Vol.2020-CE-154 No.6
2020/3/14

高校生のTwitterアカウント属性を機械学習で予測する手法の提案と評価

三村　徹[*]　辰己　丈夫[2]

概要：Twitter などの匿名で利用できる SNS は、年齢、性別、職業等の属性を隠すだけでなく、偽って利用することが可能であり、悪意をもったユーザと接触するリスクがある。そのため、危険な SNS 利用が高校生によるものである場合には、ユーザの所属を推定し、学校等で安全な利用を促す指導をする必要がある。

本研究では、Twitter の公開アカウントのうち、職業属性が高校生であるアカウントを予測することを目的とする。公開されているツイートの自然言語処理を行い、機械学習のアルゴリズムを用いて分析を行った。言語をベクトルに変換する手法において複数の方法を試行し、最大で正解率 0.75、F 値 0.75 の組合せを発見した。

キーワード：Twitter、SNS、属性推定、自然言語処理、機械学習、SVM、ネットパトロール

1. 背景と目的

1.1 SNS に関連する犯罪等の状況

2019 年の大阪家出少女誘拐事件では、小学生の被害者が Twitter 上で家出願望の情報を発信していたことから、犯罪のターゲットになったと考えられる。また、2017 年の座間 9 遺体事件では、被害者のうち 3 名が高校生であり、Twitter で自殺願望の情報を発信していたことから、犯罪のターゲットになったと考えられる。このように、Twitter 等の SNS には、事件に発展し得る情報が発信されているにも関わらず、このような事件を未然に防止することができなかった。

1.2 ネットパトロールの現状

多くの自治体においては、ネットパトロールと称して、犯罪、自殺、いじめを始め、問題のあるインターネット上の高校生等の投稿を監視している。東京都では、問題のある SNS の投稿を含め、インターネット上の投稿を年間約 7,000 件検知し、うち 40 件程度を何らかのリスクがあるとして学校等に情報提供している [1]。他の広域自治体においても、約 68 %道府県で何らかのネットパトロールを実施している。

多くの自治体での監視方法は、検索サービスで所管する学校の学校名等を入力する人海戦術であり、リアルタイムに、かつ網羅的に検知することができない。また、所管する複数の学校を、何日か置きにスケジューリングしてパトロールする場合もある。一方、リアルタイム性を重視して、クローラを用いて問題行動に関する単語で検索した場合では、アカウントの属性、特に、所属校が判別不能で、検知後に関係機関に通報することが困難である。

近年、リアルタイムかつ網羅的な検知が困難であるにも関わらずコストが高いため、納税者への説明が求められるところであり、事業の継続を見直す自治体も出ている。このため、低コストで一定の成果が期待できる技術が求められている。

1.3 高校での情報モラル教育

高校生による Twitter の情報発信では、犯罪の加害者にも被害者にもなり得る危険な情報発信をしばしば見かける。教育的な視点からは、高校生は、情報モラル教育を必要とする情報発信者とも言うことができるが、所属校や人物を特定することができないため、必要な指導を受ける機会を逃している。もし、人物を特定するに至らなくても、ある程度の集団に絞り込むことができれば、その集団全員を対象に必要な指導を行うことが可能になる。

著者の主観では、学校の教員が実施する情報モラル教育において、事前の計画による指導ではなく機会を捉えて行う指導（生徒指導）として情報モラル教育を実施する場合のきっかけは、発生した事件の大きさが大きいほど、また、事件に関与している集団を絞り込めているほど、高い動機

[*]放送大学大学院修士課程
[2]放送大学

となる。

1.4 本研究の目的

本研究では、都立高校生と思われる Twitter アカウントについて、所属校を推定することを目的とする。推定は確信度付きで行えるようにし、学校設置者である自治体が特別指導等をするかどうかを実際に判断し得る情報となることを目標とする。

ネットパトロールの完全自動化を実現するためには、高精度で判定することが必要である。しかし、高校生のツイートに対し、統計学的に類似している順に並べるだけでも、マニュアル作業の効率化に寄与することができる。

本研究が更に発展し、都立高校生での所属校推定までが可能になれば、特定の学校の中学生・高校生の所属校推定も可能になると考えられる。このような技術は、治安対策や自殺対策を始め、中等教育段階の生徒に対する情報モラル教育を補完する特別指導のための診断を、各自治体が低コストで実現する技術となり得るものである。

2. SNS の属性推定技術の動向

2.1 自然言語処理による属性推定

Twitter ユーザは、プロフィール及びツイートに自己の属性を示唆する様々な情報を自然言語で投稿する。投稿内容を分析するには、自然言語で記述された投稿を形態素解析し、出現する単語に重みを付けた特徴量を SVM で分析したり、現実世界の時空間・事物に関する辞書を作成し、紐づけたりし、属性値ごとに分類する手法がある [3]。

2.2 周辺ユーザ情報による属性推定

Twitter ユーザは、フォロー関係によりソーシャルグラフを形成している。コミュニティ検出とは、グラフにおいてノード間が密に接続しているノードの集合部分を発見する問題である。Twitter ユーザのソーシャルグラフでは、ユーザの属性に共通の特徴があり、属性値のクラスタが存在する。そのクラスタは特定のコミュニティであると考えられ、それらを抽出する手法が提案されている [4]。

また、メンション（@user の形式で特定ユーザに言及した投稿がフォロワー全員に届く）で形成されたソーシャルグラフを確認でき、そこで抽出されたコミュニティにおいて、プロフィール及びツイートを自然言語処理して精度を高める手法が提案されている [5]。

2.3 位置情報による属性推定

ジオタグとは、GPS から取得した緯度・経度の位置情報であり、Twitter のツイートには、ユーザの投稿場所を表すジオタグが付与されるものがある。ジオタグをクラスタリングし、ユーザが頻繁に訪れる場所での投稿割合を求め、職業を推定する手法がある [6]。

3. 研究手法

3.1 研究手法の概要

本研究は、分類対象とする Twitter アカウントに対し、その所属属性が高校生かどうかを、分類器により予測する問題である。その手順の概要は、次のとおりである（図 1）。

(1) 学習データ・評価データ用に、ツイートデータを収集する。
(2) ツイートデータを自然言語処理し、素性ベクトル（特徴量）を定義する。
(3) 学習データから機械学習モデルを構築し、分析器・分類器を生成する。
(4) 生成した分析器・分類器で、評価データを用いて検証する。

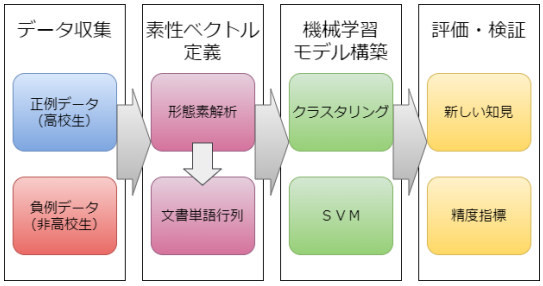

図 1 研究手法の概要

3.2 ツイートデータの収集

正例（高校生アカウント）の学習データは、教師あり学習の教師データ、教師なし学習の学習データ、正例の評価データとして使用するものである。石野 [7] によると、ある大学と関連度の深いアカウントのフォロワーから、その大学の学生を SVM で自動検出することが可能である。著者は、次の方法により、2019 年 8 月から 2019 年 11 月にかけて、都立高校を所属属性とする公開アカウント 1801 件を取得した。

(1) TwitterAPI により、高校生のフォロワーを多くもつ公式アカウントについて、フォロワー一覧を取得する。
(2) フォロワーのうち、公開アカウントで（Protected 属性が False）、スマホで利用し（Client 属性が Android か iPhone）、利用開始年が 2016 年以降（中学 3 年生以降にアカウントが作られたと想定）、1 回以上ツイートしている、の条件で抽出する。
(3) 抽出されたフォロワー一覧について、プロフィール及び直近のツイートを著者が目視で確認し、ほぼ確実に高校生であると推定されるアカウントを正例（高校生アカウント）とする。

こうして得た 1801 件のアカウントに対し、2019 年 11 月 16 日から 2019 年 11 月 17 日にかけて、TwitterAPI により直近 200 件のツイートを一斉に取得した。1801 件のア

カウントのうち、155 回以上ツイートしているユーザ（週１回以上ツイートしているユーザ）をアクティブユーザとみなし、545 名分を正例の学習データとした。

東京都 [2] の 2018 年の調査によると、東京都の中学生の Twitter 利用率 32.5%に対し、東京都の高校生の Twitter 利用率 72.4%である。Twitter 利用率は、中学 1 年生から段階的に増加すると推定され、中学 3 年生以降に作成されたアカウントを抽出することが効率的であると考えた。

負例（非高校生アカウント）のデータは、教師あり学習の教師データや、負例の評価データとして使用するものである。著者は、次の方法により、2019 年 11 月 30 日に、非高校生と見なす公開アカウント 612 件を取得した。

(1) TwitterAPI により、適当なアカウントを起点としてフォロワー一覧を取得する。
(2) フォロワーのうち、公開アカウントで（Protected 属性が False)、スマホで利用し（Client 属性が Android か iPhone)、利用開始年が 2015 年以前の条件で抽出する。
(3) 抽出されたフォロワー一覧について、さらにそれぞれのフォロワー一覧を取得する。
(4) 上記と同様の抽出を行い、負例（非高校生アカウント）とする。

こうして得た 612 件のアカウントに対し、2019 年 11 月 30 日に、TwitterAPI により直近 200 件のツイートを一斉に取得した。612 件のアカウントのうち、正例の評価データと同様に、155 回以上ツイートしているユーザ 256 名分を負例の学習データとした。

3.3 素性ベクトル（特徴量）の抽出１：形態素解析

機械学習に使用するデータは、数値（ベクトル形式）として表現する必要がある。一方、Twitter から得られるデータは、添付された画像や、投稿位置を表すジオタグを除き、スクリーンネーム、プロフィール、ツイートなど、ほとんどがテキスト形式である。テキスト形式のデータをベクトルに変換するには、テキストに表れる各単語の出現状況を特徴量とする方法がある。

はじめに、学習データ、評価データともに、日本語形態素解析システム MeCab[8] を使用し、形態素（単語）に分割する。MeCab で解析された形態素には、次のような品詞情報が含まれる。

表 1 形態素の品詞情報

感動詞	記号	形容詞	助詞
助動詞	接続詞	接頭詞	動詞
副詞	名詞	連体詞	

今回の研究では、名詞、動詞、副詞の 3 パターンにおいて、素性ベクトルを作成して検証を行うこととする。話題のトピック（トレンド）の調査を行う場合には、名詞に限定して素性ベクトルを作成する方法が主流である。しかし、著者は、動詞や副詞にも、高校生の特徴となり得るテキストが含まれていると考えている。

テキスト分析では、多くの場合でストップワード（Stop Word）を取り除く。ストップワードとは、英語の「the」、「of」、「to」などのように、あまりに多く出現するために分析にの役に立たない単語である。英語を始め、ヨーロッパ系言語のテキスト分析用のストップワードリストは、複数の研究者により GitHub 等で提供され、利用することが可能となっている。一方、MeCab による日本語の形態素解析においては、前述のとおり品詞による抽出が可能であり、分析の役に立たない「助詞」、「助動詞」、「接続詞」等をある程度取り除くことが可能である。

3.4 素性ベクトル（特徴量）の抽出２：文書単語行列

次に、文書単語行列を作成する。文書単語行列とは、文書を行にとり、単語を列にとった行列で、各行が、その文書の素性ベクトルとなっている。

表 2 文書単語行列

	単語 1	単語 2	単語 3	単語 4	単語 5	...
文書 A	·	·	·	·	·	
文書 B	·	·	·	·	·	
文書 C	·	·	·	·	·	
文書 D	·	·	·	·	·	
文書 E	·	·	·	·	·	
⋮						

素性ベクトルの要素のとりかたには、次のような方法がある。

- BoW：BoW（Bag-of-Words）とは、ある文書における、その単語の出現の有無である。BoW では、文書における単語の出現順序は考慮せず、出現の有無のみを値（0,1）とする。
- TF：TF（Term Frequency）とは、単語出現回数である。「ある文書において、その単語が何回出現したのか」を表し、次の式により定義される。

$$tf = \text{文書 A における単語 X の出現回数}$$

- IDF：IDF（Inverse Document Frequency）は、逆文書頻度である。単語が他の文書に出現しないほど高い値に、他の文書にも頻繁に出現するほど低い値になり、次の式により定義される。

$$idf = \log \frac{\text{全文書数}}{\text{単語 X を含む文書数}}$$

- TF-IDF：TF-IDF とは、TF と IDF の乗算で定義される。よって、単語出現回数と逆文書頻度の両方に比例し、出現回数が多く、かつ、他の文書に出現しない単

語を、その文書を特徴づける単語として捉えている。

今回の研究では、BoW、TF、TF-IDF の 3 パターンについて文書単語行列を作成し、検証を行うこととした。

また、文書単語行列は、文書数が増えるに連れて、出現単語数が大きくなる。著者が、Twitter ユーザの直近 200 ツイートを 1 つの文書として、文書数と文書単語行列の規模の関係を調査したところ、行列の規模は次のようになった。

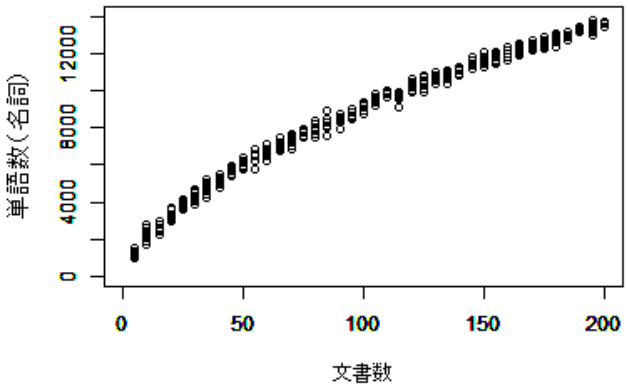

図 2　文書単語行列の規模（名詞）

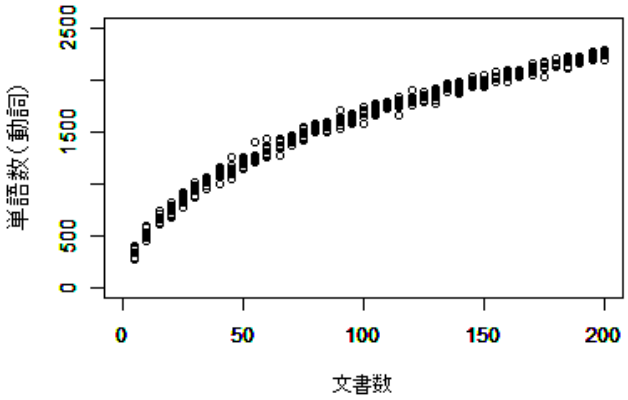

図 3　文書単語行列の規模（動詞）

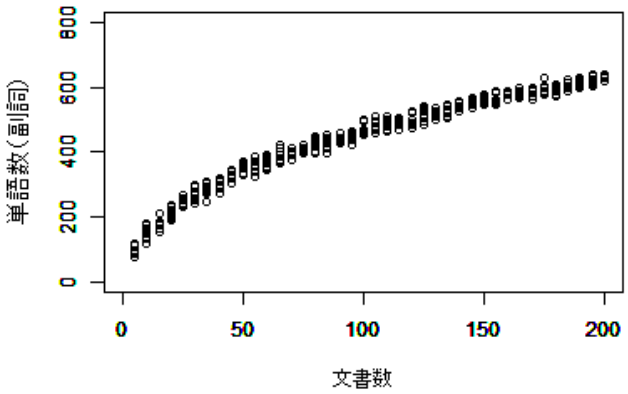

図 4　文書単語行列の規模（副詞）

ある程度十分な文書数をとった場合、文書単語行列の拡大率が収束気味になることがわかる。このことから、十分な文書数を学習データとして分類器を学習させた場合、判定対象文書に表れる単語との間に一定程度の重複が見込まれ、特徴の比較が可能であると考えられる。

3.5　クラスタリングによる分類方法

クラスタリングとは、分類対象の集合を、ベクトル相互の距離に応じて、いくつかの部分集合（クラスタ）に分割することである。クラスタリングの手法には、階層的手法と非階層的手法の 2 種類がある。

階層的手法では、まず、1 個だけの要素を含むクラスタを初期状態とする。そして、クラスタ相互の距離の近い 2 つのクラスタを逐次的に併合していく。最終的に全クラスタが 1 つのクラスタに併合されるまで繰り返すことで、ボトムアップでクラスタが形成される。クラススタリングの結果は、次のようなデンドログラムによって表される。

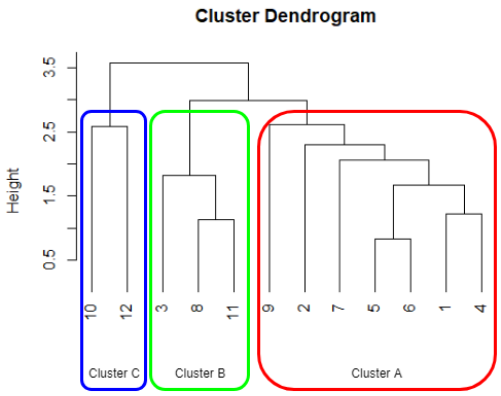

図 5　デンドログラムの例

一方、非階層的手法では、先に分割数 N を決定し、分割のよさを表す評価関数を最適にする分割を探索する方法である。

今回の研究では、正例データに対し、階層的手法でクラスタリングを行い、新しい知見の発見を目指す。後に述べる OneClassSVM による分類の精度向上のためのフィードバックが得られることを期待する。階層的手法の詳細は、Web 等により解説が容易に入手できるため本稿では省略するが、距離の定義にユークリッド距離を、距離関数には最短距離法を用いることとする。

3.6　OneClassSVM による予測方法

OneClassSVM とは、機械学習の分類アルゴリズムである SVM（Support Vector Machine）を、教師なしの 1 クラス分類に応用した手法である。正常データとして 1 つのクラスを学習させて識別境界を決定することで、その境界

を基準に外れ値を検出する。異常がほとんど発生せず、異常クラスのデータが集まらないようなシステムにおいて、異常検知を実現したい場合に有効な手法であると考えられている。

今回の研究では、正例データについては、高い精度と十分な量を確保することができている。しかし、負例データについては、正例の補集合となるデータを収集したため、特徴が多様になっていると考えられる。従って、正例データの特徴を境界内とする OneClassSVM が分類に適していると考えた。

4. 検証結果

4.1 性能評価の方法

検証結果の前に、分類器の性能評価の方法を説明する。評価データには、正・負の 2 クラスのデータがある。また、予測結果も同様に、正・負の 2 クラスのデータが存在する。評価データをどのように予測したかは、次の表のような混同行列により表すことができる。

表 3 2 クラス分類の混同行列

		分類器が予測したクラス	
		正（Positive）	負（Negative）
実際の	正例	TP（TruePositive）	FP（FalsePositive）
クラス	負例	FN（FalseNegative）	TN（TrueNegative）

その上で、代表的な評価指標として、次のものがある。

- 正解率（正確度、Accuracy）: 全体のデータの中で、正しく予測できた TP と TN の割合を、次の式で定義している。高いほど性能が良い。

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

- 適合率（精度、陽性反応的中度、Precision）: Positive と予測したデータの中で、真に Positive だったデータの割合を、次の式で定義している。高いほど性能が良く、誤りが少ないことを意味する。

$$Precision = \frac{TP}{TP + FP}$$

- 再現率（感度、真陽性率、Recall）: 実際に Positive なデータを、正しく Positive と予測できている割合を、次の式で定義している。高いほど性能が良く、Positive と予測すべきデータの回収率を意味する。

$$Recall = \frac{TP}{TP + FN}$$

- F 値（F-score、F-measure）: 適合率と再現率の調和平均を、次の式で定義している。適合率と再現率はトレードオフの関係にあり、高いほど性能とバランスが良い。

$$Fscore = \frac{2 * Precision * Recall}{Precision + Recall}$$

今回の研究では、正例と負例 20 件ずつ、計 40 件のデータで評価を行い、正解率を第 1 の評価指標にすることとした。また、適合率と再現率はトレードオフの関係にあるため、これらの調和平均を用いた F 値を第 2 の評価指標とした。

実用化する際には、検知する投稿内容との組み合わせ方に応じて、重視する評価指標を変えることも考えられる。例えば、自殺などの重大な案件に関しての投稿者属性を推定する場合には、再現率を重視して、負例を正例と判定する誤りを許容し、正例の漏れを少ないすることが考えれる。逆に、軽易な案件に関しての投稿者属性を推定する場合には、適合率を重視して、正例を負例と予測する誤りを許容し、正例としての確実性の高いものを検出することが考えられる。

4.2 文書単語行列

クラスタリング及び OneClassSVM で分析・分類する過程において、ベクトルに変換するための数値を介する必要がある。その数値の全体像を把握するのに、文書単語行列を見ればよいと考えた。

形態素の違いによる文書単語行列には、以下のとおりの特徴が見られた。

- 名詞では、特定のサードパーティサービスを利用した際に使われる単語が頻出している [表 4]。
- 動詞では、他の動詞と組み合わせて用いる補助動詞が頻出している [表 5]。
- 副詞では、口語体を平仮名表記したものが頻出している [表 6]。

表 4 TF での文書単語行列の一部（名詞）

	ave	R.1	R.2	R.3	R.4
質問	49.09	0	0	236	8
箱	31.44	0	0	142	2
笑	20.08	31	3	25	45
ん	12.15	23	17	10	12
中	10.53	1	2	49	4
の	9.63	9	8	4	7
ー	9.07	5	5	10	12
募集	9.02	0	0	46	4
匿名	8.97	0	0	46	4
人	8.75	9	8	28	8
こと	6.74	4	3	4	2
みんな	6.44	5	2	25	2
好き	4.95	1	3	25	2
最近	4.94	1	1	26	2
それ	4.51	1	2	2	1
私	4.12	2	12	3	6
数	4.10	2	1	0	0
今日	4.07	2	4	1	0
¥”	4.00	4	4	4	4
何	3.95	2	6	4	5

表 5　TF-IDF での文書単語行列の一部（動詞）

	ave	R.1	R.2	R.3	R.4
する	34.52	30.18	38.23	55.33	19.11
てる	25.28	25.45	43.78	69.23	14.25
答える	11.22	0.00	2.37	54.46	9.47
いる	10.09	16.58	9.33	16.58	9.33
なる	9.18	6.33	9.50	16.88	5.28
ある	8.46	6.33	10.55	7.39	5.28
れる	8.17	6.48	14.04	4.32	6.48
思う	7.49	3.24	8.64	3.24	2.16
比す	6.89	0.00	0.00	0.00	0.00
言う	6.22	5.63	10.13	6.75	4.50
やる	6.12	8.06	11.52	5.76	0.00
行く	5.66	2.32	12.75	2.32	0.00
すぎる	5.24	7.16	15.50	0.00	1.19
受け取る	5.07	0.00	0.00	0.00	0.00
見る	4.89	3.60	4.80	0.00	2.40
くる	4.77	5.43	6.52	0.00	3.26
わかる	3.88	3.99	1.33	0.00	0.00
くれる	3.66	2.63	5.26	0.00	0.00
くださる	3.32	10.54	0.00	0.00	3.51
できる	3.28	6.72	1.34	34.96	6.72

表 6　Bow での文書単語行列の一部（副詞）

	ave	R.1	R.2	R.3	R.4
そう	0.73	1	0	1	1
もう	0.73	1	1	0	0
どう	0.65	1	1	1	1
めっちゃ	0.62	1	0	0	1
まだ	0.61	1	1	0	1
ちょっと	0.48	1	1	1	1
ちゃんと	0.47	1	1	0	1
よく	0.47	0	0	0	1
全然	0.45	1	1	0	1
ほんとに	0.42	1	0	0	0
ずっと	0.40	0	1	1	0
多分	0.40	1	0	0	0
いつも	0.40	1	0	1	0
もっと	0.40	1	1	1	0
なんで	0.39	0	1	1	0
これから	0.38	0	0	0	0
とりあえず	0.38	1	0	0	1
本当に	0.35	1	1	0	0
さすが	0.32	0	0	0	0
やっぱり	0.32	1	1	0	0
よろしく	0.32	0	0	0	0

また、各単語の出現状況の数値化手法の違いによる文書単語行列には、以下のとおりの特徴が見られた。

- TF では、最頻出単語の数値が極端に大きく、文書間の分散も大きくなっている。[表 4]。
- TF-IDF では、TF ほど分散は大きくないが、ほぼ出現回数の影響により数値が大きくなっている [表 5]。
- BoW では、行列の大部分を、0 の要素が占める [表 6]。

4.3　クラスタリングによる分類

各ベクトル間のユークリッド距離を、最も近いものを統合していく最短距離法で、次々にクラスタを形成していく階層型クラスタリングを行った。各ベクトル間のユークリッド距離は、ベクトルの最大要素の影響を大きく受け、次のようなデンドログラムとなった。

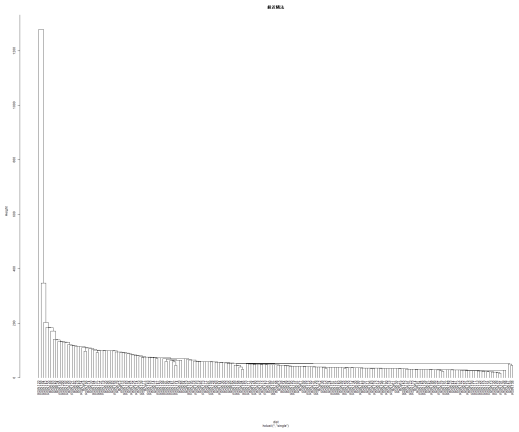

図 6　TF でのデンドログラム（名詞）

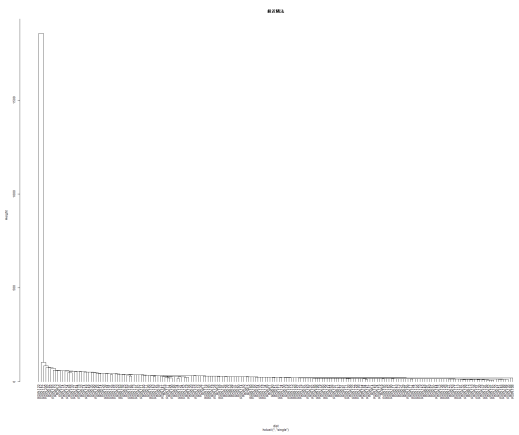

図 7　TF-IDF でのデンドログラム（動詞）

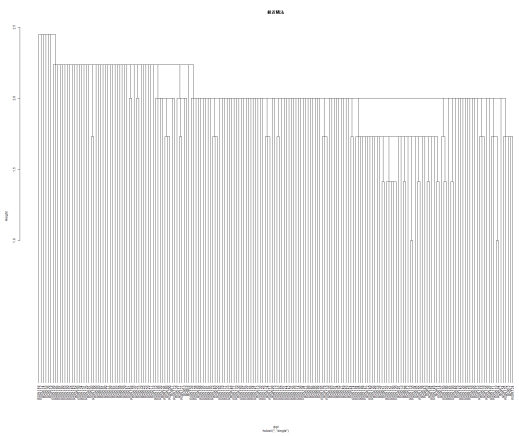

図 8　BoW でのデンドログラム（副詞）

4.4 OneClassSVM による予測

OneClassSVM は、R 言語の kernlab ライブラリ ksvm 関数により実装した。学習は、先に述べた 545 件の正例データからランダムに 200 件を抽出して学習させた。評価は、正例と負例 20 件ずつ、計 40 件のデータで評価を行った。ksvm 関数の 2 つのハイパーパラメータ sigma,nu は、グリッドサーチで最適化を行った。

文書単語行列において、全ての単語を対象としたベクトルでは、ほとんど予測することができなかった。よって、出現頻度上位 n 件だけを抽出したベクトルとし、徐々に絞り込んで最適化を行い、20〜40 単語程度で高い分類性能をもつようになった。

また、TF を要素としたベクトルでは、ほとんど予測することができず、TF-IDF を要素としたベクトルでは、ランダム以下の予測性能に留まった。

よって、BoW をベクトルの要素とし、単語数を上位 20〜40 単語に絞って詳細な評価を行った。正解率が最高となった組み合わせにおける混同行列と評価指標を、形態素ごとに次の表に示す。

表 7 名詞（上位 40 単語）抽出での混同行列

	正と予測	負と予測
正例	17	3
負例	10	10

表 8 名詞（上位 40 単語）抽出での評価指標

正解率	適合率	再現率	F 値
0.68	0.85	0.63	0.72

表 9 動詞抽出（上位 40 単語）での混同行列

	正と予測	負と予測
正例	19	1
負例	11	9

表 10 動詞抽出（上位 40 単語）での評価指標

正解率	適合率	再現率	F 値
0.70	0.95	0.63	0.76

表 11 副詞抽出（上位 21 単語）での混同行列

	正と予測	負と予測
正例	15	5
負例	5	15

表 12 副詞抽出（上位 21 単語）での評価指標

正解率	適合率	再現率	F 値
0.75	0.75	0.75	0.75

5. 考察

5.1 形態素について考察

副詞、動詞、名詞、の順に高評価となった理由は、ノイズの少なさにあると考えている。

名詞では、サードパーティサービスを利用する際に使われる形式的な名詞や代名詞が多く検出されたが、これらは高校生を特徴付ける単語とは考えにくく、サービス利用の有無を表していると考えられる。

動詞では、補助動詞が多く検出されたが、これらも高校生を特徴付ける単語とは考えにくく、属性に関係なくどの文書にも広く出現するものと考えられる。

一方、副詞では、トピックを表す単語は一切ない。省略しても意味が通じる単語ではあるが、誰かに伝えたいという強い気持ちが表れたものと考えられる。若者の SNS 利用において、『映える/栄える』や『盛る』という言葉があるが、その心理が副詞に表れているものと考えている。

5.2 文書単語行列について考察

自然言語処理での機械学習では、特徴ベクトルが高次元になりやすい。全ての単語を対象に文書単語行列を作成した際にほとんど予測することができなかった理由について、次のように考えた。

特徴ベクトルの次元が学習データ数より大きければ、どのようなラベル付けに対しても線形分離可能である。しかし、複雑な境界になってしまたために、汎化能力の低下が起こったものと考えている。

6. 研究成果と今後の展望

6.1 研究成果

今回の研究では、高校生の Twitter アカウントの属性を予測するために、次の手法を考案した。

「高校生 200 人分の直近 200 ツイートを収集し、MeCab による形態素解析で副詞を抽出して BoW モデルで文書単語行列を作成する。上位 21 単語を抽出して特徴ベクトルを生成し、OneClassSVM の機械学習アルゴリズムにて分類器を学習させる。」

本手法において、高校生の Twitter アカウントの属性推定について、正解率 0.75、F 値 0.75 で予測することができた。

6.2 今後の展望

今後は、高校生を表すの特徴量を決定する要素を絞り込むため、ヒューリスティックに探索するためのプログラムを作成する。ある程度、高校生を表す特徴量の潜在性を絞り込めたら、細かいパラメータで網羅的に探索し、予測性能を最大にするパラメータをチューニングする予定である。

参考文献

[1] 東京都教育委員会:学校非公式サイト等の監視, http://www.kyoiku.metro.tokyo.lg.jp/school/document/ict/underground_website.html (2020.02.01 アクセス).
[2] 東京都教育委員会:平成 30 年度「児童・生徒のインターネット利用状況調査」調査報告書, https://www.kyoiku.metro.tokyo.lg.jp/school/document/ict/document.html (2020.02.01 アクセス).
[3] 榊剛史, 松尾豊,“ソーシャルメディアユーザの職業推定手法の提案”, 知能と情報, vol.24 No.4, pp.773-780(2014).
[4] 中川真史, 山口祐人, 北川博之,“フォローを用いた特定地域から発信されたツイートの効率的な収集”, DEIM Forum(2018).
[5] 奥谷貴志, 山名早人, ”メンション情報を利用した Twitter ユーザプロフィール推定”, *DBSJ Japanese Journal*, vol.13-J, No.1, pp.1-6(2014).
[6] 武田直人, 関洋平, ”Twitter の投稿場所を考慮したユーザ属性推定”,DEIM Forum(2016).
[7] 石野亜耶, “大学生の Twitter アカウントの自動検出”, 広島経済大学研究論集, 第 38 巻第 3 号 (2015).
[8] 石田基広：RMeCab, http://rmecab.jp/wiki/index.php?RMeCab(2020.02.01 アクセス).