Vol.2020-SPT-36 No.44
2020/3/3

半教師ありトピックモデルによるセキュリティレポートの分類の評価方法について

杉本健太[†1]　長田侑樹[†1]　瀧田愼[†2]　古本啓祐[†3]　白石善明[†1]
高橋健志[†3]　毛利公美[†4]　髙野泰洋[†1]　森井昌克[†]

概要：自組織に応じたセキュリティ・インシデントの事前対策ならびに事後対応においてセキュリティレポートは有用だが，文書の数は増加する一方であり，自組織に関連する情報を探し出すことは容易ではない．トピックモデルなどを利用した文書の分類手法も提案されているが，セキュリティレポートの分類性能に対する定量的な評価方法は定まっていない．本稿では，半教師ありトピックモデルによるセキュリティレポートの分類性能の評価を定量的に行うことを目的とし，ケーススタディとして教師無しトピックモデルの一つである LDA と半教師ありトピックモデルの一つである SeededLDA を利用したクラスタリングの結果を用いてシード単語ごとの F 値を算出した．その結果，半教師ありトピックモデルによる文書分類の結果を評価する際には F 値が有効であることが確認された．

キーワード：LDA（Latent Dirichlet Allocation），SeededLDA，F 値

An evaluation method for classification of security reports using a Semi-supervised Topic Model

SUGIMOTO KENTA[†1]　OSADA YUKI[†1]　TAKITA MAKOTO[†2]
FURUMOTO KEISUKE[†3]　SHIRAISHI YOSHIAKI[†1]　TAKAHASHI TAKESHI[†3]
MOHRI MASAMI[†4]　TAKANO YASUHIRO[†1]　MORII MASAKATU[†1]

Abstract: Although security reports are useful in responding to security incidents according to each organization, the number of security reports is increasing, and it is not easy to find information related to each organization. Furthermore, a quantitative evaluation method for the classification performance of security reports has not been determined. In this paper, we aim to quantitatively evaluate the classification performance of security reports using a semi-supervised topic model. As a case study, we calculate the F value for each seed word using the results of clustering using LDA and SeededLDA. As a result, we show that the F value was effective when evaluating the results of document classification using a semi-supervised topic model.

Keywords: LDA（Latent Dirichlet Allocation），SeededLDA，F-measure

1. はじめに

昨今，日本だけでなく各国において，企業や組織を標的としたサイバー攻撃が高度化・多様化している．自組織に応じた事前対策を講じることとインシデント発生後の迅速な事後対応に備えることが必要不可欠である．事前対策ならびに事後対応にあたっては，サイバー攻撃の動向や影響および対策などを記載したセキュリティレポートが有用である．ただし，セキュリティレポートは多くのセキュリティベンダーが定期的に発行しているため，文書の数は増加する一方であり，自組織に関連する情報を探し出すことは容易ではない．さらに，セキュリティベンダーによって，レポートのフォーマットは様々であり，文書に付与されているラベルの基準も異なっている．そのため，セキュリティレポートを記載内容に基づいて統一された基準で分類し，セキュリティレポートから得られる情報を集約することは，インシデント対応を円滑に行う際に有益である．

文書分類の方針として，分類基準をあらかじめ定めた上で文書を振り分けるカテゴリ分類と，内容が類似した文書を集めるクラスタリングが挙げられる．セキュリティレポートのカテゴリ分類に関する研究として，Ayoade らは ATT&CK フレームワーク[1]に基づく分類手法を提案している[2]．また，セキュリティ・レポートのクラスタリングに関する研究として，永井らは話題誘導するトピックモデル（SeededLDA）[3]を利用することで，セキュリティレポートからサイバー攻撃の流行している分野や時期を分析する手法を提案している[4]．トピックモデルはクラスタリングにおける次元削減のアプローチとして注目されている．

†1 神戸大学
Kobe University
†2 兵庫県立大学
University of Hyogo
†3 情報通信研究機構
National Institute of Information and Communications Technology
†4 岐阜大学
Gifu University

ⓒ2020 Information Processing Society of Japan

Blei らにより提案された LDA(Latent Dirichlet Allocation)はトピックモデルの一種であり, Jagarlamudi らにより提案された SeededLDA はシード単語を設定することによりトピックの内容を誘導可能な手法である．永井らの手法では，セキュリティに関する用語を SeededLDA のシード単語として設定してセキュリティレポートの学習を行い，出力されるトピック分布やトピック所属確率の結果を二次元上に可視化している．しかし，セキュリティレポートの分類性能に対する定量的な評価方法は定まっていない．

本稿では，半教師ありトピックモデルによるセキュリティレポートの分類性能の評価を定量的に行うことを目的とする．ケーススタディとして教師無しトピックモデルの一つである LDA と半教師ありトピックモデルの一つである SeededLDA を利用したクラスタリングの結果を用いてシード単語ごとの F 値を算出した．その結果，半教師ありトピックモデルによる文書分類の結果を評価する際には F 値が有効であることが確認された．

2. トピックモデル

2.1 LDA（Latent Dirichlet Allocation）

容量が大きくかつ内容が一様ではない大量の文書から，目的の情報を獲得するための統計的モデリング手法の一つとしてトピックモデルがある．トピックモデルでは，文書を単語の集合と捉え，単語は単語集合の背後に存在するトピックから生成されると仮定されている．本稿では教師なしトピックモデルの 1 種である LDA（Latent Dirichlet Allocation）を用いる．LDA に文書を入力し，文書ごとのトピックの構成比率であるトピック分布とトピックごとの単語の比率である単語分布を学習し，文書の各トピックへの所属確率を推測する．

LDA では文書に現れる単語の共起情報からトピックに単語を関連付けている．しかし，LDA によって学習されたトピックがユーザにとって意味があるものとは限らない．

2.2 SeededLDA

トピックモデルに付加情報を与えて，ユーザが理解しやすいトピックになるよう誘導する SeededLDA が提案されている．付加情報として，誘導したい話題に対して，それに関連した単語を集めて，シード単語集合を設定する．SeededLDA の学習の際には，設定したシード単語やそれに類似する単語を含む文書が同じトピックに含まれやすくなるように誘導される．SeededLDA の利用により，教師付きデータを用意することなく，ユーザの指向を反映した学習が可能となり，ユーザにとって意味のあるトピックが生成されることが期待される．

3. クラスタリングの評価と F 値

3.1 教師なし学習でのクラスタリングの評価方法

クラスタリングの代表的な評価方法として，クラスタ内の凝集性を示すクラスタ内距離二乗和や，クラスタ内の凝集性に加えてクラスタ間の離散性で評価する指標で評価する Pseudo F がある．また，n次元ベクトルで表される文書の類似度を評価するコサイン類似度などの方法がある．

ただし，これら評価方法は，それぞれのクラスタ内での文書の類似度を評価する指標であるため，クラスタ間の関係は評価できない．そのため，ユーザが意図した文書が特定のクラスタにどの程度誘導されているのかを評価することは難しい．

3.2 分類器の評価方法に用いられる F 値について

機械学習や統計学における分類問題に用いられる分類器モデルの性能評価には，F 値が用いられることが多い．あるデータの分類を考えるとき，分類器の予測結果と真の結果に基づいて，次のように結果を分類する．（真の結果，予測した結果）が（正，正）のとき TP (True Positive)，（正，負）のとき FN（False Negative），（負，正）のとき FP（False Positive），（負，負）のとき TN（True Negative）とし，分類されたデータ数を求める．その数に対し，適合率（Precision）と再現率（Recall）を以下の式で求める．

$$\text{適合率} P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{1}$$

$$\text{再現率} R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{2}$$

F 値は適合率 P と再現率 R の調和平均で定義される

$$\text{F 値} = 2 \times \frac{P * R}{P + R} \tag{3}$$

F 値は 0 以上 1 以下の範囲で表され，1 に近いほど分類器としての性能が高いことを示している．

3.3 本研究のアプローチ

本研究では，教師なし学習である LDA とシード単語を設定する半教師あり学習である SeededLDA の比較を行う．シード単語を設定することで，ユーザの意図を反映したトピックが生成されているかを F 値により検証する．また，設定するシード単語による誘導のされ方の違いを検証し，半教師ありトピックモデルによる文書分類の結果を評価する方法として，F 値を用いることの有用性を確認する．

ⓒ2020 Information Processing Society of Japan

F 値を求めるために，シード単語に関連する文書，すなわちあるトピックに誘導したい文書の集合を真の結果の正のラベルとする．そして，LDA により推測された文書の各トピックへの所属確率に対するクラスタリングの結果からシード単語に関連する文書を含む文書のクラスタを予測した結果の正のラベルとする．

F 値は分類の正確性，つまり分類器の性能を示す指標であることから，F 値の大小や数値の偏りをみることで，シード単語に関連のある文書がどのように分類されたかを定量的に検証できると考えている．

4. 評価方法の提案

本章では，LDA 及び SeededLDA による分類結果を F 値を用いて評価する方法を提案する．

まず，前処理として，各文書に対して，日常的につかわれる単語であるストップワードを除去し，文書をベクトル化する．文書のベクトル化には，単語の出現回数をもとにする BoW（bag-of-word）を用いる．

LDA 及び SeededLDA による分類結果の F 値の計算は以下の手順で算出する．

Step1. 学習前のラベル付け

SeededLDA で用いるシード単語の集合を設定する．全文書に対して，シード単語およびシード単語に関連のある単語を含んでいる文書に正のラベルを付与し，それ以外の文書に負のラベルを付与する．このラベルを目標ラベルと定義する．

Step2. 入力と学習

LDA および seededLDA に文書ベクトルを入力し学習を行う．全文書の各トピックへの所属確率を得る．

Step3. クラスタリング

Step2 で得た所属確率に基づいて，K-means アルゴリズムによりクラスタリングを行う．これにより，全文書に対して，LDA 利用した場合と SeededLDA を利用した場合の２パターンのクラスタ番号を得る．

以降，LDA を利用した場合のクラスタを LDA クラスタ，SeededLDA を利用した場合を SeededLDA クラスタと呼ぶ．

Step4. クラスタリング後のラベル付け

クラスタリングの結果，目標ラベルが正である文書を含むクラスタをシード単語に関連のあるクラスタと定義し，そのクラスタに含まれる文書に正のラベルを付与し，それ以外のクラスタに含まれる文書に負のラベルを付与する．このラベルを予測ラベルと定義する．

表 1　「mirai」に関連する文書中の単語

シード単語	mirai
0	mirai
1	mirai_malware
2	that_mirai
3	mirai_botnet

表 2　「mirai」をシード単語にしたときの F 値

クラスタ番号	LDA クラスタ	SeededLDA クラスタ
0	なし	なし
1	なし	なし
2	なし	0.0895522
3	0.0555556	なし
4	なし	なし
5	なし	なし
6	なし	なし
7	なし	なし
8	0.0943396	なし
9	なし	なし
10	なし	0.173913
11	0.0338983	なし
12	なし	0.252427
13	なし	なし
14	なし	なし
15	なし	なし
16	なし	なし
17	なし	なし
18	0.510638	なし
19	なし	1

Step5. F 値の導出

（目標ラベル，予測ラベル）のラベル付けに基づいて，各クラスタで TP(正,正)，FP(正,負)，FN(負,正)，TN(負,負)の文書数を数えて，F 値を算出する．

以上のように導出した F 値を用いて，シード単語に関連する文書がどのようにクラスタリングされたかを評価する．

5. ケーススタディ

5.1 データセットと実装環境

データセットとして 2017 年度にセキュリティ会社 8 社（Symantec，TrendMicro，Cisco，Barracuda，Druva，Arbor，FireEye，Paloalto）が発行したセキュリティレポート 875 件

ⓒ2020 Information Processing Society of Japan

を使用する．

本実験で使用する LDA 及び SeededLDA は，Python 3 の GuidedLDA モジュール[5]を用いて実装し，シード単語をセットしない場合を LDA，シード単語をセットする場合を SeededLDA として扱う．

5.2 期待される結果

SeededLDA では，シード単語を設定することでユーザの意向に沿ったトピック誘導が行えるという特性から，シード単語に関連した単語を含む文書がクラスタリングによって集まると考えられる．そのため，LDA と SeededLDA の F 値を比べると，SeededLDA のほうが大きくなることが期待される．

5.3 F 値が高くなる場合

シード単語に関連した文書，すなわち正の目標ラベルが付与された文書が，少数のクラスタに集まる場合に F 値が高くなる．これは，関連文書が少数のクラスタに集まるような特徴的な話題（トピック）に関連の深い単語をシード単語として設定した場合が考えられる．

シード単語に「mirai」を設定した場合の各クラスタの F 値を検証する．表 1 に全文書中に現れた「mirai」に関連する単語を示す．また，表 2 に LDA クラスタ，SeededLDA クラスタの各クラスタの F 値を示す．「mirai」は 2017 年に登場した新種のマルウェアである．

SeededLDA クラスタ 19 において，F 値が 1 であり，中身を精査すると，目標ラベルが正であり予測ラベルも正である文書が 42 件，目標ラベルも予測ラベルも負でありクラスタのメイントピックに全く関係のない文書が 1 件含まれていた．一方で，LDA クラスタの中で F 値が最大値である LDA クラスタ 18 では，目標ラベルが正であり予測ラベルも正である文書が 23 件，目標ラベルと予測ラベルが一致しない文書が 48 件含まれていた．

クラスタごとの F 値の最大値で比較すると，SeededLDA の F 値の最大値のほうが大きい．これはシード単語を設定することで，それに関連のある文書が特定のクラスタに集まったことを意味する．以上の結果から，LDA に比べて，SeededLDA ほうが，シード単語のセットにより分類の精度は向上したといえる．

6. まとめ

本稿では，教師なしトピックモデルの一つである LDA と半教師ありトピックモデルの一つである SeededLDA を用いたセキュリティレポートの分類性能の評価を定量的に行うことを目的とし，評価の指標として F 値を用いることを提案した．SeededLDA で設定するシード単語に関連する文書を正の目標ラベルとし，LDA 及び SeededLDA を利用したクラスタリングの結果を用いて，F 値を算出した．ケーススタディとして，2017 年度に発行されたセキュリティレポートを対象としてクラスタリングを行い，半教師ありトピックモデルによる文書分類の結果を評価する指標として，F 値が有効であることが確認された．

謝辞 本研究は国立研究開発法人情報通信研究機構の委託研究「機械学習に基づくサイバー攻撃情報分析基盤技術の研究開発」により行われた．

参考文献

[1] MITRE：ATT&CK™(online), available from 〈https://attack.mitre.org/〉
[2] Ayoade, G., Chandra, S., Khan, L., Hamlen, K., & Thuraisingham, B.: Automated threat report classification over multi-source data. Proceedings, pp.236-245, 4th IEEE International Conference on Collaboration and Internet Computing, Nov, 2018.
[3] Tatsuya Nagai, Makoto Takita, Keisuke Furumoto, Yoshiaki Shiraishi, Kelin Xia, Yasuhiro Takano, Masami Mohri, Masakatu Morii, ``Understanding Attack Trends from Security Blog Posts　Using Guided-topic Model``, Journal of Information Processing, 2019, Vol.27, p.1–8.
[4] Jagarlamudi, J., Daumé III Hal., and Udupa, R., “Incorporating lexical priors into topic models,” Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pp.204-213, April 2012.
[5] “GuidedLDA” https://github.com/vi3k6i5/GuidedLDA,(参照 2020-01-29)

ⓒ2020 Information Processing Society of Japan