# 文書データの PU 学習における仮正例の抽出による分類性能向上手法

情報数理応用研究

5217C032-7　水落洋貴
指導教員　後藤正幸

# A Method of Enhancing Classification Performance on PU Learning for Document Classification by Extracting Provisional Positive Examples

HIROKI Mizuochi

## 1　研究背景・目的

近年の情報技術の発展に伴い膨大な数の電子文書データが蓄積されており，それらからユーザが目的とする情報を獲得する技術の重要性が高まっている．それらの技術の 1 つとして，ユーザが目的とする文書（以下，正例文書）を抽出することを，正例文書とそれ以外の文書（以下，負例文書）を分類する二値分類問題として捉える枠組みがある．この問題において，「正例文書は比較的少数しか与えられず，負例文書は一切与えらない．代わりに正例か負例か不明なラベルなし文書が大量に与えられる」という状況が生じることが多々ある．例えば文書推薦では，ユーザが自身の興味に基づき閲覧した少量の文書は正例として与えられるが，他の大多数の閲覧していない文書は，興味がないため閲覧していないのかそれとも興味はあるが存在を知らないなどの理由により閲覧していないのかが不明であるため，負例ではなくラベルなし文書として与えられる．このような問題設定のもとで，新たな入力文書が正例と負例のどちらであるかを判別する二値分類器を学習する問題は Positive Unlabeled Learning (以下，PU 学習) と呼ばれ，近年盛んに研究されている [1],[2].

PU 学習の中でも代表的な手法の 1 つである PEBL[3] では，半教師付き学習の枠組みを援用し，ラベルなし文書の中から正例文書と大きく異なる特徴を持つ文書を「仮負例」とみなすことで学習用文書を補ったのちに，「正例」と「仮負例」から二値分類器を学習する．この手法は，真の負例文書が正例の集合から距離的に離れている場合には良好な結果が期待できる．しかしながら，多くの場合において与えられる正例の文書数が比較的少量であるため，大半の文書を負例と分類してしまうことや，オーバーフィッティングしやすいことなどの問題があった．そのため，著者らは負例への誤分類を抑えるため，2 種の二値分類器を組合わせた方法を提案している [4]．これらの手法では，「仮負例」のラベルの与え方が重要な着眼点であり，学習用文書として与えられた「正例」とラベルなし文書の中から推定した「仮負例」を用いて二値分類器を学習する．一方で，ラベルなし文書の中には，「正例」も含まれると考えられるが，これらは学習には有効活用されていない．すなわち，何らかの方法によりラベルなし文書の中から「仮正例」を推定し，正例文書数を拡張することができれば更に分類性能を向上させることができると考えられる．そこで，本研究では，ラベルなし文書から正例文書の特徴を持った文書を「仮正例」とみなして学習を促進させることで，分類精度をさらに改善した学習アルゴリズムを構築することを目的とする．

一方，データ拡張を行うことのできるモデルの 1 つとして，近年，Generative Adversarial Network [5](以下，GAN) が提案され，高い注目を浴びている．GAN は特に画像認識の分野において，入力データと似たデータを自動生成する技術として高い性能を有している．したがって，文書データの PU 学習に対しても GAN を適用することで正例データの拡張が可能であれば，性能を改善させることができると考えられる．しかし，画像データと異なり本研究が対象とする文書データは高次元かつスパースという特性を有しているため，GAN によるデータ生成が有効であるか否かは明らかではない．

そこで，本研究ではまず，GAN による正例文書の学習と生成についてその性能を実験的に評価し，文書データ生成における GAN の有効性を検証する．その結果，GAN による文書生成は画像の生成に比べ限定的な場合にのみ効果を発揮することを明らかにする．加えて，「仮正例」を与える手法として，GAN の学習結果を用いる手法，およびあらかじめ与えられた少数の正例を用いた手法の 2 つの手法を提案し，PU 学習の枠組みで実際の文書データを用いた分類実験を行い，その有効性を示す．

## 2　準備

### 2.1　Positive Unlabeled Learning(PU 学習)

予め与えられた $N$ 件の学習用文書の集合を $\mathcal{D}_L = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N\}$，$d$ 番目の文書のラベルを $y_d \in \{-1, 0, 1\}$ とする．ただし，$y_d = 1$ の文書を正例，$y_d = 0$ の文書をラベルなし，$y_d = -1$ の文書を負例とする．また，それぞれの文書の特徴量は，形態素解析などで単語に分割したのち，それらの単語の出現頻度を要素として持つ頻度ベクトルとして表現される．すなわち，全文書中に出現する単語集合を $\mathcal{W} = \{w_i, w_2, \cdots, w_V\}$ とすると，$d$ 番目の文書は $\boldsymbol{x}_d = (x_{d1}, x_{d2} \cdots, x_{dV})$ と表現される．ただし，$w_v$ は $v$ 番目の単語，$x_{dv}$ は $d$ 番目の文書における $v$ 番目の単語の出現頻度である．

PU 学習は，$\mathcal{D}_L$ 中の $y_d = 1$ である少数の文書の集合を $\mathcal{P}$，$y_d = 0$ である大量の文書の集合を $\mathcal{U}$，$y_d = -1$ である文書の集合 $\mathcal{N} = \phi$ のもとで新たな入力文書集合 $\mathcal{D}_T$ を正例と負例のラベルのいずれかに二値分類する分類器を学習する問題である．すなわち，$\mathcal{D}_L = \mathcal{P} \cup \mathcal{U}$ を元に新規文書のラベル $\hat{y}_t \in \{-1, 1\}$ を推定する問題で，与えられた正例文書数を $N_P$，ラベルなし文書数 $N_U$ とすると，$N = N_P + N_U$ かつ $N_P \ll N_U$ である．

### 2.2　著者らの手法 [4]

既存の PU 学習手法の一つである PEBL では，まず正例文書から距離が遠い位置にあるラベルなし文書を仮負例として抽出し，与えられた正例文書と抽出された仮負例を用いて二値分類器を学習する．その後，得られた二値分類器によって負例と分類されたラベルなし文書を新たに仮負例として抽出する．この「分類器の学習」と「仮負例の抽出」を仮負例が抽出されなくなるまで繰り返し，最終的に得られる分類器を新規文書の分類に用いる．この手法により得られる分類規則ではほとんどの文書が負例と分類されてしまうことに注目し，著者らは新たな PU 学習手法として 2 つの分類器を組み合わせる手法の提案をした [4]．具体的には，まず全てのラベルなし文書を仮負例とみなして学習さ

れた二値分類器である HARD 分類器を学習する．次に，HARD 分類器の識別境界から距離の遠い $X$ 件の文書を仮負例とみなして学習された二値分類器である SOFT 分類器を学習する．そして得られた 2 つの分類器の出力の和がある値 $\alpha$ より大きいものを正例と分類する手法である．この手法は，学習用文書集合中の真のラベルが正例である文書数より負例である文書数が非常に多い場合にも有効であることが示されている．

### 2.3 Generative Adversarial Network（GAN）

GAN[5] は，2014 年に Goodfellow らによって考案された生成モデルであり，ランダムノイズを入力することで学習用データと類似したデータを自動で生成することができる．GAN は Discriminator 呼ばれる識別モデルと Generator と呼ばれる生成モデルの 2 つのネットワークモデルから構成されており，式 (1) で表される目的関数を Generator に関して最小化，Discriminator に関して最大化するという 2 つのネットワークの敵対的学習アルゴリズムにより最適化する．ただし $\mathrm{P}_{\mathrm{data}}(\boldsymbol{x})$ は真のデータが従う分布，$\mathrm{P}_{\boldsymbol{z}}(\boldsymbol{z})$ はノイズ分布，$\boldsymbol{z}$ はノイズであり，$D$ は Discriminator，$G$ は generator を意味する．

$$V(D,G) = E_{\boldsymbol{x}\sim \mathrm{P}_{\mathrm{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + E_{\boldsymbol{z}\sim \mathrm{P}_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1-D(G(\boldsymbol{z})))] \quad (1)$$

この手法は特に画像認識の分野において高い性能を示すことが知られており，人物画像や動物画像などのように複雑な画像に対しても自動で特徴量を学習し，類似した画像を生成できることで脚光を浴びている．

## 3 提案手法

### 3.1 概要

PU 学習では，学習用文書中にカテゴリ既知の文書として正例は存在するが負例は存在しないため，SVM などの通常の学習アルゴリズムにより二値分類器を学習することができない．この問題に対し，PU 学習のアルゴリズムの 1 つである PEBL[3] では，まず正例文書集合 $\mathcal{P}$ を手掛かりにラベルなし文書集合 $\mathcal{U}$ 中から正例と大きく特徴の異なる文書を抽出し，「仮負例」とみなす．これにより，擬似的に学習用文書に負例を補完することで学習を可能としている．当然ながら，このとき仮負例として抽出される文書の真のラベルはすべて負例であることが望ましい．しかし，ラベルなし文書数に比べ正例文書数が相対的に少ないだけでなく，絶対数が少ない場合には，これら少数の正例文書を利用して仮負例を抽出すると，真のラベルが正例である文書も仮負例として抽出されてしまう恐れがある．また，ごく少数の正例とそこから得られた仮負例を元に分類器を学習すると，正例にオーバーフィットしやすくなると考えられる．したがって，正例文書数を何らかの方法で拡張することができれば分類性能を向上させることができると考えられる．そこで，本研究では学習用文書に仮負例を補完するだけでなく，正例の特徴を持つと考えられる文書を「仮正例」として正例文書に加えることを考える．すなわち，正例，仮正例，仮負例の 3 種類の文書を用いて分類器を学習することで性能の向上を目指す．

少量のデータや不均衡データに対しデータの拡張を行う手法としてオーバーサンプリングがある．オーバーサンプリングは対象データの平均や分散を元にデータを生成する手法であるが，特徴量間の相関や全体としての分布を考慮しておらず，本来出現しにくいデータも多く生成されたり，逆に元データとほとんど同じデータが集中的に生成されたりする恐れがある．本研究が対象とする文書データには出現する単語同士の共起関係があること，また多くの単語の出現頻度が 0 でありごく一部の単語のみ出現することなどの特徴があり，オーバーサンプリングにより文書データを生成すると偏ったデータが生成されやすいと考えられる．また，正例の拡張のために新たに与える文書は「正例の特徴を強く持っている」ことと「すでに得られているデータと部分的に異なる要素を持つ」ことが望ましいが，オーバーサンプリングにより同じデータが大量に生成された場合は分類性能の向上にはあまり寄与しないと考えられる．したがって，より文書の特徴を考慮しており，かつバリエーションが豊富な文書を仮正例とすべきであるのでこの手法では望ましくない．

一方，GAN は類似したデータを自動で生成できるモデルとして人工知能分野で大きな注目を集めている．特に画像データに対して性能が良いとされているが，文書データに対しての当てはまりのよさは不明である．この GAN が本研究で対象としている文書データに対しても当てはまりが良いならば，PU 学習の正例文書に GAN を適用して正例文書集合に似たデータを生成し，それらを仮正例として正例を拡張することで分類性能を向上できると考えられる．しかしながら，文書により学習された GAN から生成された文書は何らかの統計的な特徴は再現されているものの，特に様々な話題が入り混じった文書集合を正例として学習した際，あらゆる話題に出現するような単語が多く出現し，正例としての特徴が弱い文書が生成されていることが確認された．したがって，GAN から生成されたデータをそのまま「仮正例」として用いることは得策ではないと考えられる．そこで，仮負例と同様に仮正例もラベルなし文書から抽出することを考え，その際に GAN から生成された文書を利用することを考える．更に，本研究ではあらかじめ与えられた正例文書集合のみを用いてラベルなし文書から仮正例を抽出することも考える．

したがって，本研究では仮正例抽出方法として (1)「GAN から生成された文書を用いた仮正例抽出」または (2)「正例文書からの距離を用いた仮正例抽出」を行った後に，正例，仮正例，仮負例を用いた分類手法を提案する．以下でより具体的に説明を行う．

### 3.2 仮正例抽出方法

本研究では，ラベルなしデータの中から，正例の可能性が高いと推測されるものに対して「仮正例」のラベルを付与することを考える．そのために，学習用文書として与えられた正例文書または GAN から生成された文書をアンカーデータの集合とみなす．そして，これらのアンカーデータとの距離を用いて個々のラベルなし文書が仮正例か否かを決定する．

まず第一に考えられる方法として，その距離の平均によりラベル付けを行う方法がある．しかし，正例文書の単語分布は均質ではなく複数の異なるトピックを持つ文書グループから構成されていると考えられる．単純に距離の平均を用いて仮正例を抽出すると，すべてのアンカーデータから均等に近い文書が仮正例として抽出されやすくなる．その結果，本来は正例の中には存在しないような，多様な単語を同時に含む文書が仮正例として抽出されてしまいやすくなると考えられる．したがって，今回のように正例の可能性が高い文書のみを仮正例として抽出することが目的である場合の方法としてはふさわしくない．そこで，ラベルなし文書ごとにアンカーデータとの間のコサイン類似度を求

め，その値が $C$ 以上となるアンカーデータ数をラベルなし文書ごとに算出し，その数が多い上位 $T$ 件のラベルなし文書を仮正例として抽出する．これにより，どこか一つでも正例が密集している領域（類似のトピックを持つと考えられる正例文書データ群）に近いラベルなし文書が特に多く抽出され，すべての領域に対して平均的に近いがどの領域にも属していないような文書が抽出されにくくなると考えられる．

本研究ではまず (1) または (2) のいずれかの方法により仮正例を抽出する．そして得られた仮正例とあらかじめ与えられた正例文書を合わせて正例とみなして著者らの手法 [4] により分類器を学習する．これにより，少数の正例しか得られていない問題設定の下でも分類性能を高く保った分類器の学習が可能になると考えられる．

## 4 実験

### 4.1 実験概要

本研究ではまず，画像データに対してすでに有効性が報告されている GAN を文書データに対して適用した際の挙動について検証する．その後，提案手法である (1)GAN の生成データをアンカーとした仮正例抽出方法，(2) 学習データ中の正例をアンカーとした仮正例抽出方法，並びに (3) 仮正例を抽出せずに行う著者らの従来手法の分類実験を行い性能の比較評価および，考察を行う．文書データとして 2000 年の読売新聞記事データを用いた．このデータは 5 つの大カテゴリからなり，その中に小カテゴリが 28 カテゴリ存在する．また，全学習用文書として 2,557 件，テスト用文書として 1,400 件を用いて実験を行った．

### 4.2 実験 1：GAN の文書データに対する適用実験

文書データを用いて GAN を学習し，得られた Generator から文書を生成することでどのような文書が生成されるかを検証する．具体的には，カテゴリを一つに限定して Generator を学習し，生成させた文書内の出現単語頻度をランキングすることで，そのカテゴリを特徴づける単語が出ているか否かを検証する．この実験を行うにあたり，大カテゴリ 1〜5 の中からそれぞれ小カテゴリを 1 つ選び，カテゴリ中の全文書を入力文書として学習し，300 件の文書を生成した．その後，300 件の生成文書中の全単語の出現頻度の和を算出し，ランキングにした．加えて，5 カテゴリを同時に入力・生成を行い同様にランキングにした．表 1 に 5 カテゴリ同時（全カテ）の単語のランキングとそれぞれのカテゴリにおける GAN からの生成文書中の単語のランキングを示す．ただし，全カテ以外，学習用文書のランキングの上位 30 件の単語を除外している．

全カテ中の単語のランキングには，カテゴリをまたがって広く用いられる漢数字や動詞の「する」といった単語が上位にあがっている．一方で政治カテゴリのランキングの特に上位には「国会」や「自民党」のように政治カテゴリの特徴を強く表すような単語が挙がっている．また，動植物カテゴリには比較的カテゴリを特徴づけるような単語というよりは全てのカテゴリで出現しうる単語が特に上位に挙がっている．動植物に関する文書では，生物の名称のように一部の文書でしか使われていないような単語がカテゴリを特徴づけており，それらの出現頻度が他のカテゴリでも使われる一般的な単語に比べ少ないためこのような文書生成の傾向となった．また，表外の順位ながら「自然」や「生息」といった言葉も比較的上位にランクしていたことを考慮すると，学習自体に問題があるというわけではなく，カテゴリを強く特徴づけるもののそのカテゴリの中でもごく限られた文書にしか出現しないような，出現頻度が低い重要単語の生成は難しいことが予想される．

表 1．生成文書の単語出現頻度ランキング

| 順位 | 全カテ | 政治 | 経済 | 社会 | 犯罪事件 | 動植物 |
|---|---|---|---|---|---|---|
| 1 | た | 国会 | 経済 | い | 調べ | の |
| 2 | し | 自民党 | 景気 | の | 火事 | だ |
| 3 | 日 | 議員 | 会 | れる | ごろ | い |
| 4 | 二 | よう | 会議 | う | 無職 | 写真 |
| 5 | 十 | 氏 | ある | 問題 | さん | よう |
| 6 | さ | 加藤 | 企業 | ら | い | ある |
| 7 | 一 | 政権 | 回復 | 目 | 分 | 町 |
| 8 | する | 衆院 | だ | だ | うえ | ない |
| 9 | いる | 党 | 的 | 今 | 駆け付け | で |
| 10 | 三 | だ | 会長 | もの | 午後 | 県 |
| 11 | れ | 会 | 雇用 | 東京 | よる | れる |
| 12 | だ | べき | か月 | 会 | 運転 | いう |
| 13 | ない | 選 | 商工 | ない | 署 | 話し |
| 14 | こと | 連立 | ため | 中 | 町 | さん |
| 15 | な | 政策 | 投資 | 無職 | 緊急 | られ |
| 16 | あっ | ため | 消費 | です | 万 | 現在 |
| 17 | 五 | 次期 | なっ | 円 | 時 | たち |
| 18 | 年 | の | 連続 | 家庭 | 木造 | 昨年 |
| 19 | ある | 委員 | 技術 | 科学 | 無事 | 花 |
| 20 | 市 | 中 | 公共 | 方 | あっ | たい |

### 4.3 実験 2：分類実験

正例とラベルなし文書が与えられた下で正例と負例を分類する問題に対し，「仮負例」と共に「仮正例」をラベルなし文書から抽出して学習に用いる提案手法の評価を行うため，実文書データを用いた分類実験を行う．ここでは，大カテゴリ 5 つの中から 1 つずつ小カテゴリを選び計 5 カテゴリを正例カテゴリとして固定し，その中から 20 件× 5 カテゴリの計 100 件を与えられた正例文書，それ以外をラベルなし文書として乱数により選ぶ操作を 20 回繰り返した際の分類性能の平均により評価する．したがって，各実験における学習用文書の内訳は，正例文書 100 件，ラベルなし文書 2,447 件（うち，456 件が正例，1,991 件が負例) となっている．また提案手法で用いるコサイン類似度の閾値を $C = 0.6$ と設定した．評価指標は ROC 曲線下面積 (AUC) とする．比較手法は，(1) 正例文書をアンカーとした仮正例抽出方法，(2)GAN 生成文書をアンカーとした仮正例抽出，(3) 著者らの従来手法 (仮正例抽出無し) の 3 つである．仮正例数 $T$ と仮負例数 $X$ を変化させて算出した AUC の結果を以下の表 2 に示す．仮正例数が 0 の行が著者らの手法，それ以外の場所は上段が (1) の手法，下段が (2) の手法を意味しており，太字は仮負例数を固定した下での最も分類性能の高い箇所を示している．また，仮正例数 20 で固定した下で仮負例数を変化させたときの AUC の推移を図 2 に示す．

表 2 より，いずれの仮正例数・仮負例数においても学習用文書として予め与えられた正例をアンカーとして仮正例を抽出する手法が優れていた．また，仮正例数が少ないときは仮負例数を多くした方が性能が高く，仮正例数が多いときは仮負例数を少なくした方が性能が高くなることが見て取れる．分類性能には，単にどれだけ多くの仮正例・仮負例を抽出できるかということだけではなく，抽出した仮正例･仮負例がどれだけ真の正例・負例ラベルを正しく反映しているかも大きくかかわっている．したがって，単純に仮正例･仮負例を多く抽出すれば性能が上がるというわけではない．また，多く抽出することと正しく抽出することはトレー

ドオフの関係にある．そのため，仮正例数が多い場合には負例を仮正例として誤って抽出してしまい，その後の仮負例の抽出にも悪影響を及ぼすため，仮負例数が少ないときに性能が高くなると考えられる．また，仮正例数が少ない場合には誤抽出も少ないため，比較的多くの仮負例を抽出することで性能を上げることができると考えられる．本研究における問題設定は予め与えられている正例数が少ないことを想定しているので，少ない正例数に誤った仮正例を追加してしまうと性能の悪化が激しいと考えられる．このことから仮正例数が 5 という小さい数で仮負例数が 1,000 という大きい数字の時に最も性能が高くなったと考えられる．

また図 2 より，GAN により仮正例を抽出する手法は 1,000 件前後を境に性能の劣化が激しくなっているがほかの 2 手法は劣化がそれほど大きくない．これは，GAN により生成された文書が統計的性質を部分的に再現できてはいるものの完全に再現できているわけではないため，劣化に大きく寄与してしまうためだと考えられる．

表 2．仮正例数と仮負例数を変化させた分類性能評価

| 仮正例数 | 仮負例数 100 | 200 | 500 | 1000 | 2000 |
|---|---|---|---|---|---|
| 0 | 0.7566 | 0.7656 | 0.7929 | 0.8054 | 0.8024 |
| 5 | 0.7656 | 0.7751 | 0.7985 | **0.8119** | **0.8051** |
| | 0.7633 | 0.7726 | 0.7972 | 0.8068 | 0.7982 |
| 10 | 0.7711 | 0.7791 | **0.8013** | 0.8102 | 0.8047 |
| | 0.7637 | 0.7733 | 0.7968 | 0.8043 | 0.7964 |
| 15 | 0.7709 | 0.7815 | **0.8013** | 0.8087 | 0.8024 |
| | 0.7636 | 0.7722 | 0.7943 | 0.8015 | 0.7922 |
| 20 | 0.7712 | **0.7822** | 0.8000 | 0.8061 | 0.7991 |
| | 0.7625 | 0.7721 | 0.7913 | 0.7984 | 0.7875 |
| 50 | **0.7718** | 0.7814 | 0.7945 | 0.7998 | 0.7901 |
| | 0.7604 | 0.7681 | 0.7842 | 0.7888 | 0.7747 |

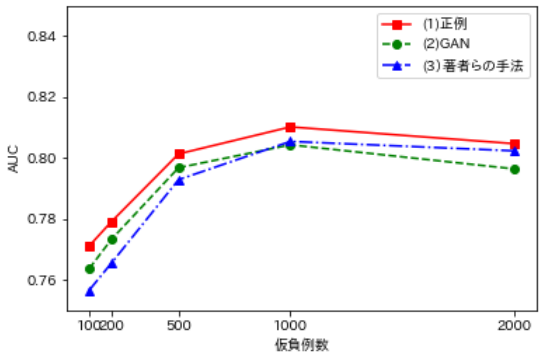


図 1．仮負例数を変化させた AUC の推移

### 4.4 考察

本研究における主な着眼点は，GAN によって文書のような高次元スパースなデータの生成はうまく行えるのか，また，生成された文書を用いて少数のデータの拡張を行い，分類性能の向上をさせることができるのかということであった．本研究で判明したこととして，GAN による文書の生成は学習に用いる文書の単語出現頻度に強く依存しており，実験 1 における政治カテゴリのような特徴的な単語が同カテゴリ内の複数の文書に出現するようなデータセットに対しては学習がうまく行われると考えられる．しかし，複数カテゴリを合わせて正例とした場合のように，多様な特徴的単語が少数の文書で集中的に使われているような文書データセットを用いて学習を行った際は，あらゆる文書に出現しているような一般的単語ばかり生成してしまい，カテゴリを特徴づける文書という意味での生成は難しい結果となった．

したがって，高次元スパースな文書の集合で，かつ内包している話題が多様なデータセットを GAN により学習する際には，予めデータセットをクラスタリングしてそれぞれ学習を行うことや tf-idf などの出現する文書数を考慮した単語表現の利用とそれに合わせたパラメータチューニングなどが重要となると考えられる．現状では GAN により生成された文書では与えられた正例文書を超える性能を実現することはできていないが大きく劣っているわけではないため，文書の特性に特化した GAN を考案することで性能を超える可能性は十分にあると考えられる．

また，本研究では仮正例の抽出方法として，与えられた正例のみを用いる方法と GAN の生成文書のみを用いる方法の 2 つの手法を提案したが，それぞれを別々に行うのではなく，例えば与えられた正例文書集合と GAN の生成文書集合を共にアンカーとみなして仮正例を抽出するという手法も考えられる．本研究においては GAN の文書生成性能の検証が大きな柱の一つとなっているので別々に検証を行ったが，このような手法により特に与えられた正例文書数は少ないが出現単語の傾向が文書間でさほど違いがないようなデータセットに対しては高い性能を示すことが予想される．

## 5 まとめと今後の課題

本研究では PU 学習の枠組みにおける文書分類問題において，得られている正例文書に近い文書をラベルなし文書から抽出し，元の正例と合わせて補強をすることで分類性能を高める提案を行った．その際，学習用文書中の正例文書をアンカーとして仮正例を抽出する手法では性能を高めることができたが，GAN という近年提案された生成モデルにおいても一定の成果が得られることがわかった．今後の課題としては，仮正例と仮負例の抽出を同時に行う手法の提案や文書に特化した GAN の学習方法の考案などが挙げられる．

## 参考文献


[1] Elkan, C., and Noto, K., “Learning Classifiers from Only Positive and Unlabeled Data,” *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 213–220, 2008.

[2] Liu, B., Dai, Y., Li, X., Lee, W. S., and Yu, P. S., “Building text classifiers using positive and unlabeled examples, ” *In Proc. The 3rd IEEE ICDM*, pp. 179–188, 2003.

[3] YU, H., Han, H., and Chang, K.C.-C., “PEBL: Positive Example Based Learning for web page classification using SVM,” *Proc, ACM Special Interest Group on Knowledge discovery and Data Mining*, pp. 239–248, 2002.

[4] Mizuochi, H., Okayama, S., Kumoi, G., and Goto, M., “A Study on Semi-supervised Learning Using a Small Number of Positive Example Documents and Unlabeled Documents,” *15th Asian Network for Quality Conference*, ICT-15, 2017.

[5] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y., “Generative Adversarial Networks,” *In Advances in neural information processing systems*, pp. 2672–2680, 2014.