マルチラベル分類への教師あり対照学習適用における ラベル共起を用いた負例への重み付け

小林 稜馬 [1]　李 吉屹 [2]
[1] 北海道大学 大学院 情報科学院 [2] 北海道大学 大学院 情報科学研究院
kobayashi.ryoma.n8@elms.hokudai.ac.jp

概要

マルチラベル分類では，ラベル間に部分的な関連が存在し，この関係性を捉えた表現学習が難しい．既存の InfoNCE 形式の教師あり対照学習は正例にのみ重み付けを行い，負例は一様に扱っていた．そこで本研究では，正例だけでなく，負例にもラベル類似度に基づく重みを導入した MCACR（Multi-Label Contrastive Attraction and Contrastive Repulsion）を提案し，負例の強さを段階付けて扱う．AAPD と RCV1-v2 において，MCACR は既存手法と同程度の Micro-F1 を維持しつつ，Macro-F1 を向上させることを確認した．

1　はじめに

マルチラベル分類（Multi-Label Classification）は，感情分析，ニュース分類，タグ推薦など幅広い応用を持つ重要な課題である [1]．マルチラベル分類では各ラベルが互いに独立ではなく，共起や包含などの依存性が存在するため，ラベル間関係を踏まえた表現学習が求められる．対照学習（Contrastive Learning）は，類似インスタンスを近づけ，非類似インスタンスを遠ざけることで表現空間を構築する枠組みであり，その教師あり変種である教師あり対照学習（Supervised Contrastive Learning; SupCon）はマルチクラス分類 において高い有効性が示されている [2]．この枠組みはマルチラベル分類にも導入されており，ラベルを一つでも共有するインスタンス対を正例とみなし，そのラベル合致度に基づいて寄与を重み付けするのが一般的である [3, 4]．

一方で，多くの手法は負例を一様に扱う．しかしマルチラベル分類では，共有ラベルのないインスタンス間にもラベル集合の部分的関連があり得るため，一様な負例処理は意味的に近い対まで分離し，表現空間をラベル分布の構造から乖離させ得る．

本研究では，マルチラベル分類における教師あり対照学習において，正例に対する既存の重み付けに加え，負例に対してラベル共起に基づく重みを導入する MCACR（Multi-Label Contrastive Attraction and Contrastive Repulsion） を提案する．これにより，負例に対してもラベル共起に基づく類似度を考慮した学習を可能にする．

本研究の貢献は次のとおりである．

1. マルチラベル分類における従来の SupCon に基づく手法が負例を一様に扱うことの限界を，マルチクラス分類との対比を通じて整理した．
2. インスタンス対のラベルペアの共起に基づき，負例の寄与を動的に調整する対照学習の拡張を提案した．
3. AAPD および RCV1-v2 において，提案手法が既存手法よりも Macro-F1 の改善を示すことを確認した．

2　関連研究

SupCon [2] は，クラス内のインスタンスを集約しつつ，異なるクラスを遠ざけることで表現を最適化する枠組みであり，マルチクラス分類で有効性が示されている．近年は，クラス不均衡やロングテールに適合させるための重み付け [5] に加え，クラスタ中心をプロトタイプとして扱う PCL [6] のような拡張も提案されている．これらの手法はマルチクラス分類での頑健性や識別性能向上に寄与しているが，ラベル間の部分的な共起を考慮する設計には至っておらず，マルチラベル分類にそのまま適用すると負例処理が過度に一様になる．

SupCon をマルチラベル分類に導入する流れでは，ラベル集合の類似度を陽に扱う損失関数が提案されている [7, 4, 3]．Jaccard 係数に基づく複数の重み付け戦略が提示され，ラベル重複の度合いに応じて正例寄与を制御する設計が検証されている [7]．ラベル集

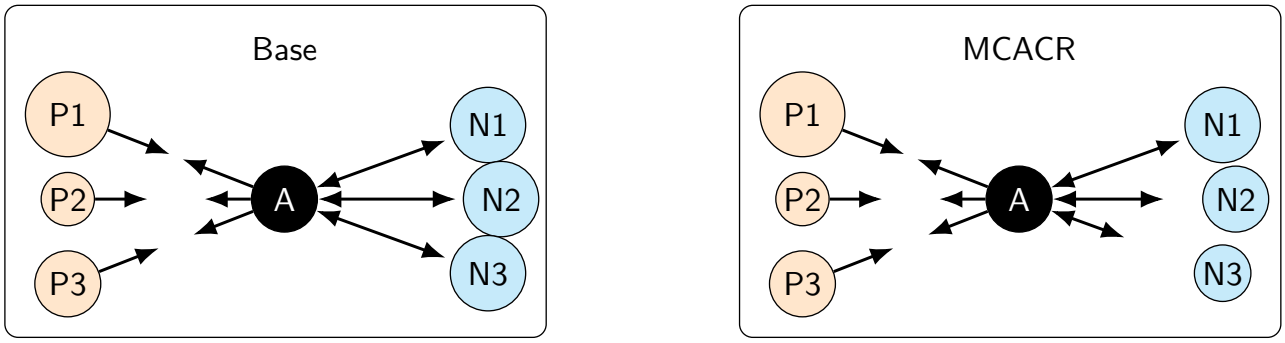

図 1 MCACR の概要. Base は正例にのみラベル類似度に基づく重みを入れ，負例は一様に扱っていたのに対し，MCACR は負例にもラベル類似度に応じた重み付けを導入する.

合の重なり率に応じた SupCon 拡張も提案され, 類似ラベルを共有するインスタンスを段階的に引き寄せる枠組みが報告されている [4]. さらに, マルチラベル分類におけるクラス不均衡やロングテール分布に対処するため, ラベルの出現頻度に応じた重み付けと Jaccard 係数を組み合わせた手法が提案されている [3]. しかし, いずれも負例の寄与は「ラベルを共有しないインスタンス」を一括で扱う設計に留まっており, ラベル集合間の意味的距離に応じて負例を差別化する余地がある. 本研究はこのギャップに着目し, 負例への重み付けを通じてラベルの共起を埋め込みに反映させる点で既存研究と異なる.

3 提案手法

3.1 定義と記法

まず，記法を導入する．バッチ内のインスタンスのインデックス集合を B，ラベル数を L とする．バッチ内の i 番目の文書を x_i とし，エンコーダ g_θ により表現 $\mathbf{z}_i = g_\theta(x_i)$ を得る．対応するラベルベクトルを $\mathbf{y}_i \in \{0,1\}^L$ で表し，$\mathbf{y}_i$ の j 成分を y_i^j とする．インスタンス i が持つラベルのインデックス集合を $Y_i = \{l \mid y_i^l = 1, l = 1, 2, \ldots, L\}$ とする．各ラベル l に対応する学習可能なプロトタイプを $\mathbf{z}_{|B|+l}$ とし，$\mathcal{P} = \{|B|+1, \ldots, |B|+L\}$ をプロトタイプのインデックス集合とする．文書埋め込みとプロトタイプ埋め込みはいずれも同一空間の表現として $\|\mathbf{z}_i\|_2 = 1$ となるように L_2 正規化して扱う [2]．さらに，インスタンス i に対する正例集合を $P_{\text{inst}}(i) = \{j \in B \setminus \{i\} \mid Y_i \cap Y_j \neq \emptyset\}$，負例集合を $N_{\text{inst}}(i) = \{j \in B \setminus \{i\} \mid Y_i \cap Y_j = \emptyset\}$ とする．また，プロトタイプに対する正例集合を $P_{\mathcal{P}}(i) = \{|B| + l \mid l \in Y_i\}$，負例集合を $N_{\mathcal{P}}(i) = \{|B| + l \mid l \notin Y_i\}$ とする．以上をまとめて，正例集合を $P(i) = P_{\text{inst}}(i) \cup P_{\mathcal{P}}(i)$，負例集合を $N(i) = N_{\text{inst}}(i) \cup N_{\mathcal{P}}(i)$ と定義する．

本研究では，入力文書 x_i から得られる表現 $\mathbf{z}_i$ の学習に着目する．学習では，以下で定義する目的関数 L を最小化する．評価では，$\mathbf{z}_i$ に分類器 h を適用して $\hat{\mathbf{y}}_i$ を出力し，分類性能を評価する．

3.2 マルチラベル分類のための教師あり対照学習

マルチラベル分類 に対する SupCon に基づく手法 [4, 3, 7] では，ラベルを一つでも共有するインスタンス対を正例とみなし，Jaccard 係数に基づいて重み付けする. 本研究では，プロトタイプも用いて対照学習を行い，このアプローチをベースラインとして次式で定義する：

$$L_{\text{Base}} = -\frac{1}{|B|} \sum_{i \in B} \frac{1}{\sum_{j \in P(i)} w_{ij}} \sum_{j \in P(i)} w_{ij} \log \frac{e^{\mathbf{z}_i^\top \mathbf{z}_j / \tau}}{\sum_{k \in (B \cup \mathcal{P}) \setminus \{i\}} e^{\mathbf{z}_i^\top \mathbf{z}_k / \tau}},$$

ここで，w_{ij} は

$$w_{ij} = \begin{cases} \dfrac{|Y_i \cap Y_j|}{|Y_i \cup Y_j|} & (j \in P_{\text{inst}}(i)), \\ 1 & (j \in P_{\mathcal{P}}(i)) \end{cases}$$

であり，τ は温度パラメータである．このアプローチでは正例のラベル関係は考慮されるが，負例を一様に扱うため，ラベル集合間に意味的な関連性を持つ負例も強制的に遠ざけてしまう可能性がある.

対照学習における正例項と負例項を分離し，それぞれを独立に最適化する CACR（Contrastive Attraction and Contrastive Repulsion）が提案されている [8]. CACR は遠い正例・近い負例を強く動かすことで，特にロングテール分布を持つマルチクラス分類での性能向上が報告されている. アンカー i に対する正例集合 $P(i)$ と負例集合 $N(i)$ に対し，正例の引力（CA）と負例の斥力（CR）を分けて計算し，そ

れらの差を損失とする：

$$L_{\mathrm{CACR}} = \frac{1}{|B|}\sum_{i\in B}\big(l_{\mathrm{CA}}(\mathbf{z}_i) - l_{\mathrm{CR}}(\mathbf{z}_i)\big),$$

$$l_{\mathrm{CA}}(\mathbf{z}_i) = \sum_{j\in P(i)} \frac{e^{d_{\tau^+}(\mathbf{z}_i,\mathbf{z}_j)}\, w^+_{ij}}{\sum_{j'\in P(i)} e^{d_{\tau^+}(\mathbf{z}_i,\mathbf{z}_{j'})}\, w^+_{ij'}}\, c(\mathbf{z}_i,\mathbf{z}_j),$$

$$l_{\mathrm{CR}}(\mathbf{z}_i) = \sum_{k\in N(i)} \frac{e^{-d_{\tau^-}(\mathbf{z}_i,\mathbf{z}_k)}\, w^-_{ik}}{\sum_{k'\in N(i)} e^{-d_{\tau^-}(\mathbf{z}_i,\mathbf{z}_{k'})}\, w^-_{ik'}}\, c(\mathbf{z}_i,\mathbf{z}_k),$$

ここで $c(\mathbf{z}_i,\mathbf{z}_j) = \|\mathbf{z}_i - \mathbf{z}_j\|_2^2$，$d_\tau(\mathbf{z}_i,\mathbf{z}_j) = \|\mathbf{z}_i - \mathbf{z}_j\|_2^2/\tau$ であり，元の CACR では $w^+_{ij} = w^-_{ij} = 1$ として一様な重みを用いる．

3.3 CACR のマルチラベル拡張

本研究では，既存のマルチラベル SupCon が正例のみにラベル集合類似度を反映し，負例を一様に扱う点に着目する．この設計は，ラベル集合間に部分的な関連が残る負例まで過度に分離し，表現空間がラベル分布の構造から乖離する可能性がある．そこで CACR を基盤とし，負例にもラベル集合に基づく非一様な重みを導入して斥力の強度をラベル共起に応じて連続的に調整することで，ラベル空間の構造を埋め込みに反映させる マルチラベル CACR（MCACR）を提案する．

正例重みには前節で用いた Jaccard 係数 w_{ij} をそのまま用い，$w^+_{ij} = w_{ij}$ とする．負例に対しては，まず各ラベルペア (l,m) について，学習データ全体での二つのラベルの共起強度を NPMI [9] を用いて測り，その後に全ラベルペアで平均する．類似度が低いほど重みが大きくなるように定義する：

$$w^-_{ij} = 1 - \frac{1}{|Y_i|\,|Y_j|}\sum_{l\in Y_i}\sum_{m\in Y_j}\frac{\mathrm{NPMI}(l,m)+1}{2}.$$

なお，$k \in N_{\mathcal{P}}(i)$ のときは $w^-_{ik} = 1$ とする．距離項が支配的にならないように，重みを $\alpha \geq 1$ で冪乗して差を強調する．最終的な損失は次式となる：

$$L = \frac{1}{|B|}\sum_{i\in B}\big(l^{\mathrm{MLC}}_{\mathrm{CA}}(\mathbf{z}_i) - l^{\mathrm{MLC}}_{\mathrm{CR}}(\mathbf{z}_i)\big),$$

$$l^{\mathrm{MLC}}_{\mathrm{CA}}(\mathbf{z}_i) = \sum_{j\in P(i)} \frac{e^{d_{\tau^+}(\mathbf{z}_i,\mathbf{z}_j)}\,(w^+_{ij})^\alpha}{\sum_{j'\in P(i)} e^{d_{\tau^+}(\mathbf{z}_i,\mathbf{z}_{j'})}\,(w^+_{ij'})^\alpha}\, c(\mathbf{z}_i,\mathbf{z}_j),$$

$$l^{\mathrm{MLC}}_{\mathrm{CR}}(\mathbf{z}_i) = \sum_{k\in N(i)} \frac{e^{-d_{\tau^-}(\mathbf{z}_i,\mathbf{z}_k)}\,(w^-_{ik})^\alpha}{\sum_{k'\in N(i)} e^{-d_{\tau^-}(\mathbf{z}_i,\mathbf{z}_{k'})}\,(w^-_{ik'})^\alpha}\, c(\mathbf{z}_i,\mathbf{z}_k).$$

正例のラベル重複度と負例のラベル非類似度の双方を重み付けに反映させることで，意味的に近い負例を過度に分離せず，遠い負例のみを強く押し出す表現学習を実現する．

表 1 データセットの統計量（$\bar{L}$ はインスタンス当たりの平均ラベル数，$\bar{W}$ は文書当たりの平均単語数）

Dataset	|train|	|val|	|test|	L	$\bar{L}$	$\bar{W}$
AAPD	37,820	6,677	11,343	54	2.4	163
RCV1-v2	19,676	3,473	781,274	91	3.2	241

4 実験

本節では，提案手法の有効性を検証するための実験について述べる．まず，実験で使用したデータセットについて説明する．次に，提案手法と比較するベースライン手法および評価指標について記述する．最後に，実験設定の詳細を示す．

4.1 データセット

- AAPD [10]: arXiv の論文概要データセット．
- RCV1-v2 [1]: Reuters のニュース記事データセット．先行研究 [3] と同様に train セットで 30 件以下の出現数を持つラベルを削除し，train セットをさらに train/val に分割して使用した．

4.2 実験設定

全手法で RoBERTa-base [11] をエンコーダとし，最大入力長 300 トークン，[CLS] トークンの最終層の隠れ状態を使用する．最適化には AdamW [12] を用い，バッチサイズは 128，学習率は $5e^{-5}$，weight decay を 0.01，ドロップアウト率を 0.1 に設定した．

本研究では，対照学習を用いない手法 (RoBERTa)，Multi-Label SupCon に基づく手法 (Base)，および MCACR を比較する．また CACR の負例重み付けを一様にした MCACR (w/o w^-) も評価する．転移学習では，射影ヘッドは使用せず，エンコーダー出力に対して線形層とシグモイド関数からなる分類ヘッドを付加した上で，binary cross-entropy（BCE）損失によりエンコーダを含むモデル全体を学習する．

学習手順は手法により以下の通りである．

- RoBERTa：対照学習を行わず，直接転移学習を行う．予備実験でスコアが 40 エポック程度までに収束したため，本稿では 40 エポック学習する．学習率は総更新数の 5% をウォームアップとするコサインスケジューラを用いた．
- Base, MCACR (w/o w^-), MCACR：
  - 対照学習：射影ヘッド（2 層 MLP，隠れ次元 256）を付加して 80 エポック学習し，学習率は総更新数の 5% をウォームアップとするコサインスケジューラを用いた．

表 2 AAPD データセットにおける分類性能（Hamming Loss: HL は $\times 10^3$）

Loss	Micro-F1↑	Macro-F1↑	HL↓
RoBERTa	73.09 ± 0.28	59.07 ± 0.62	**23.16 ± 0.30**
Base	**73.34 ± 0.17**	59.48 ± 0.29	23.24 ± 0.12
MCACR (w/o w^-)	73.02 ± 0.03	58.99 ± 0.11	23.65 ± 0.12
MCACR	73.32 ± 0.14	**59.96 ± 0.34**	23.83 ± 0.08

表 3 RCV1-v2 データセットにおける分類性能（Hamming Loss: HL は $\times 10^3$）

Loss	Micro-F1↑	Macro-F1↑	HL↓
RoBERTa	87.44 ± 0.19	74.56 ± 0.51	8.79 ± 0.12
Base	**87.78 ± 0.18**	74.76 ± 0.99	**8.57 ± 0.10**
MCACR (w/o w^-)	87.62 ± 0.27	74.46 ± 1.41	8.70 ± 0.13
MCACR	87.70 ± 0.06	**75.28 ± 0.12**	8.69 ± 0.04

– 転移学習：10 エポック学習する．学習率は線形スケジューラを用い，ウォームアップは行わない．

Base の温度は $\tau = 0.1$ とする．MCACR, MCACR (w/o w^-) では $\tau^+ = 1$ に固定し，$\tau^- \in \{0.5, 1.0, 2.0, 3.0\}$ を探索した．指数は $\alpha \in \{1, 5, 10, 20\}$ を探索した．

評価指標 評価指標として Micro-F1，Macro-F1，Hamming Loss を用いる．報告値は val Micro-F1 で最良のハイパーパラメータを選択した後，異なる 3 シードでテストセット上での性能を評価し，平均と標準偏差を算出した．

5 結果と考察

本節では，まずベースラインと提案手法の分類性能を AAPD および RCV1-v2 で比較し，続いて提案損失の主要ハイパーパラメータ（指数，距離項，負例温度）の影響を ablation で検証する．

5.1 分類性能の比較

表 2 および表 3 に，AAPD および RCV1-v2 データセットにおける各手法の分類性能を示す．

表 2 は AAPD における分類性能を示しており，MCACR は Micro-F1 を概ね維持しつつ Macro-F1 を改善する傾向が見られる．MCACR (w/o w^-) は Base に近い結果である．提案では，Jaccard 係数による正例重みと NPMI に基づく負例重みを併用し，負例の強さを段階付ける設計としているため，ラベル非類似度の差を反映した学習が Macro-F1 に寄与していると考える．一方で Hamming Loss はわずかに悪化しており，補助指標とのトレードオフが示唆される．表 3 の RCV1-v2 についても同様の傾向である．

表 4 指数 α の影響（RCV1-v2, Hamming Loss: HL は $\times 10^3$）

α	Micro-F1↑	Macro-F1↑	HL↓
1	87.36	73.99	8.85
5	87.52	74.69	8.76
10	**87.71**	**75.53**	**8.68**
20	87.68	75.25	8.69

表 5 距離に基づく重み付けの有無による性能差（RCV1-v2，Hamming Loss: HL は $\times 10^3$）

重み付け	Micro-F1↑	Macro-F1↑	HL↓
w/ $\exp(\pm d_\tau)$	**87.71**	**75.53**	**8.68**
w/o $\exp(\pm d_\tau)$	87.51	74.41	8.77

表 6 τ^- の違いによる性能比較（RCV1-v2，Hamming Loss: HL は $\times 10^3$）

τ^-	Micro-F1↑	Macro-F1↑	HL↓
0.5	87.71	**75.53**	8.68
1.0	**87.76**	75.42	**8.64**
2.0	87.56	73.70	8.73
3.0	87.69	75.12	8.66

5.2 Ablation Study

MCACR の主要ハイパーパラメータ（指数 α，距離に基づく重み付けの有無，負例温度 τ^-）について，RCV1-v2 で挙動を確認した．

表 4 は指数 α を変えた結果を示しており，$\alpha = 10$ で最良になり，ラベル類似度の相対的な差を強調することが有効であることを確認した．$\alpha = 10$ が最良となったため，距離に基づく重み付けの寄与を検証する目的で，重み計算における指数項 $\exp(\pm d_\tau)$ を定数 1 に置換した設定と比較した． 表 5 に示すように，距離に基づく重み付けが，特に Macro-F1 の改善に寄与することが確認できた．表 6 は負例温度 τ^- の結果を示しており，α に比べてスコアの変動が小さい．これはラベル類似度とのバランスによって温度の影響が相対的に抑えられているためと考えられる．

6 おわりに

本研究では，マルチラベル分類における教師あり対照学習において，負例の重み付けを導入する MCACR を提案した．AAPD および RCV1-v2 を用いた実験から，負例の強さを段階付ける設計が有効である可能性を示した．今後は，より高度な重み設計や冪乗パラメータの自動推定に取り組む．

参考文献

[1] David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li. Rcv1: A new benchmark collection for text categorization research. **J. Mach. Learn. Res.**, Vol. 5, p. 361–397, December 2004.

[2] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In **Proceedings of the 34th International Conference on Neural Information Processing Systems**, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[3] Alexandre Audibert, Aurélien Gauffre, and Massih-Reza Amini. Exploring contrastive learning for long-tailed multi-label text classification. In **Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2024, Vilnius, Lithuania, September 9–13, 2024, Proceedings, Part VII**, p. 245–261, Berlin, Heidelberg, 2024. Springer-Verlag.

[4] Pingyue Zhang and Mengyue Wu. Multi-label supervised contrastive learning. **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 38, No. 15, pp. 16786–16793, March 2024.

[5] Jianggang Zhu, Zheng Wang, Jingjing Chen, Yi-Ping Phoebe Chen, and Yu-Gang Jiang. Balanced contrastive learning for long-tailed visual recognition. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**, pp. 6908–6917, June 2022.

[6] Junnan Li, Pan Zhou, Caiming Xiong, and Steven C. H. Hoi. Prototypical contrastive learning of unsupervised representations, 2020.

[7] Nankai Lin, Guanqiu Qin, Gang Wang, Dong Zhou, and Aimin Yang. An effective deployment of contrastive learning in multi-label text classification. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Findings of the Association for Computational Linguistics: ACL 2023**, pp. 8730–8744, Toronto, Canada, July 2023. Association for Computational Linguistics.

[8] Huangjie Zheng, Xu Chen, Jiangchao Yao, Hongxia Yang, Chunyuan Li, Ya Zhang, Hao Zhang, Ivor Tsang, Jingren Zhou, and Mingyuan Zhou. Contrastive attraction and contrastive repulsion for representation learning. **Transactions on Machine Learning Research**, 2023.

[9] Gerlof Bouma. Normalized (pointwise) mutual information in collocation extraction. In **Proceedings of the Biennial GSCL Conference 2009**, Potsdam, Germany, 2009. GGSCL.

[10] Pengcheng Yang, Xu Sun, Wei Li, Shuming Ma, Wei Wu, and Houfeng Wang. SGM: Sequence generation model for multi-label classification. In Emily M. Bender, Leon Derczynski, and Pierre Isabelle, editors, **Proceedings of the 27th International Conference on Computational Linguistics**, pp. 3915–3926, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

[11] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[12] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In **International Conference on Learning Representations**, 2019.