Weighted Asymmetric Loss for Multi-Label Text Classification on Imbalanced Data

安田有希 [1]　宮崎太郎 [1]　後藤淳 [2]
[1]NHK 放送技術研究所　[2]NHK 財団

{yasuda.y-hk,miyazaki.t-jw}@nhk.or.jp　goto.j-fw@nhk-fdn.or.jp

掲載号の情報

31 巻 3 号 pp. 1166-1192.
doi: https://doi.org/10.5715/jnlp.31.1166

概要

マルチラベルテキスト分類 [1](MLTC) は、自然言語処理技術を現実世界に適用するうえで重要なタスクの一つである。MLTC は事前定義されたラベルから適切なラベルサブセット $y^{(i)} = \{y_1^{(i)}, y_2^{(i)}, \cdots, y_N^{(i)}\} \in \{0,1\}^N$ を文書 $x^{(i)} \in X$ に割り当てるタスクであり、そのために関数 $f : \chi^{(i)} \rightarrow y^{(i)}$ を学習することが目的である。ここで、X は文書のセット、N は事前定義されたラベルの総数、$\chi^{(i)}$ ($\chi^{(i)} \subset \mathbb{R}^d$) は i 番目の文書 $x^{(i)}$ から作成された d 次元の特徴を表す。

MLTC 用のデータセットでは、概念の粒度が細かく、かつ多数のラベルを扱うためラベルの分布が不均衡になる場合が多い。そのため、MLTC のデータセットにおいてはラベル分布がロングテールとなることが一般的であることが指摘されている [2]。このような分布の特徴を持つデータではデータ中に出現する頻度が低いラベル、すなわち低頻度ラベルに対する分類精度が低くなりやすい。これは、モデルが低頻度ラベルを負例として学習することが多いことに起因すると考えられる。

本研究では、MLTC における不均衡データの精度改善を目的とした損失関数である Weighted Asymmetric Loss（WASL）を提案する。WASL は、ラベルの出現頻度の差による影響とラベル空間の大きさに起因する負例由来の損失値による影響を緩和することを狙いとしている。ラベルの出現頻度の差を是正するために Class-balanced loss[3](CBL) をもとにした重みを導入した。また、負例由来の損失値を緩和するために、Asymmetric Loss[4](ASL) から着想を得た重みを導入した。さらに、ラベルの共起情報を用いたラベル平滑化を導入し、低頻度ラベルのサンプル数の少なさを補うことで精度改善を図った。

表 1　実験結果 (一部抜粋)

	Reuters-21578	
Methods/Metrics	macro-f1	micro-f1
RoBERTa w/ BCE	0.5862	0.9027
RoBERTa w/ CBL	0.5886	0.8990
RoBERTa w/ ASL	0.6443	0.9043
RoBERTa w/ WASL (proposed)	**0.6691**	**0.9110**

提案手法の有効性を検証するために、MLTC のベンチマークデータを用いた評価実験を実施した。実験結果より、提案手法がベースライン手法の精度を統計的に有意に上回ったことが確認された。表 1 に実験結果の一部を抜粋して示す。ここで BCE は Binary Cross-Entropy Loss を指す。また、追加実験としてラベル分布が均衡なデータセットを用いた実験を実施した。追加実験の結果から提案手法とベースライン手法に有意な差はなく、提案手法が不均衡データセットに対して有効であることを確認した。

参考文献

[1] Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification: An overview. **International Journal of Data Warehousing and Mining**, Vol. 3, No. 3, pp. 1–13, 2007.

[2] Sophie Henning, William Beluch, Alexander Fraser, and Annemarie Friedrich. A survey of methods for addressing class imbalance in deep-learning based natural language processing. In **Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics**, pp. 523–540, 2023.

[3] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In **Proceedings of the IEEE/CVF conference on computer vision and pattern recognition**, pp. 9268–9277, 2019.

[4] Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In **Proceedings of the IEEE/CVF International Conference on Computer Vision**, pp. 82–91, October 2021.