E-029

ニュース原稿におけるラベル共起情報に基づくラベル平滑化手法

Label smoothing based on co-occurrence information of labels in news articles

安田 有希† 後藤 淳†
Yuki Yasuda Jun Goto

1. はじめに

放送局では日々大量のニュース原稿が制作され、データベースに蓄積されている。蓄積された原稿データを有効活用するためには、原稿の内容を示すラベルなどのメタ情報の付与作業が必要である。しかしながら、大量の原稿データに人手でラベルを付与することには多大な労力を要する。したがって、分類作業の自動化は重要な課題である。

ニュース原稿は扱うトピックが多岐に渡るため、内容を記述する複数のラベルを付与する必要がある。これは、マルチラベルテキスト分類[1]というタスクで表現できる。

マルチラベル分類分野にはモデルが不均衡データセット[2]の影響を受けやすいという問題が存在する。不均衡データセットとは、データセットにおける各ラベルの出現数が著しく異なるデータセットであり、多くのラベルを扱うニュース原稿のデータセットではラベルの不均衡が特に起こりやすい。例として、本研究で作成したデータセットのラベル出現頻度を図 1 に示す。

このようなラベル分布のデータセットを用いた学習では、出現頻度が低いラベルと対応する文書サンプルが少ないことから、モデルは低頻度ラベルとそれに対応する入力に対して過剰に適合してしまい、低頻度ラベルに関する精度が低下する。本研究では、この問題に取り組むためにデータセットにおけるラベル同士の共起情報を用いたラベル平滑化手法を提案する。

2. 関連研究

ニューラルネットワークを用いたマルチラベル分類では、ラベルを$y \in \{0,1\}$というベクトルとして表したハードターゲットを教師データとして使用することが一般的である。しかしながら、ハードターゲットをそのまま使用することで、モデルがラベルの頻度の影響を直接受けてしまう。その影響を緩和するためにネガティブラベル、すなわちラベルに対応する y の値が 0 のラベル値に一定の値を分配し、教師データを平滑化する手法が取られている。ただし、このような古典的なラベル平滑化では、すべてのネガティブラベルに対して一定の値が分配され、マルチラベル分類の特性を加味しているとは言えない[3]。

マルチラベル分類では、複数のラベルが文書に割りあてられることから、実際にはラベルは相互に意味的な関係を持っていると考えられ、ネガティブラベルであってもポジティブラベルに近いものもあれば、そうでないものも存在すると考えられる。たとえば、「スポーツ」というラベルと「野球」というラベルは包含関係にあり、データの中に同時に出現する可能性が高い。そこで、各ラベルの共起情報をもとにラベルを平滑化する手法が取られている[4]。しかしながら、既存手法ではラベルの PPMI スコアを足し合わせた数値を平滑化に利用しており、学習サンプルごとにそれぞれのラベルへの平滑化の度合いは一定である。

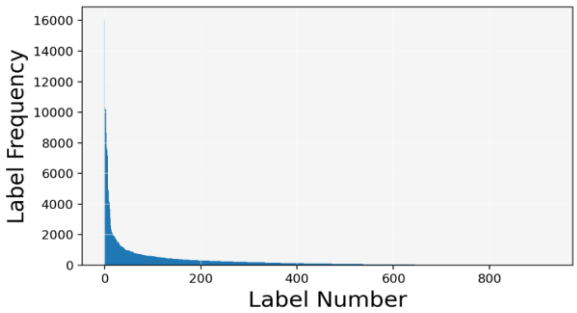

図 1　各ラベルの出現頻度

3. 提案手法

本提案手法における学習方法を図 2 に示す。本提案手法では、分類モデルとは別に各ラベルの平滑化の度合いを出力する平滑化モデルを構築する。平滑化モデルへの入力として、教師ラベルおよび教師ラベルと共起するラベルを用いる。平滑化モデルの教師データはハードターゲットの教師ラベルである。ソフトターゲットは、1 エポックごとに平滑化モデルの出力とハードターゲットを組み合わせて構築される。一方で、分類モデルは構築されたソフトターゲットをもとに学習する。なお、平滑化モデルは学習時のみに利用され、推論時には分類モデルのみを用いて推論する。

3.1 平滑化モデル

平滑化モデルは線形変換と self-attention 機構、Convolutional Neural Network(CNN)によって構成される。平滑化モデルへの入力はラベル列である。ラベル列は教師ラベルと教師ラベルと共起するラベルから構成される。例えば、「スポーツ」と「野球」というラベルがポジティブのサンプルを学習する場合、平滑化モデルへの入力は「スポーツ、野球、…引退、賭博」というラベル列となる。この時、「引退」や「賭博」などのラベルは学習データ中で「スポーツ」や「野球」というラベルと共起している関連ラベルである。入力されたラベルはまずラベル埋め込みよって任意の次元の特徴量に変換される。次にラベルの埋め込み表現は self- attention 機構に入力され、それぞれの類似性が加味された特徴量へと変換される。そして、self-attention 機構の出力を CNN と次元数調整用の線形変換機構に入力し、ラベル数に相当する次元の特徴量に変換する。最後に sigmoid 関数を使って 0~1 の値に調整したスコアを平滑化モデルの出力とする。平滑化モデルは、Binary Cross

† NHK 放送技術研究所　NHK Science & Technology Research Laboratories

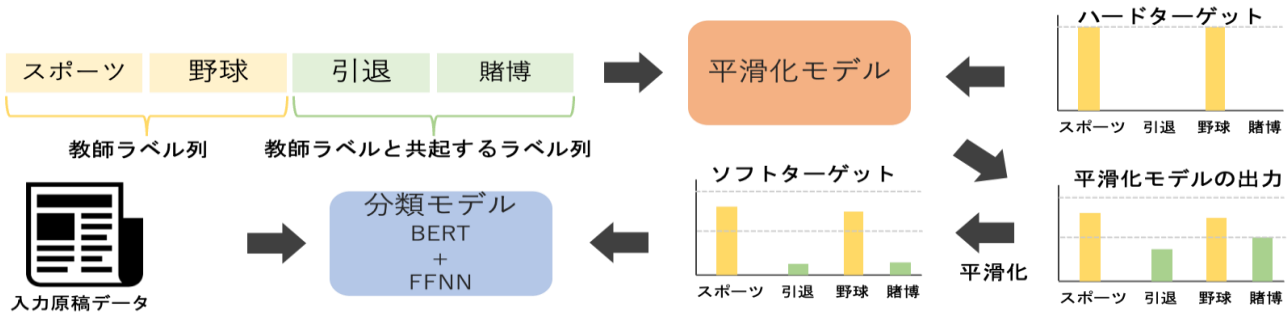

図 2　本研究の学習方法

Entropy 関数を使用して計算された損失値を最小化することによって学習する。

3.2 平滑化手法

分類モデルが学習するためのソフトターゲットは、疎なベクトルであるハードターゲットと平滑化モデルの出力を足し合わせることによって構築する。教師ラベルがポジティブである場合は要素の値を減少させ、教師ラベルがネガティブである場合は、要素の値を増加させる。ソフトターゲットy_k'は以下の通りに定式化される。

$$y_k' = y_k + \alpha(2\delta_{k,l} - 1)s(l')_k,$$

ここで、$\delta_{k,y}$はディラックのデルタ関数、kとlはラベル、$s(l')$は平滑化モデルの出力、αは平滑化モデルのスコアをスケーリングするハイパーパラメータである。

3.3 分類モデル

本研究では、文書を分類するモデルとして 2 層の Feed-forward neural network (FFNN)を備えた BERT を採用した。文書$x \in X$は Word Piece によってトークン化され、BERT へ入力される。BERT の出力である[CLS]ベクトルは FFNN に入力され、ラベル数次元のスコアベクトルに変換される。シグモイド関数を介したスコアの値が 0.5 を超えるラベルをポジティブラベルとして出力する。

4. 実験

提案手法の有効性を検証するために、評価実験を行った。本実験の評価指標として、micro-f1 と macrof-1 を設定した。

4.1 比較手法

本実験では、以下の 2 つの手法を比較手法として用いた。1 つ目が平滑化を使わずに分類モデルをそのままハードターゲットで学習した手法（平滑化なし）である。2 つ目が教師ラベルの PPMI スコアを足し合わせ、静的に平滑化の度合いを決定する手法（既存手法）である[4]。

4.2 データセット

本実験では、複数のトピックラベルを持つという特徴とラベルの出現頻度が不均衡であるという特徴を備えたデータセットを構築した。文書Xは NHK NEWS WEB で実際に公開された記事データであり、記事数は 48,493 である。一方で、ラベルは$y \in \{0,1\}^{1052}$とし、全ての記事Xの内容を十分に記述できる一般名詞から構成されている。例として、記事のジャンルを指す「スポーツ」というラベルや具体的な競技名を指す「野球」というラベルが挙げられる。また、それぞれの記事データに付与されるラベルはアノテータ 2 人の合意を持って決定された。本実験ではデータセットをランダムで学習データ 34,647 と開発データ 3,849、テストデータ 9,998 に振り分けた。

4.3 実験結果

表 1 にラベルセット全体の結果を示す。さらに、低頻度ラベルの結果を表 2 に示す。表中の各スコアは異なるランダムシードを用いて 3 回実験を行った結果の平均を記載している。表 1、2 によると提案手法は全体のラベルおよび低頻度ラベルに関してベースラインを上回っていることがわかる。また、micro-f1、macro-f1 共に提案手法が既存手法を上回っていることから提案手法が既存手法に比べて幅広いラベルを当てられていることを示している。

表 1　ラベル全体の実験結果

Method	Macro-f1	Micro-f1
平滑化なし	0.352	0.709
既存手法	0.398	0.709
提案手法	**0.409**	**0.720**

表 2　低頻度ラベルの実験結果

Method	Macro-f1	Micro-f1
平滑化なし	0.109	0.442
既存手法	0.120	0.493
提案手法	**0.146**	**0.516**

5. おわりに

ニュース原稿のマルチラベルテキスト分類において低頻度ラベルの精度が低いという課題を示した。その課題に対するアプローチとしてラベルの共起情報をもとに平滑化モデルを使ってラベルの平滑化を行う手法を提案した。実験結果より、提案手法の有効性が認められた。今後の展望としてさまざまなモデル、データセットで評価実験を実施し、手法の汎用的な有効性を確認する。

参考文献

[1] T. Ggorios, I. Katakis, “Multi-label classification: An overview.”, IJDWM, vol3, no.3, pp.1-13 (2007).
[2] H. Haibo, E. A. Garcia, “Learning from imbalanced data.”, IEEE Transactions on knowledge and data engineering, vol.21, no.9, pp.1263-1284 (2009).
[3] B. Guo, S. Han, X. Han, H. Huang, T. Lu, “Label confusion learning to enhance text classification models.”, AAAI, vol.35. no.14, pp. 12929-12936 (2021).
[4] 安田 有希, 石渡 大智, 宮﨑 太郎, 後藤 淳, “マルチラベル分類における共起情報を用いたラベル平滑化手法”, 音声言語情報処理 (SLP), vol.2021, no.30, pp.1-6 (2021).