CE-002

ビジネスメール詐欺対策としての送信者推定システムに関する研究
An Email Author Identification System Based on Machine Learning for Business Email Compromise

榑松理樹‡　山崎隆平‡　小笠原諒‡　羽倉淳‡　藤田ハミド‡
Masaki Kurematsu　Ryuhei Yamazaki　Ryo Ogasawara　Jun Hakura　Hamido Fujita

1. はじめに

ビジネスメール詐欺（Business Email Compromise，以後 BEC と略記）とは，IPA の報告書[1]によれば，“巧妙な騙しの手口を駆使した，偽の電子メールを組織・企業に送り付け，従業員を騙して送金取引に係る資金を詐取するといった，金銭的な被害をもたらすサイバー攻撃”である．米国連邦捜査局によれば，2013 年 10 月から 2018 年 5 月までに米国インターネット犯罪苦情センターを含む複数の情報源に報告された BEC の発生件数は 78,617 件，被害総額は約 120 億 US ドル（未遂を含む）[2]にのぼっている． また，日本在住の法人組織の情報セキュリティ・社内 IT・経理責任者ら 1,030 人を対象としたトレンドマイクロ社の「ビジネスメール詐欺に関する実態調査 2018」[3]では，全体の約 4 割がビジネスメール詐欺の攻撃を受けた経験があることが報告されている．これらの結果から BEC が日本国内の企業・組織に対する脅威であることがうかがえ，対策が必要となっている．

このような BEC への対策として，前出の IPA の報告書[1]では，取引先とのメール以外の方法での確認，社内規程の整備，普段とは異なるメールに注意，ウィルス・不正アクセス対策などがあげられている．しかし，これらの多くは人手で行う必要があり，その労力は大きい．そのため，その負荷軽減が必要である．

一方，電子メールに対する分類研究としては，SPAM 電子メール検出（以後，SPAM 検出と略記）がある．これらの研究においては，電子メール中の語や文字に基づき構築した文書ベクトルに機械学習アルゴリズムを適用して生成した分類器で分類することが主流である．近年精度が高いものも報告されている．この考え方は，BEC にも適用できる可能性は高い．

以上の背景から，我々は，BEC 対策の 1 つである「普段とは異なるメールに注意」という点に着目し，SPAM 検出で用いられる方法を援用することでこれを支援するシステムの構築を試みる．本研究では，普段と異なるメールとは，文体や語用の傾向が，同一送信者（メールアドレス）から送られてきた過去のメールにおける傾向と異なっているメールとして捉える．この考えに基づき，文体や語用を特徴として捉え，文書ベクトルを作成し，機械学習アルゴリズムに適用することで識別器を構築する．この識別器を用い，新規メールの判定を試みる．以下，2 章においては提案手法の詳細を示す．3 章においてオープンデータを用いた評価実験の結果を示すとともに，考察を述べ，5 章で本研究をまとめる．

2. 手法

提案手法の概要を図 1 に示す．図 1 に示すように本提案手法は，識別器を構築する学習フェーズと識別を行う識別フェーズから構成される．以後，各フェーズについて説明する．

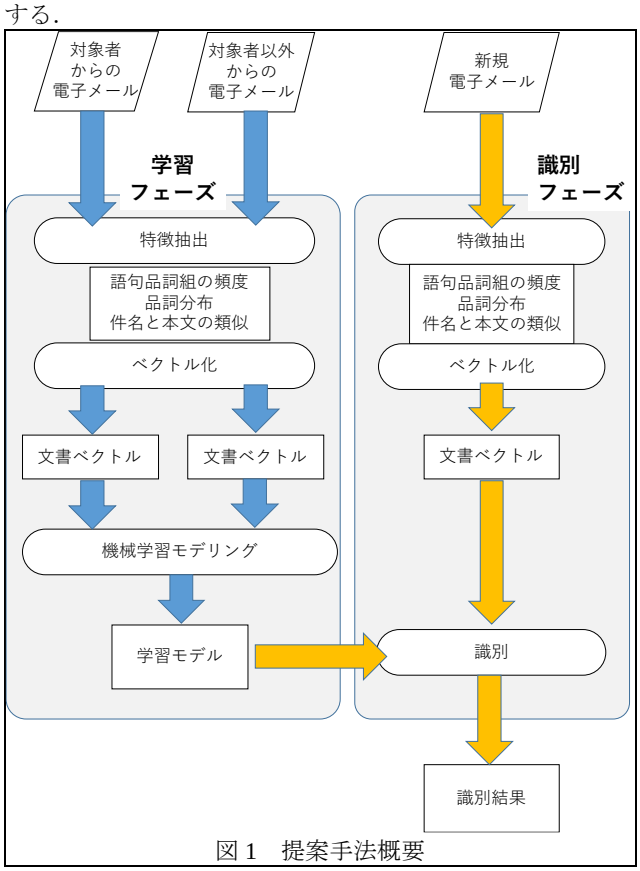

図 1　提案手法概要

2.1 学習データ

識別器を構築するデータとしては，受信したメールを用いる．本研究では，同一アドレスから送信されてきたメールに対し，普段と同じ，すなわち過去に送られてきたメールと文体や語用の傾向が同じか異なるかを判断する．そのことから，対象とする特定の送信者から送られてきた過去のメールを正解データ，そのほかの送信者から送られてきたメールを負解データとし，以後の処理を行う．

2.2 特徴抽出・文書ベクトル化

電子メールは，送信者や送信先のメールアドレス，件名，本文などから構成される．これらのうち，本研究では，件名および本文から特徴抽出に，送信者のメールアドレスを学習ラベルとして利用する．特徴抽出を行うにあたり，電子メール本文に対し，次にあげる処理を行う．

(1) 短いメールの削除：本文が数単語のみしかないメールをデータから削除する．これは，単語数が少ないと有用な特徴を得られないと考えたためである．

(2) 電子メールの本文前半部の抽出：電子メールには他人のメールが引用されている場合がある．この部分は同一送信者識別に対してはノイズとなるため取り除くことが好ましい．一方で引用方法は人によって異なるため，引用部分を自動で抽出するのは困難である．また，自らの経験に基づけは，受信者がその電子メールが同一人物から送られてきたかの判断をするのはメールの前半分であると考えられる．以上のことから，本研究では，メール本文から閾値で決めた個数の単語を抽出し，本文としては，この部分を用いる．

以上の処理を施した本文に対し，次にあげる点を特徴として抽出する．

(1) 語句品詞組の頻度：同一送信者であれば，語用が類似することが考えられる．また同じ語句であっても品詞が異なる場合も考えられる．本研究では，これらの点を考慮し，語句と品詞の組を作り，その組の頻度を特徴としてとらえる．具体的には訓練データにおいて，閾値以上出現する語句を抽出する．この特徴は，SPAM 検出においても利用されている一般的な項目である． SPAM 検出においては，単語のみを利用することが多いが，本研究では品詞と組みにすることで，同一単語でも用法が異なる場合は別に扱うことを試みる．また，同じ理由からステミングは行わない．

(2) 品詞毎の頻度：同一送信者であれば，文体が類似することが考えられる．この特徴をつかむために，本研究では品詞毎の頻度を特徴としてとらえる．具体的には訓練データにおいて，全品詞の出現数をカウントし，その分布を利用する．ただし，データ全体で出現しなかった品詞については利用しない．

また，上記に加え，件名について以下の値を特徴として抽出する．

(3) 件名と本文の類似度：電子メールの件名の内容は，人によって異なる．本文の概要を書く人もいれば，タイトルのみの人，あるいはつけない人もいる．この点から，本研究では，本文と件名との関係も送信者の特徴を示すという仮説をたてた．この仮説に基づき，件名に出現する単語のうち，本文に出現する割合を，件名と本文の類似度として求め，利用する．

以上の特徴を各電子メールから抽出し，それらを要素とする文書ベクトルに変換する．この文書ベクトル集合に対し，次節で述べる機械学習を適用し，識別子を構築する．

2.3 モデリング

モデリングにおいては，教師あり機械学習アルゴリズムに文書ベクトルを適用することで識別器を構築する．本研究では，k 最近傍法（以後，kNN と表記）[4]，サポートベクターマシン（以後，SVM と表記）[5]，ナイーブベイズ分類器（以後，NBC と表記）[6]，決定木 [7]を用いている．これらのアルゴリズムを選択した理由は，SPAM 検出などの同様の研究で使用されているためである．以下，各アルゴリズムおよび利用方法について簡単に説明する．

（1）kNN

kNN は代表的な分類アルゴリズムの 1 つである．単純で効果的な手法であるが，広く使われている．未知データと最も近い k 個内のデータ中でもっとも多いクラスを回答とする．基本的には訓練フェーズが不要である．特徴ごとに値の範囲が異なる場合，距離に対し値の大きな特徴の影響が大きくなる．その点を抑えるため，本研究では，特徴に最小最大正規化を施した後，kNN を適用する．

（2）SVM

SVM は，線形入力素子を利用し，分類を行う機械学習手法である．これは，インスタンスと特徴値を表す多次元世界にプロットされたデータの間に境界を形成する面として示される．SVM の目的は，超平面と呼ばれるこの境界面を作成することである．これにより，空間が分割され，両側に均質なパーティションが作成される．SVM 学習は，最近傍学習と線形回帰モデリングの両方の側面を組み合わせたものととらえることができ，この組み合わせは非常に強力であることから，SVM は，非常に複雑な関係をモデル化することが可能となる．線形分類を実行することに加え，カーネルトリックを使用して非線形分類を効率的に実行し，それらの入力を高次元の特徴空間に暗黙的にマッピングすることができる．

本研究においては，kNN と同様に最小最大正規化を施したデータに SVM を適用する．

（3）NBC

NBC は，ベイズの定理を分類問題に適用するための単純かつ最も一般的な方法である．NBC はシンプル，高速，そして非常に効果的でという長所をもち，ノイズが多くデータが欠落している場合にうまく機能する傾向がある．また，トレーニングに必要な例が比較的少ない場合でも，非常に多い場合でもうまく機能する．NBC における学習モデルは，クラスごとの特徴の出現を示す尤度表となる．NBC はカテゴリカルデータに適しているため，本研究においては，語句品詞組の頻度を，0 の場合は’N’，1 以上の場合は’Y’の 2 値に変換する．また，品詞ごとの頻度，件名と本文の類似度については，最小最大正規化を施した後，0.5 以上は’Y’，0.5 未満は’N’の 2 値に変換する．これらを元に尤度表を作成する．

（4）決定木

決定木は強力な分類器であり，木構造を利用して，機能間の関係と潜在的な結果をモデル化する．このアルゴリズムは，多くの問題で機能する汎用の分類器になる．数値データのみでなく，カテゴリカルデータと混在している場合でも利用できる．また，学習で得られたモデルは数学的な背景がなくても解釈できるという利点を持つ．

本研究においては，2.2 で得られた文書ベクトルをそのまま適用する．

2.4 識別

識別フェーズにおいては，モデリングによって得られた 4 つの学習モデルを未知データに適用し，該当ユーザか否かの判別を行う．kNN，SVM に適用する際には，最小最大正規化を行うが，このとき，最大値および最小値は，訓練データの最小値最大値を用いる．NBC を適用する際も，訓練データの値を元に 2 値化を行い，識別を行う．決定木においては，特に前処理を行わず，そのまま識別を行う．

また，各モデルの結果のうち，訓練データに対する精度の高い 3 つのモデルにより多数決を行った結果に基づく識別も行う．訓練モデルにおける Kappa 係数の高いものを選択する理由は，訓練データと未知データとの差が小さい，

訓練データに近いほうがよいと判断するためである．なお，Kappa 係数については，3 章にて述べる．

3. 評価

3.1 概要

前述した識別手法の有用性を評価するために，本研究ではオープンデータといくつかのツールを用いた評価実験を行った．以下，評価実験の内容について述べる．

(1) データセット

本研究では，データセットとして UC Berkeley Enron E メール分析プロジェクト[8] において，1 つ以上の分類ラベルが付与された Enron Email データセット[9]の一部を利用した．Enron Email データセットとは，CALO プロジェクトによって収集および作成されたデータであり，約 150 人のユーザ，主に Enron 社の上級管理職からのデータがフォルダにまとめられている．本データは，連邦エネルギー規制委員会によって Web 上で公開されている． UC Berkeley Enron E メール分析本プロジェクトは，バークレー大学の ANLP（Applied Natural Processing Language Processing）コースで行われたものであり，約 1700 の電子メールが学生によってラベル付けされている．

本研究においては，この約 1700 のメールの中から，文字数が少ないメールを取り除いた 1286 個のデータおよび，最も頻繁に現れる送信者を識別対象者としてデータを作成した．また，単語数としては，文頭から 100 単語を基準とし，文の途中で 100 単語目になる場合は，文末までを含めた．なお文末の判断には，ピリオドのほか，改行を用いている．

このデータに対し， 4-Fold クロスバリデーションを実行した．データの割合を表 1 に示す．

表 1　データセットの構成

Fold	Fold1	Fold2	Fold3	Fold4	All
対象者	237	237	237	236	947
対象者以外	172	173	172	172	689
Total	409	410	409	408	1286

(2) ハイパーパラメータの設定

各学習モデルにおけるハイパーパラメータの設定は表 2 に示すとおりである．なおメールの文書化にあたっては，単語ごとの品詞の切り出しには TreeTagger[9]を，機械学習については R[10]のライブラリを用いた．

表 2　ハイパーパラメータの設定

kNN	距離：ユークリッド距離，k=3
SVM	カーネル関数："radial"，$e(-3\|u-v\|^2)$
NBC	Laplace =1
決定木	ノードの最小アイテム数=7 分割の閾値= 0.01

(3) 評価基準

実験結果の評価としては，式(1)で求める精度（ACC）および式(2)で求める Kappa 係数[11]を用いる．

$$ACC = M / N \quad 式(1)$$

$$Kappa\ 係数 = (ACC - P_{期待}) / (1 - P_{期待}) \quad 式(2)$$

式(1)において， M は，実際のクラスラベルと識別したクラスラベルが等しいデータ項目数を示し， N は，テストデータセット内のデータ項目数を示す．

式(2)において， $P_{期待}$は，訓練データにおいて無作為に選んだ際の各カテゴリの出現確率であり，式(3)によって求める．

$$P_{期待} = \Sigma_k (p_{正解=k} \cdot p_{予測=k}) \quad 式(3)$$

精度はよく知られている標準的な測定基準であるため，多くの研究で使用されるが，実際のクラスラベルの割合に影響を受けやすい．たとえば，あるクラスラベルを持つデータ項目の割合が他よりも大きい場合，このクラスのデータ項目の数がデータ項目の数に近くなるため，精度が高くなりやすい．この点を考慮するため，偶然の一致について調整した Kappa 係数も評価として本研究では用いる．

3.2 実験結果

各機械学習アルゴリズムを使用したときの訓練データの精度と Kappa 係数を表 3 に示す．この結果に基づいて多数決に用いるモデルを選択する．SVM と NBC を使用する際の精度と Kappa 統計は他のものよりも高いため，これらを用いる． k NN と決定木においては，決定木が k NN よりも精度が高いケースは 3 回である．また，精度の平均は決定木が高い．この結果から，決定木を用いる．

表 3　訓練データの精度と Kappa 係数

Fold	基準	kNN	SVM	NBC	決定木
Fold1	ACC	0.803	0.975	0.874	0.822
	Kappa 係数	0.573	0.948	0.745	0.625
Fold2	ACC	0.823	0.977	0.879	0.802
	Kappa 係数	0.619	0.953	0.753	0.583
Fold3	ACC	0.811	0.976	0.883	0.813
	Kappa 係数	0.591	0.951	0.761	0.612
Fold4	ACC	0.800	0.983	0.893	0.832
	Kappa 係数	0.565	0.965	0.781	0.647
平均	ACC	0.809	0.978	0.882	0.817
	Kappa 係数	0.587	0.954	0.760	0.617

次に，4 つの機械学習アルゴリズムおよび投票によるテストデータの検出結果を平均と SD とともに表 4 に示す．

表 4　テストデータの精度と Kappa 係数

項目	基準	kNN	SVM	NBC	決定木	投票
Fold1	ACC	0.650	0.831	0.814	0.760	0.848
	Kappa	0.220	0.646	0.619	0.489	0.683
Fold2	ACC	0.680	0.832	0.798	0.780	0.844
	Kappa	0.285	0.652	0.590	0.541	0.679
Fold3	ACC	0.680	0.826	0.824	0.763	0.873
	Kappa	0.278	0.635	0.641	0.503	0.734
Fold4	ACC	0.676	0.806	0.809	0.772	0.828
	Kappa	0.266	0.591	0.608	0.522	0.640
平均	ACC	0.672	0.824	0.811	0.769	0.848
	Kappa	0.262	0.631	0.615	0.514	0.684
SD	ACC	0.014	0.012	0.011	0.009	0.018
	Kappa	0.029	0.028	0.021	0.022	0.039

3.3 評価・考察

実験の結果，SVM，NBC および投票は，精度が 0.8 を超えたことから有用となる可能性は高い．また，Kappa 係数も 0.6 以上であり，精度は十分といえる．また，投票がもっとも高くなったことから複数のモデルの結果を統合することは有用である可能性は高い．決定木についても，精度の平均が 0.7，Kappa 係数の平均が 0.5 を上回ったことから，SVM などよりも低いものの，有用である可能性は高い．一方 kNN においては，精度の平均が 0.67 であるのに対し，Kappa 係数は 0.26 と低い．このことから，kNN は他手法に比べ，有用ではない．

決定木が SVM などと比べて低い理由としては，決定木が，一部の特徴間の関係を用いているのに対し， NBC は関係に焦点を合わせないものもすべての特徴を利用しており， SVM はすべての特徴間の関係に焦点を当てているためと考えられる．このことから，一部の特徴を利用するのではなく，すべてを利用するほうが良いと考えられる．

kNN についていえば，訓練データとテストデータとの差が大きい．kNN は単純でトレーニングデータに依存するため，この点が結果に影響を与えたと考えられる．

次に投票について考察を加える．本方法の性能は，最も良かったが，1 つの機械学習アルゴリズムを使用する場合よりも絶対的な優位性は言えず，組み合わせ方によっては結果が悪化する場合がある．実際，決定木の代わりに kNN を用いた場合，kNN 単独より性能は向上するが， SVM 単独よりも低下する．すなわち，低い性能のアルゴリズムが高い性能のアルゴリズムの負荷となっている．このことから，提案した投票方法は機械学習アルゴリズムの選択にその精度が依存するといえる．

また，本研究では，特徴として，品詞分布および件名と本文の類似度を用いている．その有用性を評価するために，同一データに対し，語句品詞の出現情報のみを NBC [14]で識別した結果と比較を行った．結果，本提案手法は平均で 0.05 下回った．このことから，今回追加した特徴が識別精度向上に貢献していない可能性がある．

今後の課題としては，第 1 に実験結果の更なる詳細分析があげられる．第 2 に実験の分析結果に基づき，本提案手法を修正することが挙げられる．現時点では，(1)用語の選択方法を見直し，(2)照合方法の変更が挙げられる．(1)については，現時点では出現位置以外での単語の絞り込みを行っていない．結果，ノイズが含まれる可能性がある．これに対し，出現頻度やストップワードなどを考慮した単語の絞り込みを行う．(2)については，シソーラスや単語ベクトルなどのリソースを用いて同義語との照合を行う．さらに，これらの変更を加えた手法を，より大きなデータセットによって評価することがあげられる．

4. おわりに

近年，BEC は SPAM メールのような大きな問題になってきている．セキュリティ会社や政府機関などが注意を促すとともに防衛対策が示されているが，それらのほとんどはコンピュータ化されていない．本研究では，そのような対策のコンピュータ化の第一歩として，対策の 1 つである「普段と違うメール」の検出を，過去に受信したメールから機械学習を用いて構築した学習モデルを用いて行う手法を提案した．

本手法では，対象者をあらかじめ定義し，機械学習を利用して対象者から電子メールが送信されるかどうかを予測するモデルを作成する．機械学習を適用するために，各電子メールを，件名と本文の類似度，品詞分布，本文の最初の部分の極貧仕組み出現からなる特徴ベクトルに変換する．識別フェーズにおいては，各機械学習単独での結果の他，訓練データにおいて精度の高いものによる投票を用いた識別を行う．

本手法の有用性を，オープンデータセットとツールを使って評価した．機械学習としては，kNN，SVM，NBC，決定木を用いた．評価実験の結果，SVM，NBC，投票においては精度の平均が 0.8 以上，Kappa 係数の平均が 0.6 以上と高い結果を得た．このことから，本手法は有効である可能性を示せた．しかし，データサイズは大きくないため，より大きなデータでの検証が必要である．

今後の課題として，シソーラスなどのリソースを活用することや，犯罪捜査などに用いられる著者特定の手法の適用することで，本手法を改善することがあげられる．さらに．これらの変更を加えた手法の有用性を，より大きなデータセットを使って評価することも今後の課題である．

参考文献

[1] IPA，“ビジネスメール詐欺「BEC」に関する事例と注意喚起（続報）～ あらゆる国内企業・組織が攻撃対象となる状況に～”，https://www.ipa.go.jp/files/000068781.pdf，(2018)

[2] FBI IC3，“BUSINESS E-MAIL COMPROMISE THE 12 BILLION DOLLAR SCAM”，https://www.ic3.gov/media/2018/180712.aspx，(2019.6.12 アクセス)

[3] Trend Micro，“ビジネスメール詐欺に関する実態調査 2018”，https://www.trendmicro.com/ja_jp/about/press-release/2018/pr-20180814-01.html，(2018)

[4] T.M. Cover and P.E. Hart，“Nearest neighbor pattern classification.”, IEEE Trans. Inform. Theory, IT-13(1), 21-27(1967)

[5] E. Bernhard, M. GUYON Isabelle, N. VAPNIK Vladimir，“A training algorithm for optimal margin classifiers”， Proceedings of the fifth annual workshop on Computational learning theory, ACM, 144-152(1992)

[6] M. E. Maron:“Automatic Indexing，An Experimental Inquiry”，J.ACM, Vol.8, 404-417(1961)

[7] Quinlan, J. R.，“Induction of Decision Trees, Machine Learning”，Vol.1, No.1, 81-106(1986)

[8] UC Berkeley Enron Email Analysis, http://bailando.sims.berkeley.edu/enron_email.html (2019.6.12 アクセス)

[9] Enron Email Dataset, https://www.cs.cmu.edu/~./enron/ (2019.6.12 アクセス),

[10] Helmut Schmid，“Probabilistic Part-of-Speech Tagging Using Decision Trees”. Proceedings of International Conference on New Methods in Language Processing, 44-9，（1994），http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/ (2019.6.12 アクセス)

[11] The R Foundation，https://www.r-project.org/，(2019.6.12 アクセス)(1997)

[12] J.R. Landis and G.G. Koch : The measurement of observer agreement for categorical data, Biometrics， 33，159-174(1977)

[13] Brett Lantz: Machine Learning with R second edition, Packt Publishing, Birmingham, West Midlands, England, 84-124(2013)