異なる粒度が混在する教師データを対象とした階層型マルチラベル分類モデルに関する研究

1X20C112-6　宮島健悟
指導教員　後藤正幸

1. 研究背景と目的

マルチラベル分類は，各入力データに対して複数のクラスラベルを予測するタスクである．実データセットでは，複数のラベル間に意味的な階層構造が存在することが多く，このような階層構造を考慮したモデルを適用することが望ましい．ラベル間の意味的な階層構造を考慮したマルチラベル分類モデルの 1 つとして，Multi-label Box Model[1]（以下，MBM）が提案されており，教師データに全ての階層のクラスラベルが付与されている場合に対して有効性が示されている．実用例として，医療や生物の画像データにおいて MBM が適用されている．これに対して本研究では多数のユーザーが記事の投稿と共にラベルを付与することが出来る投稿 Web サービスのデータへの適用について考える．このデータでは，先の例と同様に付与されるラベル間に意味的な階層構造があるが，ラベルの付与基準が異なる投稿者が任意で記事にラベルを付与するため，全ての階層のラベルが付与された記事のデータはほとんど存在しない．そのため，このようなデータをそのまま用いて MBM の学習を行うと，ラベルの予測精度が低下する恐れがある．

そこで本研究では，Bidirectional Encoder Representations from Transformers[2]（以下，BERT）を導入することにより，不足している階層のラベルを補完して MBM を学習するフレームワークを提案する．これにより，不足していた階層のラベルが補完された状態で MBM を学習することが出来るため，ラベルの予測精度の向上が期待できる．本稿では，一部の階層のラベルが不足しているデータに対して従来手法および提案手法を適用した場合の精度を比較した評価実験を通じ，提案手法の有効性を示す．

2. 準備

2.1. 階層型マルチラベル分類

マルチラベル分類とは，1 つのデータに対して複数のラベルを予測する分類問題である．その中でも，ラベル間に存在する階層構造を考慮したマルチラベル分類のことを階層型マルチラベル分類という．階層型マルチラベル分類に関する研究では，埋め込み方法や損失関数などの工夫が盛んに行われている．

2.2. Multi-label Box Model

本研究では，埋め込み方法を工夫して階層型マルチラベル分類を行う手法である MBM を扱う．MBM は，ボックス空間にデータとラベルを埋め込むことで階層型マルチラベル分類を行う手法である．ボックス空間とは，データが超立方体（ボックス）として表現された空間のことを指す．全てのラベルと入力データをボックスで表現し，入力データと該当するラベルのボックスが重なるように学習することで，ラベル間の階層構造を明示的に与えずとも，階層構造を考慮したマルチラベル分類が可能である．MBM の概要図を図 1 に示す．

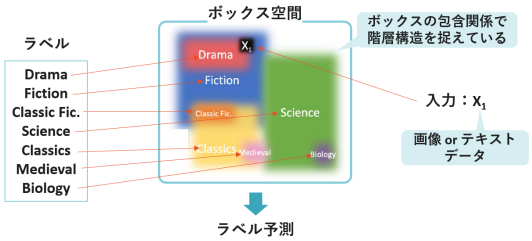

図 1: MBM の概要図 [1]

MBM では，全てのラベル $\mathcal{L} = \{\ell_1, ..., \ell_L\}$，入力データ $\boldsymbol{X} \in \mathcal{X}$ をそれぞれボックスで表現した後，両者を同じボックス空間に埋め込む．はじめに，ラベルを d 次元のボックスで表現する．この操作 $\mathrm{Box}_{\boldsymbol{\psi}} : \mathcal{L} \to \mathbb{I}^d$ を式 (1) に示す．なお，$\boldsymbol{\psi}^-, \boldsymbol{\psi}^+ \in \mathbb{R}^{L \times d}$ は学習されるパラメータである．

$$\mathrm{Box}_{\boldsymbol{\psi}}(\ell_i) := \prod_{j=1}^{d} [\psi^-_{i,j}, \psi^-_{i,j} + \log(1 + \exp(\psi^+_{i,j}))] \quad (1)$$

次に，入力データ $\boldsymbol{X}$ を d 次元のボックスで表現する操作 $\mathrm{Box}_{\boldsymbol{\theta}} = I^d \circ \mathcal{F}_{\boldsymbol{\theta}} : \mathcal{X} \to \mathbb{I}^d$ を行う．ここで，$\mathcal{F}_{\boldsymbol{\theta}} : \mathcal{X} \to \mathbb{R}^d$ はパラメータ $\boldsymbol{\theta}$ を持つニューラルネットワークであり，入力データ $\boldsymbol{X}$ を $\boldsymbol{z} = (z_1, ..., z_d) \in \mathbb{R}^d$ に変換する．$I^d : \mathbb{R}^d \to \mathbb{I}^d$ は $\boldsymbol{z}$ を d 次元のボックスで表現する関数であり，式 (2) で定義される．なお，δ はボックスの大きさを調整するハイパーパラメータである．

$$I^d(\boldsymbol{z}) := \prod_{j=1}^{d} [z_j - \delta, z_j + \delta] \quad (2)$$

最後に，全てのラベルと入力データを同じボックス空間に埋め込む．この操作を式 (3) に示す．なお，$\boldsymbol{y} = (y^{(1)}, ..., y^{(L)}) \in \{0, 1\}^L$ は正解ラベル，$\boldsymbol{\psi} \in \mathbb{R}^{L \times 2d}$ は $\boldsymbol{\psi}^-$ と $\boldsymbol{\psi}^+$ を結合したパラメータ，λ は近似ベッセル体積を計算する関数である．

$$P_{\mathrm{MBM}}(y^{(i)} = 1 | \boldsymbol{X}; \boldsymbol{\psi}, \boldsymbol{\theta}) = \frac{\lambda(\mathrm{Box}_{\boldsymbol{\psi}}(\ell_i) \widetilde{\cap} \mathrm{Box}_{\boldsymbol{\theta}}(\boldsymbol{X}))}{\lambda(\mathrm{Box}_{\boldsymbol{\theta}}(\boldsymbol{X}))} \quad (3)$$

MBM のパラメータ $(\boldsymbol{\psi}, \boldsymbol{\theta})$ は与えられる教師データ $\mathcal{D} = \{(\boldsymbol{X}_1, \boldsymbol{y}_1), ..., (\boldsymbol{X}_N, \boldsymbol{y}_N)\}$ に対して式 (4) で定義される負の対数尤度損失 L_{nll} を最小化するように学習される．

$$L_{\mathrm{nll}}(\boldsymbol{\psi}, \boldsymbol{\theta}; \mathcal{D}) = -\sum_{n=1}^{N} \sum_{i=1}^{L} \log P(y_n^{(i)} | \boldsymbol{X}_n; \boldsymbol{\psi}, \boldsymbol{\theta}) \quad (4)$$

2.3. Bidirectional Encoder Representations from Transformers

本研究では，不足している階層のラベルを補完するために，双方向の Transformer エンコーダを用いた自然言語処理モ

デルである BERT を用いる．BERT を用いて生成した文書の埋め込みベクトルのコサイン類似度を計算することで，文書同士の類似度を算出することが可能である．

3. 提案手法

3.1. 着想

MBM では，一部の階層のラベルが不足しているデータを学習に用いた場合，ラベルの予測精度が低下してしまう．この問題を解決するために，BERT を用いて求めた記事（文書）同士の類似度を利用して，不足していると考えられるラベルを MBM の学習前に補完するフレームワークを提案する．

3.2. 提案フレームワーク

本研究で提案するフレームワークでは，ラベルを補完する対象データとの類似度が最も高いデータに付与されているラベルのうち，対象データに存在しないラベルを補完する．この操作を加えることで，不足しているであろうラベルを補完することが可能になると考えられる．提案フレームワークのイメージを図 2 に示す．また，この提案フレームワークのアルゴリズムを以下に示す．

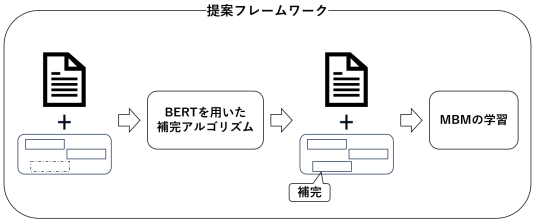

図 2: 提案フレームワークのイメージ図

提案フレームワークのアルゴリズム

Step.1 BERT を用いて全ての文書データをベクトル化し，$(\boldsymbol{x}_1, ..., \boldsymbol{x}_N)$ とする．$i = 1$ とする．

Step.2 対象データ $\boldsymbol{x}_i$ に付与されているラベルを 1 つでも含むデータを全て抽出する．

Step.3 対象データ $\boldsymbol{x}_i$ と Step.2 で抽出したデータの類似度を計算する．

Step.4 類似度が最も高かったデータのラベルのうち，対象データ $\boldsymbol{x}_i$ に不足しているラベルを補完する．

Step.5 $i < N$ であれば，$i = i + 1$ を行い，Step.2 へ戻る．

Step.6 補完後のデータを用いて MBM を学習する．

4. 評価実験

本章では，提案手法の有効性を明らかにするために，一部の階層のラベルが不足しているデータに対し，提案手法を用いた場合と従来手法を用いた場合の精度を比較する．

4.1. 実験条件

本実験では，ニュース記事とその記事のジャンルを表すラベルが含まれるデータセット 20news[3] を用いる．20news では，各記事に対し，階層構造を持つラベルが必ず 2 つ以上付与されている．実験では，ある階層のラベルが不足している教師データを再現するために，教師データの 90%に対してラベルをランダムに 1 つずつ欠損させる．なお，実験に使用するデータは 11,270 件，ラベルは 36 種類であり，文書データは 1-hot ベクトルに変換した．また MBM については，バッチサイズを 4，エポック数を 10，学習率を 0.0001，最適化手法を AdamW，$\delta = 10^{-5}$ として学習を行った．評価指標には，F 値 (F1 score)，Mean Average Precision(MAP)，Constraint Violation(CV) を用いる．ここで，F1 score は Recall（再現率）と Precision（適合率）を調和平均した値であり，値が高いほど良い性能を示す．MAP はテストデータの各クラスの平均適合率の平均を表す評価指標であり，これも値が高いほど良い性能を示す．CV は階層構造を考慮した予測を行っているかを確認する評価指標であり，値が低いほど良い性能を示す．

4.2. 実験結果と考察

実験結果を表 1 に示す．なお,「MBM（完全データ）」は全ての階層のラベルが付与された教師データによる MBM の学習を,「MBM（欠損データ）」は一部の階層のラベルが欠損している教師データによる MBM の学習を,「提案（欠損データ）」は提案手法に基づいた MBM の学習を表す．

表 1: テストデータに対する実験結果

評価指標	MBM（完全データ）	MBM（欠損データ）	提案（欠損データ）
F1 score ↑	83.27(± 1.20)	54.33(± 2.20)	**78.51(± 1.30)**
CV ↓	6.96(± 0.92)	13.51(± 1.58)	**7.27(± 0.89)**
MAP ↑	93.99(± 0.34)	91.30(± 0.71)	**92.02(± 0.57)**

表 1 より，全ての評価指標で提案（欠損データ）の方が MBM（欠損データ）よりも精度が高いことが分かる．特に，F1 score では，提案（欠損データ）は MBM（欠損データ）よりも大幅に高い精度向上が観測されたことから，提案手法の有効性が確認できる．一方で，MBM（完全データ）と提案（欠損データ）を比較した結果，提案（欠損データ）は MBM（完全データ）と精度が同等になったとは言えない．このことから，補完後のラベルは真のラベルと同一のラベルをまだ完全に付与できていないため，ラベルの補完方法に改善の余地があると考えられる．

5. まとめと今後の課題

本研究では，BERT と MBM を用いて，一部の階層のラベルが不足した教師データを対象とした階層型マルチラベル分類を行うフレームワークを提案した．また，実験では提案手法が従来の MBM を上回る精度を達成し，提案手法の有効性を確認できた．今後の課題としては，より精度を向上させるための補完方法の工夫や，提案手法の異なるドメインへの適用などが挙げられる．

参考文献

[1] Dhruvesh Patel, Pavitra Dangati, Jay-Yoon Lee, Michael Boratko, and Andrew McCallum. Modeling label space interactions in multi-label classification using box embeddings. *ICLR 2022 Poster*, 2022.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[3] Ken Lang. Newsweeder: Learning to filter netnews. In *Proceedings of the Twelfth International Conference on Machine Learning*, pp. 331–339, 1995.