

高校生 Twitter アカウントの SVM による分類 における必要データ量と分類精度についての考察

三村 徹* 辰己 丈夫²

概要：著者のこれまでの研究において、Twitter の公開アカウントのツイート本文のデータから、職業属性が高校生であるアカウントを OneClassSVM で分類し、0.75 の正解率での分類を実現した。現在は、精度向上のため、SVM で 2 クラス分類を試みているが、実用化するには十分な学習データ量を確保することが課題となっている。

本研究では、職業属性が高校生であるアカウントの分類において、学習データの量と分類精度の関係を調査した。学習用データ、評価用データの他に、パラメータ調整用データが必要であることを考慮した上で、0.7 程度の正解率の実現には、少なくとも 100 件程度の学習データが必要となること、正解率の精度向上には、評価用データ 40 件程度では不十分であることを確認した。パラメータ調整用データの必要量については、今回の検証では知見を見出すことはできなかった。

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。さらに、警察庁 [1] によると、2018 年における SNS 等に起因する被害児童の現状において、Twitter に起因する被害が約 4 割と最も多く、被害者の約半数は高校生であった。

1.2 ネットパトロールの現状

Twitter をはじめ、匿名で利用できる SNS は、犯罪等に関する情報を発信していても、発信者の特定に至るまでが困難であり、発信者のモラル教育以外、有効な対策がないのが現状である。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 [2]。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

*放送大学大学院修士課程

²放送大学

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んでいるほど、高い動機となる。

1.4 本研究の目的

本研究の大構想は、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している。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指している。

本稿においては、前述の目的のうち、Twitter アカウントの職業属性が高校生かどうかを高精度で判定することを目標としている。また、機械学習には大量のデータを必要とすることから、判定に必要なデータ数を調査することも、同時に目標としている。

2. 関連研究の動向

2.1 機械学習を用いたサイバーパトロールシステム

安彦 [4] は、Twitter での薬物売買に関する投稿を収集し、それらを教師データとして機械学習を行い分類器を作成した。投稿本文の形態素解析により抽出した名詞群に対して BoW (Bag of Words) の手法を用いて単語ごとのラベルを割り振り、特徴語コーパスを作成し、単語の出現頻度から特徴ベクトルを作成している。

2.2 自然言語処理によるユーザ属性推定

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 [5]。

2.3 周辺ユーザ情報によるユーザ属性推定

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、属性値のクラスが存在する。そのクラス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 [6]。

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 [7]。

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、パラメータ調整データ、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) パラメータ調整用データで、分析器・分類器を最適化する。
- (5) 最適化済みの分析器・分類器で、評価データを用いて検証する。

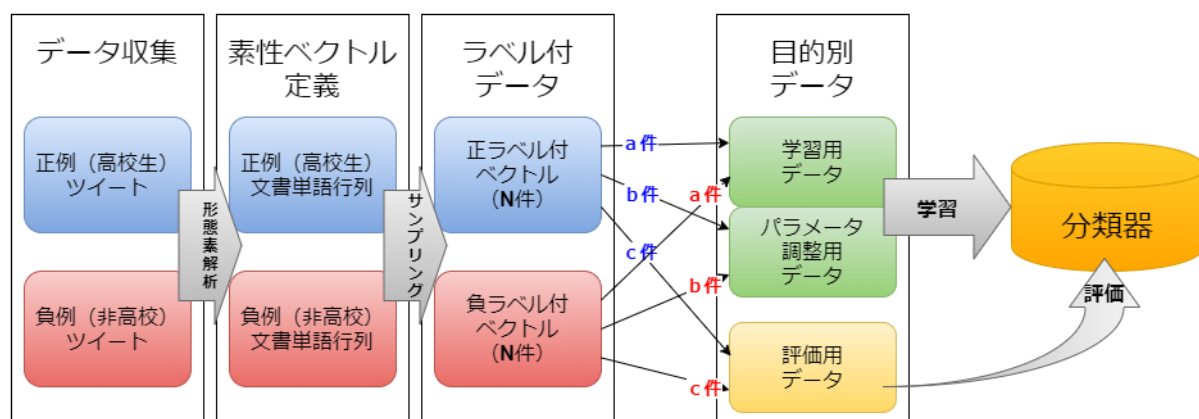


図 1 研究手法の概要

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 [8] 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 [3] 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。（図 2）。

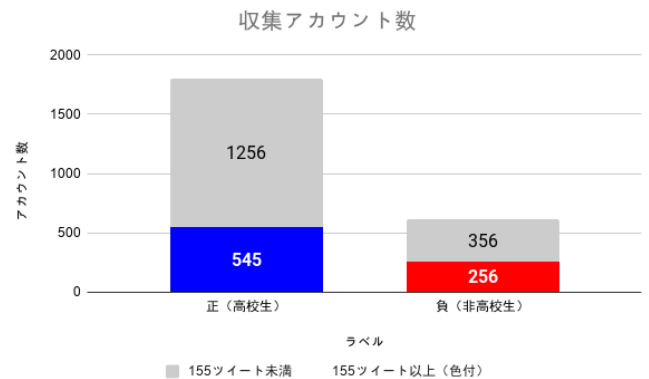


図 2 収集アカウント数

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の特徴量とする方法がある。

はじめに、学習用データ、パラメータ調整用データ、評価用データを区別せずに、日本語形態素解析システム MeCab[9] 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	

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役に立たない単語である。英語を始め、ヨーロッパ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」等がある程度取り除くことが可能である。

これまでの著者の研究 [10] において、名詞、動詞、副詞、それぞれの品詞ごとに素性ベクトルを作成し、分類を試みた結果、副詞を利用した分類が高精度となることがわかつ

ている。よって、今回は副詞の出現状況の特徴量とする方法を採用している。

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）。

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、その文書の特徴づける単語として捉えている。

これまでの著者の研究 [10] において、BoW、TF、IDF、TF-IDF、それぞれの方法で分類を試みた結果、BoW を利用した分類が高精度となることがわかっている。よって、今回は BoW を用いて特徴量とする方法を採用した。

また、文書単語行列においては、これまでの著者の研究から、出現率上位 16 単語（副詞）のみを特徴量とした [表 3]。

表3 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

3.5 交差検定パターンとその目的

収集できたデータ量が十分でない場合、同じデータを複数使用する必要があるため、交差検定により測定する。今回は、交差検定のパターンは次の I~III の 3 パターンとして、5 回の平均値を測定する。使用するデータの数は、正例、負例それぞれ 120,160,200 ずつの 3 通り行った。

- (1) パラメータ調整用データは、評価用データと兼ねる。
本来、評価用データは、分類器作成（パラメータ調整用データを用いること）から独立させるべきものであるが、パラメータがジャストフィットした際の精度（精度の理想値）として、比較用に検証する（図3）。



図3 交差検定パターン I

- (2) パラメータ調整用データは、学習用データと兼ねる。
学習用データでパラメータ調整を行うということは、学習用データが機械学習にとって「質の良い」データである場合に、汎化性能を発揮するものと思われる。分類性能を発揮するために必要な学習データの量を検証する（図4）。



図 4 交差検定パターン II

(3) 学習用データ、パラメータ調整用データ、評価用データを、それぞれ独立させる。分類器の作成・調整に使用するデータは、学習用データとパラメータ調整用データである。評価用データは、完全に分類器の作成・調整とは無関係なデータであり、純粋な汎化性能を測定できる方法である (図 5)。

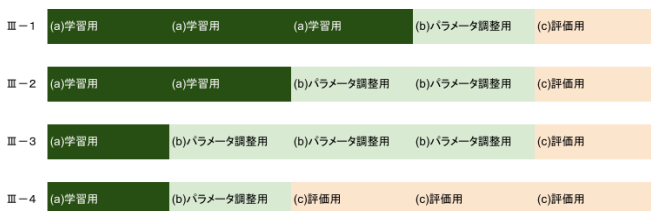


図 5 交差検定パターン III

3.6 SVM の実行環境

SVM は、R の library(kernlab)、ksvm 関数により実装している。type = "nu-svc"、kernel = "rbfdot" (ガウシアンカーネル) とし、ハイパーパラメータは、sigma, nu の 2 つである。実行環境は、次のとおりである。

表 4 実行環境	
OS	Windows10 Home 64bit
CPU	Intel Core i7-4770 3.40GHz
メモリ	8.00GB
R	v3.5.2
RStudio	v1.1

3.7 性能評価の方法

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。

		分類器が予測したクラス	
		正 (Positive)	負 (Negative)
実際の クラス	正例	TP (TruePositive)	FP (FalsePositive)
	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。

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

上記の評価指標の中で、今回は正解率を採用する。

4. 検証結果

4.1 交差検定パターン I の結果

パターン I では、評価用データに対してパラメータ調整を行った後、評価を行う。いわゆる「カンニング」をする方法である。学習データ数が多いほど、正解率が高くなる傾向にあるが、増加率の鈍化傾向は見られる。学習データ 100 件程度を超えると、正解率 0.7 以上になっている (図 6)。

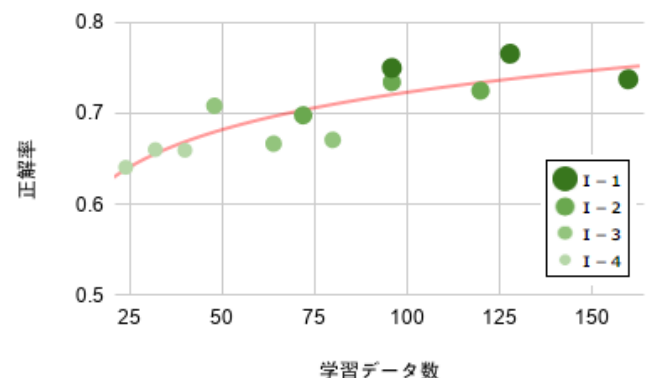


図 6 学習データ数と正解率

また、評価データの配分を多くした方法ほど正解率が低

くなっているが、これは、学習データの配分と評価データの配分がトレードオフになっているためである（図7）。

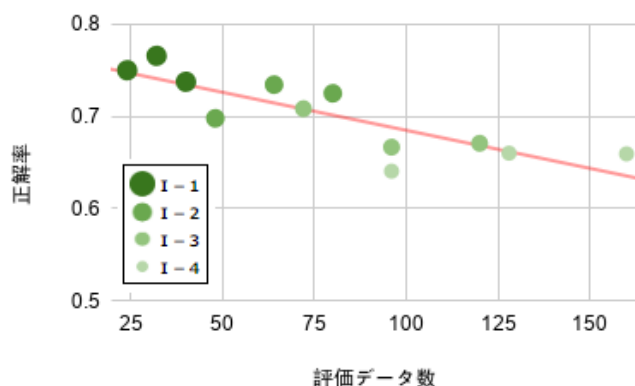


図7 評価データ数と正解率

4.2 交差検定パターンIIの結果

パターンIIでは、学習用データでパラメータの調整を試みたが、ほぼ全てのパラメータに対して100%に近い分類性能を示した。そのため、最適なパラメータを探索できないという問題が発生した。そこで、 $\sigma = 0.1$, $\nu = 0.1$ として固定し、評価を行った。

パターンIに比べ、正解率が低くなっている。パターンIと同様、学習データ数が多いほど、正解率が高くなる傾向にあるが、増加率の鈍化傾向は見られる。パターンIと比較して、正解率にバラつきがある（図8）。

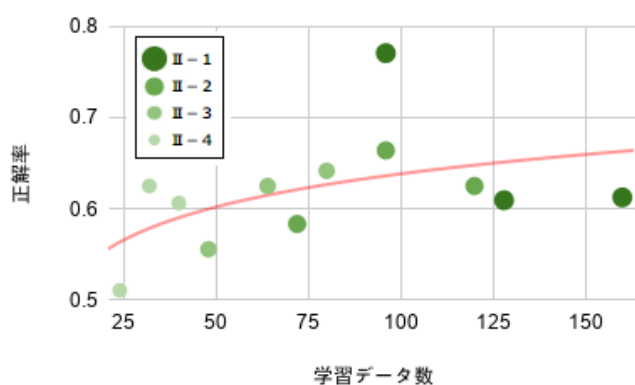


図8 学習データ数と正解率

また、パターンIと同様、評価データの配分を多くした方法ほど正解率が低くなっているが、これは、学習データの配分と評価データの配分がトレードオフになっているためである（図9）。

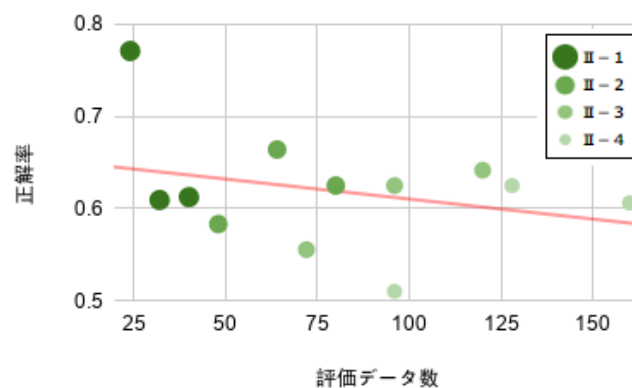


図9 評価データ数と正解率

4.3 交差検定パターンIIIの結果

増加や減少の傾向はほとんど見られず、バラつきがある（図10）。

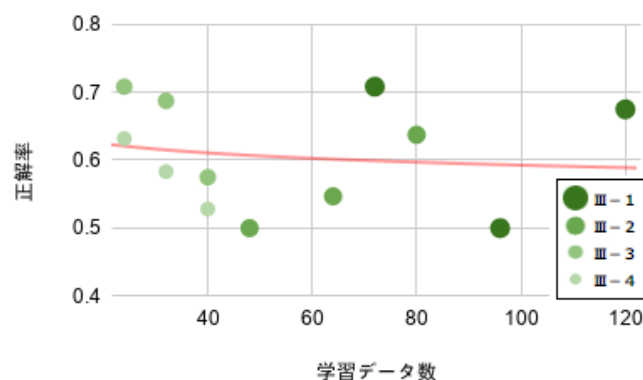


図10 学習用データ数と正解率

十分な評価用データを確保できていないと、バラつきが大きい。40件に満たない評価用データでは、特にバラつきが大きくなっている。（図11）。

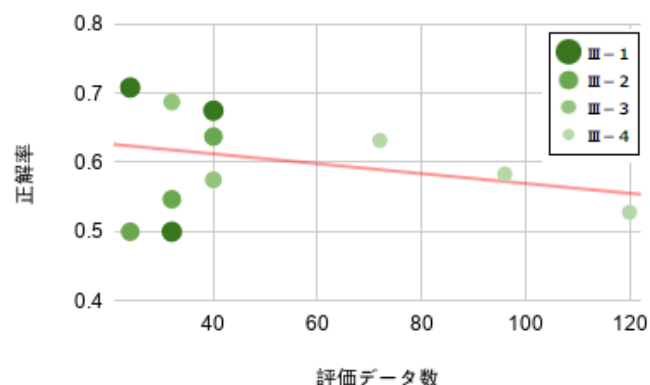


図11 評価用データ数と正解率

パラメータ調整用データと正解率の関係は、特に傾向を

見いだすことはできない。パラメータ調整用データが少ないと正解率が低くなるわけではない（図 12）。

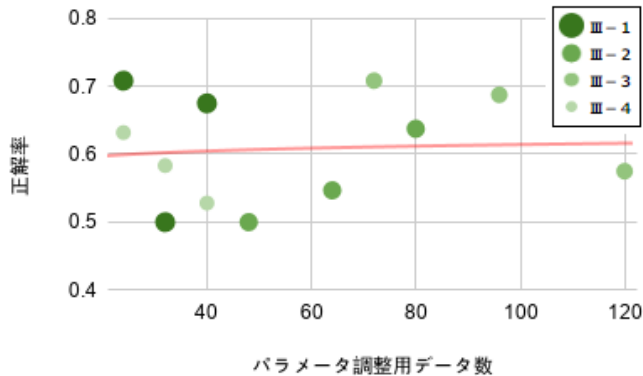


図 12 パラメータ調整用データ数と正解率

5. 考察

5.1 学習用データと正解率との関係

交差検定パターン I から、正解率は、学習データ量もしくはパラメータ調整が多いほど高くなることがわかる。そして、評価用データが多くても、正解率には影響しないと考えられる。

また、学習データ数が 100 件を超えると、0.7 以上の正解率となる分類器の作成が可能であることが分かる。学習データ数が多いほど正解率が向上すると思われるが、これ以上増やしても正解率が上がらなくなるポイント（増加の鈍化ポイント）については、データ不足から発見には至らなかった。正例、負例それぞれ 500 件程度のデータがあれば、そのポイントに近づくことができると考えている。

5.2 評価用データと正解率のバラつきとの関係

評価用データ数が十分でないと、ランダムに選択しているデータセットと分類器の相性による誤差が大きくなると思われる。同様に、分類器の正解率と学習データの相性についても、誤差の原因となっていると考えられる。このことを逆に利用して、分類器の汎化能力を引き出すデータセット（良質のデータ）を抽出する方法があるのではないかと考えている。

5.3 パラメータ調整用データと正解率の関係

交差検定パターン II から、パラメータ調整を行わないと、正解率が 0.6 程度に留まり、パラメータ調整が分類器の汎化能力に影響を与えることが分かる。交差検定パターン III からは、正解率の向上に必要なパラメータ調整用データ量は、見出すことはできなかった。データ量よりも、データの質が分類性能に影響を与えていると考えている。

6. 研究成果と今後の展望

6.1 研究成果

高校生 Twitter のアカウント属性を機械学習で分類するのに必要なデータ量のうち、アカウントの収集については、次のように整理した。

- (1) 検出アカウントのうち、ラベル付け可能なアカウントのみを利用できる。
- (2) ラベル付け可能なアカウントのうち、十分なツイート数のあるアカウントのみを利用できる。
- (3) 十分なツイート数のあるアカウントのうち、分類に有効なアカウントを選択して利用したい。

今回、正例のラベル付きアカウント 1801 件、負例のラベル付きアカウント 612 件を取得していた。しかし、十分な情報量である 155 回以上ツイートしているユーザを抽出する段階で、正例 545 件、負例 256 件にまで、データが減ってしまった。

また、正例、負例のそれぞれのデータについて、学習用、パラメータ調整用、評価用のデータが必要である。学習用データとして 100 件程度、評価用データとして 50 件程度は必要と考えられる。パラメータ調整用としては、必要であることは確かであるが、量については見出すことはできなかった。

さらに、精度の高い分類器を作成するには、その中でも良質のデータセットを選べることが望ましい。

以上のことから、高校生 Twitter アカウントを SVM で分類するにあたり、正例、負例それぞれ 200 件程度のデータが必要である。また、精度の高い分類器を作成するためには、さらにその何倍かのデータが必要であり、そのデータを収集するためには、さらにその何倍かのアカウント探索が必要である。

6.2 今後の展望

今回、負例のデータ数 256 件がボトルネックとなり、SVM の精度を向上させるためのデータ選択が十分にできない状況であった。しかし、正例のデータ数 612 件は、OneClassSVM での精度向上を試みるに十分なデータ量であると考えられる。今後は、OneClassSVM において、精度の高い分類器に必要なデータ（良質のデータ）の抽出を行う予定である。

参考文献

- [1] 警察庁：平成 30 年における SNS 等に起因する被害児童の現状 (2019).
- [2] 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).
- [3] 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](https://www.kyoiku.metro.tokyo.lg.jp/school/document/ict/document.html) (2020.02.01 アクセス).
- [4] 安彦智史, 加藤諒, 北川悦司, “機械学習を用いた薬物売買におけるサイバーパトロールシステムの開発”, 情報処理学会論文誌, Vol.61 No.3, pp.535-543 (2020).
 - [5] 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).
 - [6] 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).
 - [7] 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).
 - [8] 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).
 - [9] 石田基広: RMeCab, [http://rmecab.jp/wiki/index.
php?RMeCab](http://rmecab.jp/wiki/index.php?RMeCab)(2020.02.01 アクセス).
 - [10] 三村徹, 辰己丈夫, “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”, 情報処理学会研究報告, CE154(2020).