

ニューラルネットワークを用いたスパムメールフィルタリングの 方式検討と実装

三宅一平[†] 尾花賢[†]

法政大学情報科学部[†]

1. まえがき

既存のスパムメールのフィルタリングの手法にはサポートベクターマシン (SVM) やベイジアンフィルタリングがあり, いずれも精度と再現性を用いた指標であるF値が90%以上で高い性能である. ベイジアンフィルタリングは分類器の生成, 推論時間は非常に高速だが, F値は他の手法より劣ってしまう. また, SVMは一般的には線形分離不可能なデータの分類に適していない. しかし, メール本文の語彙は非常に多様であるため, スパムメールのフィルタリングにおいては線形分離不可能なデータも分類可能であるニューラルネットワークがより適切な手法だと考えられる. また, ニューラルネットワークの学習時間は既存手法に比べると遅いが, 複雑な関数を近似できるため既存手法より高い検出性能であることが期待される. しかし, 現状ニューラルネットワークを用いた手法の研究は進んでいない.

本手法ではニューラルネットワークを用いたスパムフィルタリングを提案し, 層の数やニューロン数, dropout層のニューロンを無効にする確率などのパラメータを様々な組み合わせで試行することで最適なネットワークモデルを検討した. また, 複数のBag of Words手法のベクトル化手法を比較し, 入力データの次元数における性能の変化についても評価し, 最適な分類器を検討・実装・評価した.

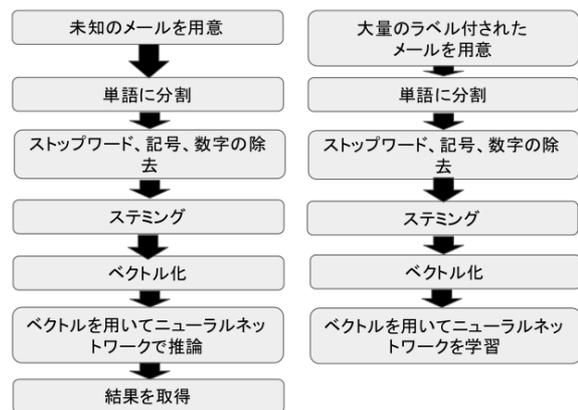


図1 提案手法の処理の流れ

2. 提案手法

本研究で提案する手法は学習と推論の2フェイズで構成される. 学習は, (1)大量のメール本文, ラベルを用意, (2)メールのトークン化, (3)トークンから特徴ベクトルを生成, (4)特徴ベクトルと対応したラベルを用いてニューラルネットワークを学習させる. 推論のフェイズは, (1)メールを用意, (2)メールのトークン化, (3)トークンから特徴ベクトルを生成, (4)特徴ベクトルを用いてニューラルネットワークで分類, (5)結果を取得する. 図-1は学習, 推論の処理の流れを示している. 以下, 各処理を詳細に説明する.

はじめに, 各メールの本文に題名を加えたドキュメントに単語の分割, ストップワードや記号, 数字の除去, ステミング, URLの処理を行う.

次に, 下処理し単語に分割されたドキュメントから特徴ベクトルを生成する. 特徴ベクトル化

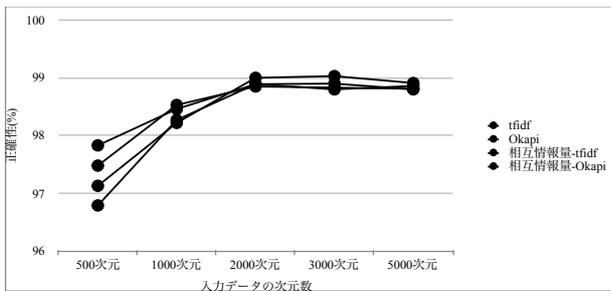


図2 ベクトル化手法ごとの比較結果

手法として以下の4つの方式の比較を行った。

- ・単語選定に相互情報量, 重み付けはTFIDF
- ・単語選定に相互情報量, 重み付けはOkapi BM25+
- ・単語の選定, 重み付けどちらもTFIDF
- ・単語の選定, 重み付けどちらもOkapi BM25+

Okapi BM25+はTFIDFと似ているが, 短いドキュメントの単語を重要視するという特性がある。また, 相互情報量を用いた単語の選定はラベルと単語の相互情報量が高い単語を学習対象にする手法だ。比較結果を表2に示す。

これらの結果から提案方式では最良の結果が得られた手法を採用する。具体的にはラベルと単語の相互情報量の高い単語の上位2000語を学習対象にして, 重み付け手法はBag of Words方式のOkapi BM25+を採用する。

次に, 生成したベクトルをニューラルネットワークに学習・推論させる。ニューラルネットワークは全結合層, Dropout層, 活性化関数を1ブロックとし, ブロックを3層重ねた構成とする。入力データの次元数と同じ隠れ層が2層ある構成である。本手法では偽陽性を下げるためにSpamと判断する閾値を45%に設定した。

3. 評価実験

3.1 実験環境

データセットはEnron Spam datasetの1&2, 3&4, 5&6の3組を用いる。テストデータを20%, 検証データを10%, 残りの70%を訓練データとする。

また, 既存研究との比較のためにLingSpam datasetのLemmatizeフォルダを用いる。テストデータを10%, 検証データを10%, 残りの80%を訓

練データとする。Spamが全体の約17%である。

どちらも英語のデータセットであり, メールの題名, 本文で構成されており, スпамかそうでないかのラベルづけがしてある。

3.2 実験結果

テストデータを分類した実行結果を表1に示す。本研究と同じLingSpam datasetを使用しているHaoら[1]の手法と比較すると正確性が0.46%, 精度が3%向上していることから, 既存手法と比べても高い検出性能であることがわかる。

表1 実験結果

	Enron 1&2	Enron 3&4	Enron 5&6	Ling Spam
Accuracy	97.9	98.91	98.66	98.96
Precision	97.66	99.25	99.2	100
Recall	97.66	98.67	98.96	93.75
False Positive	1.91	0.82	2.17	0.0

4. むすび

本研究ではニューラルネットワークを用いたスパムメールのフィルタリングを検討・実装・評価した。各パラメータごとに試行をし, 最適な分類器を検討し実装したところ, 既存手法より優れた検出性能となった。しかし, 分類器の最適なパラメータはデータセットによって異なることがあり, データセットの内訳などを考慮して決める必要があることがわかった。

今回はBag of Words方式の重み付けでベクトル化したが, 分散表現を用いて時系列を考慮できる分類器の再帰型ニューラルネットワークを学習させることでより性能が向上することを期待する。

参考文献

[1] Stephen Robertson, Hugo Zaragoza, "The Probabilistic Relevance Framework: BM25 and Beyond", Foundations and Trends in Information Retrieval archive Volume 3 Issue 4, April 2009 Pages 333-389