋・河原研究室が開発した形態素解析ツールである。特徴として、文字コードが UTF-8 に対応している点、また WEB テキストから自動獲得された辞書、Wikipedia から抽出された辞書を使用できる点が挙げられる。また、JUMAN は MeCab よりも単語の意味分類を細かく実施される。

Janome は Python で書かれている形態素解析ツールである。名前の由来は「蛇の目」からきている。特徴として、Python のみで書かれているため、Python 上での利用が非常に簡単であることが挙げられる。

2.2 単語の重み付け

単語の重み付けは、テキストデータを数値のベクトルに変換する工程である。形態素解析で得られた単語の集まりを対象としてベクトル化していく。ここで一般的に採用されるのは、テキスト文書の文脈を無視し、単語の出現頻度のみに注目する BOW (Bag of Words)モデルである。
重み付けによく採用されるのは TF-IDF 重みである。TF-IDF とはテキストデータ内の単語の重要度を評価する手法である。tfidf は $\mathrm{tfidf} = \mathrm{tf}(t, d) \times \mathrm{idf}(t, d)$ と表す。

TF (Term Frequency)は単語の出現頻度を表す項目である。以下は TF の定義である。

$$\mathrm{tf}(t, d) = \frac{\text{文書 d における単語 t の出現頻度}}{\textit{文書d における全単語の出現頻度の和}}$$

すなわち、文書 d において単語 t がどのくらい出現したかである。

IDF (Inverse Document Frequency)は逆文書頻度を表す項目である。逆文書頻度とは単語がレアなものなら高い値を、色々な文書によく出現する単語なら低い値をとるものである。

$$\mathrm{idf}(t, d) = \log \frac{\textit{全文書数} n}{1 + \textit{単語} t \textit{を含む文書数} df(t)}$$

この数式の分母に 1 を足しているのは、トレーニングデータに出現するすべての単語に 0 以外の値を割り当てて、ゼロ割を回避するためである。IDF では多くの文書に存在しうる頻出単語の重みを減らし、その文書にしかない重要な単語の重みを増やすことで、分類の機械学習に適したデータを作成することができる。

2.3 データのクレンジング

上記の処理で得られたデータはまだ機械学習の教師データとして使えない。その原因は自然言語には多くの分類の邪魔となるノイズがあるためである。このノイズとは例えばテキスト中の「～は」などの非常に頻出する言葉であったり、メールなどに記されているメールアドレス、WEB サイトの URL など目的とする分類との関連性が薄かったり、そもそも自然言語でないものである。

分類の精度を高めるために、テキストデータをクレンジング(洗浄)することが重要である。多くの場合、ストップワードを指定することで取り除く手法がとられている。ストップワードとは特定の単語を指定すうことで、その単語が文書中に出てきてもそれを削除し、不要な単語を考慮せずに処理を行うことができる。

3. 機械学習アルゴリズム

本研究で使う機械学習アルゴリズムを説明する。まず、産業界において広く使用されている分類アルゴリズムの一つであるロジスティック回帰について紹介する。また、非線形分類にも適しているランダムフォレスト分類モデルについて説明する。

3.1 ロジスティック回帰

ロジスティック回帰分析では、与えられた説明変数の値の組に対する、ある現象の発生する条件付確率をロジスティックス回帰式としてモデル化する。ロジスティック回帰

は分類モデルとして、非常に実装しやすいものの、線形分離可能なクラスに対してのみではあるが高い性能が発揮される。

ロジスティック回帰の概念を理解するために、まずオッズ比から見ていこう。オッズ比は事象の起こりやすさを表すもので、$\frac{p}{1-p}$と書くことができる。この場合、p は正事象の確率を表す。正事象は必ずしも良いことを意味するわけでなく、予測したい事象を表す。その場合はロジット関数を定義できる。この関数は単にオッズ比の対数となる。

$$\text{logit(p)} = log\frac{p}{1-p}$$

ロジット関数は、0 よりも大きく 1 よりも小さい範囲の入力値を受け取り、実数の全範囲の値に変換する。この関数を使って、特徴量の値と対数オッズの間の線形関係を表すことができる。

$$\text{logit}\big(\text{p(y = 1|x)}\big) = \sum_{i=0}^{m} w_i\, x_i = w^T x$$

ここで p(y=1|x)は、特徴量 x が与えられた場合にサンプルがクラス 1 に属するという条件付き確率を表す。

ここで実際に重要なのはサンプルが特定のクラスに属している確率を予測することである。これはロジット関数の逆関数であり、ロジスティック関数とも呼ばれる。特徴的な S 字を描くことからシグモイド関数と呼ばれることもある。

$$\Phi\,(\text{z}) = \frac{1}{1+e^{-z}}$$

この場合の z は総入力を表す。つまり重みとサンプルの特徴量の線形結合であり、以下のように計算できる。

$$\text{z} = w^T x$$

シグモイド関数を図に表すと図 1 のようになる。

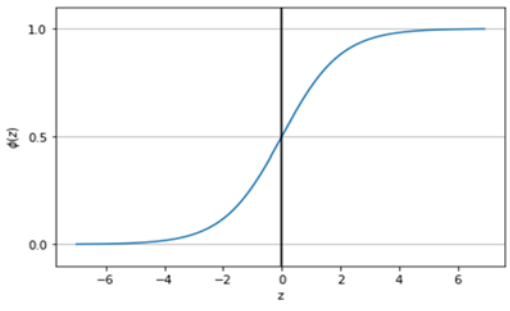

図 1 シグモイド関数

ロジスティック回帰では、活性化関数が先ほどのシグモイド関数となり、入力から出力までを図にすると図 4 のようになる。

特徴量 x が重み w でパラメータ化されるとすれば、このシグモイド関数の出力は、サンプルがクラス 1 に属している確率 Φ(z)=P(y=1|x;w)であると考えられる。例えば、あるサンプルに対して Φ(z)=0.8 が算出される場合は、このサンプルが正事象である確率が 80%であることを意味する。また、逆の自走である確率は 20%である。あとは、量子化器 (単位ステップ関数)を使って予測された確率を二値の結果に変換する。

多くの場合、予測されるクラスラベルに関心があるだけでなく、クラスの所属関係の確率を見積もることも重要である。例えば気象予報において雨が降るかどうかだけでなく、降水確率も発表するためにロジスティック回帰が使用される。

3.2 ランダムフォレスト

アンサンプル学習は、弱学習機と呼ばれるあまり精度の高くない学習器を複数用いて、その結果を組み合わせて、精度向上を図る機械学習法である。アンサンプル学習では、異なるサンプルから単純なモデルを複数生成し、それらを統合することにより、全体としての精度を実現す津モデルを構築する。アンサンプル学習の中には、与えられたデータセットから、ブットストラップによって、複数の学習データセットをサンプルとして生成し回帰・分類を行うバッギング(Bagging)やランダムフォレスト(Random Forest)などの学習法がある。

ランダムフォレストは、複数の木(tree)によって構成される機械学習アルゴリズムである。ここでの木は、決定木のことで、それぞれの決定木の性能はあまり高くなく、それらを複数組み合わせることにより、高い予測精度を持つ学習器となる。ランダムフォレストでは、決定木として二分決定木が主に用いられ、ランダムフォレストのアルゴリズムを以下に示す。

1) 与えられたデータセットから n 個のブートストラップ・サンプル $B_1, B_2, \ldots, B_n$ を作成する。ただし、構築したモデルを評価するために 1/3 のデータを除いてサンプルする。除いたデータを OOB (out of bag)データと呼ぶ。
2) B_k (k=1,2,…,n) における M 個の変数の中から m 個の変数をランダムサンプリングする。M は、データセット中の変数の数を表し、m は、m=root(M)が多く用いられる。
3) ブートストラップ・サンプル B_k の m 個の変数を用いて、未剪定の最大の決定木 T_k を生成する。
4) n 個のブートストラップ・サンプル Bk の決定木 Tk について、OOB データを用いてテストを行い、推測誤差を求める。
5) その結果を統合し、新たな分類器を構築する。分類の問題では多数決をとる

4. 実験評価

4.1 データセット

今回実験で使用されるデータセットは、令和元年度本学学生向けに配信された連絡通知 200 件である。うち、就職関連 160 件、就活以外 40 程度である。

形態素解析解析器に MeCab を使用し、新語・固有表現に強い mecab-ipadic-NEologd 辞書を使って形態素解析を行っている。mecab-ipadic-NEologd は形態素解析エンジン MeCab と共に使う単語分かち書き辞書で、週 2 回以上新更新され、新語・固有表現に強く、語彙数が多く、しかもオープンソース・ソフトウェアである という特徴がある。今回の実験では対象となる品詞は以下に限定する。

- 一般名詞　例：学校、会社、情報
- 固有名詞　例：九産大、福岡、リクナビ、安倍晋三
- サ変接続名詞　例：説明、求人、連絡、選考
- 副詞可能名詞　例：以下
- 形容動詞語幹　例：多数、詳細、無関係

4.2 評価方法

本実験では、単一のデータセットをランダムに 2 分割し、一方のデータセット（フィットデータ）を用いて判別モデルを構築し、もう一方のデータセット（テストデータ）を用いてその判別精度を求める交差検証を行う。分割におけるテストデータの割合を α と表す。

また、k 分割交差検証(k-fold cross-validation)という交差検証法で評価している。k 分割交差検証では、非復元抽出(例えば、2 回おみくじをひくとき 1 回目にひいたおみくじを 2 回目にひくとき戻さないおみくじの引き方をいう)を用いて、トレーニングデータをランダムに k 個に分割する。そのうちの k−1 個をモデルのトレーニングに使用し、1 個をテストに使用する。この手順を k 回繰り返すことで k 個のモデルと性能評価を取得する。

次に、個々のサブセットに基づいてモデルの平均性能を計算する。これは、トレーニングデータの再分割に大きく左右されない性能評価を取得するためである。一般に、k 分割交差検証はモデルのチューニングに利用される。つまり、満足のいく汎化し性能が得られる最適なハイパーパラメータ値を見つけ出すために使用される。つまり、満足のいく汎化性能が得られる最適なハイパーパラメータ値を見つけ出すために使用される。満足のいくハイパーパラメータ値が見つかったら、トレーニングデータセット全体でモデルを再びトレーニングし、トレーニングセットからは独立したテストデータを使って最終的な性能評価を得ることができる。

例えば、k=10 としたときトレーニングデータは 10 個に分割される。そのうち 9 個がトレーニングに使用され、残りの 1 個がモデルを評価するためのテストデータとして使用される。このとき得られた性能をE_iとする。これを 10 回くりかえし次の E を計算する。k 分割交差検証で標準的に用いられる k の値は 10 である。また、ここの E がこの分類器の性能となる。

$$\mathrm{E} = \frac{1}{10}\sum_{i=1}^{10} E_i$$

分割数を増やすとともに処理負荷が高くなる。本研究のように比較的小さなトレーニングデータを扱っている場合は分割数を 5 にしてより正確な性能評価ができることを確認した。

4.3 実験結果

テストデータの割合 α を 0.10 から 0.05 ずつ増やしながら、ランダムフォレスト（RandomForest）とロジスティック回帰の分類モデルを訓練し、予測精度を評価する。結果は図 2 でしている。横軸はテストデータの割合（α）、縦軸は交差検証で得られた分類制度の平均値である。

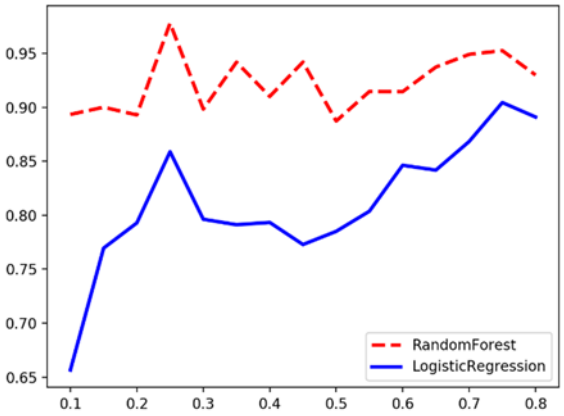

図 2　交差検証による分類精度の評価結果

実験結果から、ランダムフォレストはロジスティック回帰分析より、高い分類精度が得られることが分かった。また、テストデータの割合が 0.25 のとき、つまり、3/4 をフィットデータとして 1/4 をテストデータとする場合に最も高い予測精度が得られることもわかった。

5. おわりに

本研究では大学からの重要連絡通知を見逃すことを防ぐため、機械学習を用いた連絡通知の分類器を作成することを試みた。今回は特に就職活動中の学生を想定し、連絡通知を「就活関連」と「就活以外」に分類する。また、分類器にランダムフォレストとロジスティック回帰分析を使用したが、ランダムフォレストの方が今回の問題に適していることが分かった。

参考文献

[1] 佐藤直, 渡邊隆志, 類似度を利用した迷惑メールフィルタリング, 第 77 回全国大会講演論文集, pp.455 - 456, 2015 年
[2] 杉井学, 松野浩嗣, 機械学習によるスパムメールの特徴の決定木表現, 情報処理学会研究報告. マルチメディア通信と分散処理研究会報告, pp.183−188, 2007 年