RFC を対象にしたクラスタリングと教師有り学習分類器を用いた関心度フィルタの評価

学修番号 23745118
氏名 高嶋隆一

概要

IETF [1] が発行する技術文書として RFC [2] が存在する。RFC はインターネット上で公開されているが、その技術はインターネットに関連する多岐な分野に及ぶ。 本レポートではまず、非階層的クラスタリングを用いてデータをカテゴリ毎に分割する事を試みる。また、教師データとして RFC に自身の関心の有無の情報を付加した教師有り学習を行い、テストデータに対して自身の関心の有無が判定できるかの評価を行う。

1 趣旨

エンジニアや研究者は日々発行される最新の技術文書や論文から情報を入手し、その知識を更新していく必要がある。一方で、発行される文書の数は多く、全ての文書に対して目を通した上で自己に関係が深いものを選別するのは困難である。

本レポートでは技術文書の一例として RFC を対象として、機械学習の手法を用いた文書の選別を検証する。

2 挑戦・学び

2.1 挑戦

技術文書に対する機械学習による分類について、実際に Python を用いた下記の実装を行いその手法の有効性を検証する。

- 非階層的クラスタリング[3]を用いた文書のカテゴリ毎の分割
- 教師有り学習分類器を用いた関心度フィルタ

2.2 学び

機械学習による技術文書の分類に関する有効性と、用いる機械学習アルゴリズムの適不適について学ぶ。

3 内容本体

3.1 非階層的クラスタリングを用いた文書のカテゴリ毎の分割

3.1.1 対象データの取得

今回の非階層的クラスタリングの対象としては RFC 全体のうち、XML 化されており 2023 年 12 月までに発行されている 836 の RFC を対象とした。

rfc-editors.org 検索画面から条件を満たす RFC8656 以降を含む 2019 年から 2023 年を対象として検索を行い、XML ファイルを取得する。

3.1.2 データの読み込みと前処理

Python3 を用いて pandas[4]のデータフレームとして、RFC 番号、タイトル、アブストラクト、キーワード、著者名の羅列を該当の RFC の XML タグから抽出する。

その後、spaCy[5]の言語モデルを用いて Bag of Words(以下 BoW)[6], TF-IDF[7]の手法を用いてテキストのベクトル化を行う。

3.1.3 非階層的クラスタリング

エルボー法[8]、シルエット分析[9]を用いたクラスタ数の推定を行う。

その後に推定されたクラスタ数で非階層的クラスタリングを行う。実行結果としてクラスタ毎の RFC のサンプル抽出、BoW の単語別頻度、TF-IDF の単語別の値を参照し、生成されたクラスタの特徴を評価する。

3.2 教師有り学習分類器を用いた関心度フィルタ

3.2.1 対象データの取得

3.1.1 で取得したデータを用いる。 利用するフィールドについては同様である。

上記に加え、自身で interest フィールドを追加した。このフィールドは、「興味があり読む」ものを 1 とし、「興味が無く読まない」ものを 0 としたものである

3.2.2 アルゴリズム毎の分類精度の比較

アルゴリズム毎の精度を測定し最適なアルゴリズムを選択する。今回は下記 3 つのアルゴリズムを比較する。

- ナイーブベイズ[10]分類器
- ロジスティック回帰[11]分類器

- サポートベクトルマシン[12]分類器（以下 SVM）

比較の観点としては分類精度、計算機上の実行時間を用いる事とし、各アルゴリズムのモデル作成時にはグリッドサーチ[13]により最適なパラメータを求めるものとする。

また、教師データとテストデータはそれぞれのアルゴリズムで同一のものを用い、データ全数 836 に対して教師データとテストデータの比率をホールドアウト法により 1 対 1 の比率で分割するものとする。

3.2.3 関心度フィルタの検証

前項で最も優れた結果を残したアルゴリズムを用いた分類器による関心度フィルタの検証を行う。

実環境を想定し、最新のRFCに対して過去学習内容から関心度の有無を判定する。今回は、2023 年 10 月から 12 月までの 3 ヶ月分 46 個の RFC をテストデータとする。教師データしては 2023 年 9 月以前のものを用い、学習期間による分類精度の差異を確認する。

4 実験結果と評価

4.1 非階層的クラスタリングを用いた文書のカテゴリ毎の分割

4.1.1 クラスタ数の推定

BoW, TF-IDF でベクトル化されたテキストに対するエルボー法の評価結果を図 1, 図 2 に示す。

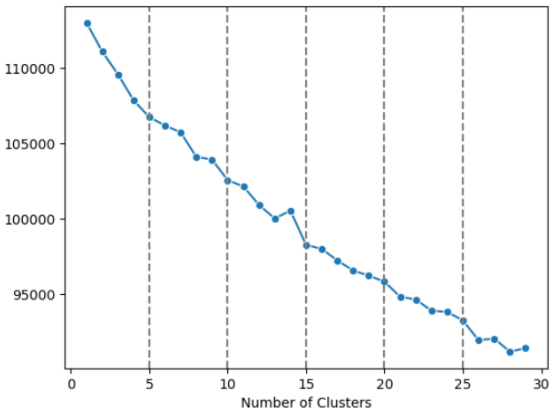

図 1. エルボー法評価結果 BoW

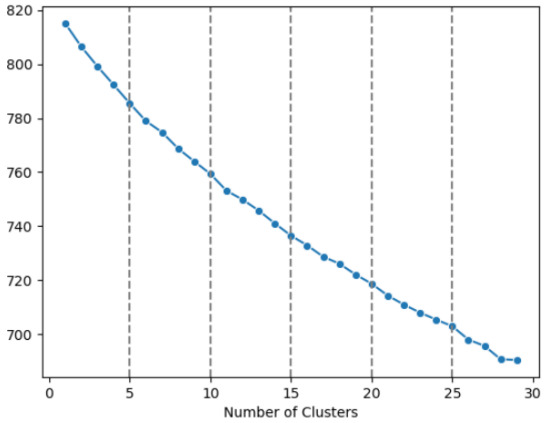

図 2. エルボー法評価結果 TF-IDF

エルボー法では BoW に関してはクラスタ数 13, 14 近傍の角度が大となり[図 1]、TF-IDF では 28 近傍が大となっている[図 2]。

次に BoW, TF-IDF でベクトル化されたテキストに対するシルエット分析を実施した結果のクラスタ数とシルエット係数についての評価結果を表 1 に示す。

表 1. シルエット分析結果

クラスタ数	シルエット係数	
	BoW	TF-IDF
5	0.048	0.014
6	0.056	0.016
7	0.051	0.017
8	0.013	0.019
9	0.050	0.021
10	0.033	0.022
11	0.050	0.025
12	0.046	0.026
13	0.053	0.028
14	0.040	0.028
15	0.009	0.031
16	-0.005	0.032
17	-0.003	0.034
18	-0.001	0.035
19	0.001	0.034
20	0.002	0.037
21	-0.002	0.037
22	-0.002	0.040
23	0.002	0.042
24	0.003	0.039
25	0.016	0.041

26	0.030	0.044
27	0.009	0.043
28	0.004	0.047
29	0.010	0.045

シルエット解析では BoW に関しては小数点 2 桁未満の差だがクラスタ数 13 が極大になっており、TF-IDF ではほぼ単調増加であり特徴のあるクラスタ数は見られない。

エルボー法、シルエット分析の結果総合した結果、クラスタ数 14 として非階層的クラスタリングを実施する。

4.1.2 非階層的クラスタリング実行結果

非階層的クラスタリングの結果の特徴をまとめたものを表 2 に示す。

項番 0、1、3、6、7、9、10、12、13 の様に特徴的に技術分類毎にクラスタリングされたものも存在するが、項番 2、4、5、8、11 の様に複数の技術分類が混合されるクラスタが存在する結果となった。

表 2. クラスタ毎の特徴

項番	単語の特徴	推定される分類
0	rtc, sctp, session 等	RTP をトランスポートに用いた各種プロトコルの技術標準
1	certificate, eap, x.509 等	証明書やそれを用いた標準
2	link, cache, proxy 等	統一性なく複数の分野が混合
3	path, pce, wavelength	MPLS の PCE(Path Caluclation Element) 関連の技術標準
4	ipv6, plane, detnet, icn 等	IPv6 が関連する複数の分野が混合
5	tls, quic, 1.3, hash, eddsa 等	TLS や暗号化が関連する複数の分野が標準
6	dns, dnssec, zone, nsec3	DNS やドメインに関するセキュリティ技術標準
7	model, yang 等	YANG モデルそのものや、YANG を用いた各種プロトコルのデータモデル
8	pen, eap, diameter 等	暗号化が関連する複数の分野が混合
9	mpls, segment, routing 等	SR-MPLS, SRv6 等 Segment Routing
10	dot, signal, threat 等	DDoS 関連
11	detnet, dscp, intent 等	統一性なく複数の分野が混合
12	token, authentication 等	認証や認証を必要とするトランスポートの技術標準
13	http, cache, cbor 等	比較的モダンな HTTP のメッセージ標準

4.2 教師有り学習分類器を用いた関心度フィルタ

4.2.1 アルゴリズム毎の分類精度の比較

表 3 に各アルゴリズムのグリッドサーチ結果を示す。ホールドアウト法を用いたモデル作成時には、ロジスティック回帰が優れた ROC AUC スコアを示している。

表 3.各アルゴリズムのグリッドサーチ結果

アルゴリズム	ROC AUC スコア	最適なパラメータ
ナイーブベイズ	0.900	α:0.1
ロジスティック回帰	0.917	C:1.0
カーネル SVM	0.898	C:10.0, γ:1.0, カーネル:RBF

次に、表 4 に教師データ、テストデータに関する比較を示す。実行時間は JupyterNotebook のマジックワード%%time を用いた CPU 時間、実時間であり参考値とする。実行結果から、各アルゴリズム共に教師データの分類精度に比較してテストデータに分類精度が低く、過学習を起こしている事が分かる。

今回の検証では、テストデータに対する分類精度が最も高く、実行時間についても他のアルゴリズムと比較して特に大きな実行時間を必要しないナイーブベイズを関心度フィルタに用いる事とする。

表 4.各アルゴリズムの分類精度の比較

アルゴリズム	分類精度		CPU times /Wall time
	教師データ	テストデータ	
ナイーブベイズ	0.976	0.852	1.43s /1.26s
ロジスティック回帰	0.864	0.797	2.4s /730ms
カーネル SVM	0.995	0.840	6.51s /1min 20s

4.2.2 ナイーブベイズ分類器を用いた関心度フィルタの検証

前項の結果から、ナイーブベイズ分類器を用いて関心度フィルタの検証を行った結果を表 5 に示す。教師データ、テストデータに関しては 3.2.3 に示した方式に従うものとする。

表 5.ナイーブベイズ分類器の学習期間毎の分類精度

教師データ	分類精度	
	教師データ	テストデータ
過去 3 箇月, 57 個	0.982	0.696
過去 6 箇月, 90 個	0.989	0.696
過去 9 箇月, 127 個	0.992	0.761
過去 12 箇月, 163 個	1.000	0.739
過去 24 箇月, 356 個	0.983	0.761
過去 36 箇月, 612 個	0.979	0.804
過去 46 箇月, 790 個	0.980	0.783

この結果から、テストデータに対する分類精度が前項のランダム抽出した 1:1 比率のホールドデータ法での学習時の 0.852 と比較して、極端に悪い値となっている。一方で教師データに対しては最大 1.000 と過学習がより進んでいる結果となっている。

4.2.3 追加検証: 時系列による学習時のデータの偏り

前項では、時系列で最新のデータに対して過去のデータによる学習を用いた判定を行った場合、ランダム抽出時よりも分類精度が劣るという結果になった。

RFC は技術の標準化文書であり、発行時期により文書の技術分類の偏りが発生する可能性がある。また時間経過につれて新規の技術分類が発生する。 その為、今回テストデータとして採用した 2023 年 10 月から 2023 年 12 月の 3ヶ月の 46 個が偏った内容である、もしくは新規の内容である場合には、ランダムに教師データとテストデータを抽出した場合に比較して悪くなる可能性がある。

上記のデータの偏りの発生という仮説の検証の為、全項目比較結果のデータの最長比較となる「教師データ 46ヶ月(790 個): テストデータ 3ヶ月 (46 個)」の比率でテストデータをランダム抽出し、5 回試行した結果を表 6 に示す。

表 6. 最長学習時間の比率を用いた
ランダム抽出データによる比較

試行回数	分類精度	
	教師データ	テストデータ
1	0.972	0.848
2	0.971	0.826
3	0.976	0.826
4	0.968	0.957
5	0.975	0.804
平均値	0.972	0.852

試行回数毎にばらつきはあるものの、分類精度はアルゴリズム比較時の値 0.852 に近い値となった。

上記の結果から、4.2.2 で仮定した今回テストデータとして採用した 2023 年 10 月から 2023 年 12 月の 3ヶ月の 46 個が偏った内容である、もしくは新規の内容であるという仮説は蓋然性があると考えられる。

5 考察

5.1 非階層的クラスタリングを用いた文書のカテゴリ毎の分割

5.1.1 今回実装に関する考察

4.1.2 の結果より、特徴的に技術分類毎にクラスタリングされたものも存在するが、複数の技術分類が混合されるクラスタが存在する結果となった。

複数の技術分類が混合されるクラスタが存在する理由としては、個々のRFCが目的としている技術の単語の他に利用している要素技術の単語も入ってくる事が考えられる。他の技術に依存する割合が低い様な技術分類は特徴あるクラスタとして識別されるが、汎用的な技術、例えば暗号化や IPv6 等の要素技術を使うものは要素技術をもってクラスタが生成されてしまう。

これらの問題は、クラスタ数をある程度増やせばクラスタ毎の特徴は確立すると思われるが、一方で 1 つの技術分類が要素技術により複数のクラスタに分かれてしまう事にも繋がる。この事から RFC の様に相互に依存関係がある技術文書について、単純に非階層的クラスタリングを用いて分類する事は困難であるという知見を得た。

5.1.2 改善手法

改善手法として、ひとつにはストップワードの調整が考えられる。しかし、RFC の場合には頻出単語もその技術単語そのものに関する RFC 等も存在する為、単純に除去する事は難しい。 当初設定したテーマである非階層的クラスタリングについての有効性を否定する形になるが、今回の RFC の様な文書の分類の場合

- 非階層的クラスタリングと階層的クラスタリングを併用し、その文書の特徴を捉える
- 教師有り学習を用いる分類器のアルゴリズムを利用する

等の形式が有効であると考えられる。

5.2 教師有り学習分類器を用いた関心度フィルタ

5.2.1 今回実装に関する考察

今回は迷惑メールフィルタを参考にし、RFCに対して自身で興味有り、興味無しの 2 値のフィールドを付加して関心度フィルタを実装した。分類器に用いるアルゴリズムとしてはナイーブベイズ、ロジスティック回帰、SVM とした。各分類器に対して分類精度を比較した結果、ナイーブベイズが最も優れた分類精度と実用的な実行速度を示した。

続いて、ナイーブベイズ分類器を用いた関心度フィルタを実装したが、テストデータとして検証時の最新 3 ヶ月のデータを、教師データとしてそれより過去のデータを用いた結果、前項のベンチマーク時に較べ劣る分類精度となった。

関心度フィルタの評価対象とした RFC は技術文書であり、発行時期により文書の技術分類の偏りが発生する可能性がある。偏りが発生しているかどうかの判定を行う為、前項と同じデータ比率でランダムに抽出したデータに対して繰り返し分類精度の評価を行った結果、アルゴリズム比較時と同様の値が得られ偏りが発生している事が検証できた。

今回のRFCの判定精度としては、分類精度が 0.852 前後という事で比較的実用に近い関心度フィルタが実装できたと考えられるが、実用的には改善が必要になる。

5.2.2 改善手法

実際に利用するケースを想定すると、時系列によるデータの偏りは日常的に発生すると考えられる。これらへの改善手法としては下記が考えられる。

5.2.2.1. インタラクティブなインタフェースの導入による継続的な学習

迷惑メールフィルタの様に実際の判定が間違っていた場合には再度学習を行わせる。偽陰性の場合には定期的に「興味なし」と分類されたものの一覧を見る事になるが、リアルタイムに発行された全ての RFC に目を通すよりは時間が節約できると考えられる。

5.2.3 データウィンドウの使用

古いデータによる過学習への対応として、一定期間、例えば過去 6 ヶ月、1 年の様に現在時刻を元にした一定期間の過去の記録を消去する事が考えられる。これは、時期による偏りへの対処だけではなく、自己の関心度に変更があった場合にも有効であると考えられる。

5.2.4 過去のデータのサンプリング

古いデータによる過学習への対応として、過去のデータ程サンプリングレートを低くして学習させるというアプローチも考えられる。一方で、RFC 程度の母数ではサンプリングによる偏りも考えられる為、実際の有効性については検証する必要がある。

参考文献

[1] JPNIC, IETF とは, https://www.nic.ad.jp/ja/basics/terms/ietf.html, visited on 2024 (ウェブ参照)

[2] JPNIC, インターネット 10 分講座:RFC, https://www.nic.ad.jp/ja/newsletter/No24/090.html, Jul. 2003. (visited on 2024) (ウェブ参照)

[3] 追川修一, 非階層的クラスタリング, pp27-37, データ分析実践特論 第 2 回クラスタリング(1), 2023

[4] pandas: powerful Python data analysis toolkit, https://github.com/pandas-dev/pandas/blob/main/README.md, visited on 2024

[5] spaCy: Industrial-Strength Natural Language Processing IN PYTHON, https://spacy.io/, visited on 2024

[6] 追川修一, Bag of Words, pp7-9, データ分析実践特論 第 3 回クラスタリング(2), 2023

[7] 追川修一, TF-IDF:単語の重要度の評価, pp10-11, データ分析実践特論 第 3 回クラスタリング(2), 2023

[8] 追川修一, エルボー法, pp25, データ分析実践特論 第 2 回クラスタリング(1), 2023

[9] 追川修一, シルエット分析, pp36-37, データ分析実践特論 第 2 回クラスタリング(1), 2023

[10] 追川修一, アルゴリズム:ナイーブベイズ, pp3-15, データ分析実践特論 第 6 回分類(2), 2023

[11] 追川修一, ロジスティック回帰量, pp5-15, データ分析実践特論 第 4 回分類(1), 2023

[12] 追川修一, アルゴリズム: SVM, pp22-31, データ分析実践特論 第 6 回分類(2), 2023

[13] 追川修一, パラメータのチューニング:グリッドサーチ, pp30, データ分析実践特論 第 6 回分類(2), 2023