

階層分類タスクにおけるスパースな回答に頑健な統合手法

榎本 昌文[†] 竹岡 邦紘^{††} 小山田昌史^{††} 董 于洋^{††} 岡留 剛^{††}

[†] 関西学院大学 大学院 理工学研究科 〒 669-1337 兵庫県三田市学園 2 丁目 1 番地

^{††} 日本電気株式会社 データサイエンス研究所 〒 211-8666 神奈川県川崎市中原区下沼部 1753

E-mail: [†]mossa@kwansei.ac.jp, ^{††}{k_takeoka,oyamada,dongyuyang}@nec.com, ^{†††}tokadome@acm.org

あらまし 低コストなラベリングを実現するクラウドソーシングでは、高品質なラベルを得るために、複数人のワーカーからの回答を統合することで品質管理を行う方法が広く採用されている。しかし、実際の応用では、コストの制約でワーカーに問い合わせをできないことや、意欲または自信の低いワーカーが回答を行わないことによって、各サンプルに対して十分な数の回答を収集できず、スパースな回答が得られる。特にマルチラベル分類タスクの場合、サンプルと候補ラベルの全ての組に対して、ワーカーに回答させることは難しいので、回答はよりスパースになる。階層分類タスクはマルチラベル分類タスクの一つだが、ラベルの階層情報を制約とすることで、回答のスパース性に対処した効果的な回答統合ができる。本研究では、階層分類タスクにおけるスパースな回答に頑健な統合手法を提案する。提案手法では、階層構造を潜在空間に埋め込み、サンプルの表現に適切な事前分布を置くことによって、回答のスパース性と階層構造の両方を考慮した回答統合を行う。本研究では、階層構造の情報を活用しつつ、ワーカーに対するサンプルとラベルの割り当てに依存しない回答の生成プロセスをモデル化する。人工データを用いた実験を行い、ラベルの推定において、提案手法が既存手法に比べて有効であることを示した。

キーワード クラウドソーシング, 品質管理, 回答統合, 階層分類

1 はじめに

階層分類は、図 1 に示すように、事前に定義されたラベルの階層構造に従って、1 つのサンプル x に 1 つの葉ラベルまでのパスにある全てのラベル z を割り当てる分類である。階層分類は、画像や文書、タンパク質の機能の分類のために広く利用されている [1]。階層分類を行うモデルの構築には、ラベル付きのデータセット $\{x, z\}$ が必要である。一般的に、正しいラベル z を付与するためには、細かい知識が必要であったり、ラベルの種類が多いため、高いコストがかかる。このようなラベル付きデータセット作成のコストを削減するために、クラウドソーシングを利用することが多い。クラウドソーシングを利用することによって、専門家によるアノテーションにかかる時間やお金を抑え、低コストでラベル付きデータセットを得ることができる [2]。Amazon Mechanical Turk¹等のクラウドソーシングサービスがラベル付きデータセットの収集をサポートしている。

クラウドソーシングによって収集した回答 y は、低コストで得られる反面、間違いが含まれることがあるなど品質が低く、それを補う仕組みが必要となる。そこで、複数のワーカーに同じタスクを依頼することによって品質を高めることが研究されている。具体的には、複数のワーカーから得られた回答 y を統合し、各サンプルに対応するラベル z を推定する。これは回答統合と呼ばれる手法であり、クラウドソーシングにおける品質管理の一例である。例えば、回答の多数決がもっとも単純な手法である。ワーカーの熟練度を考慮するなど多数決よりも高精度に統

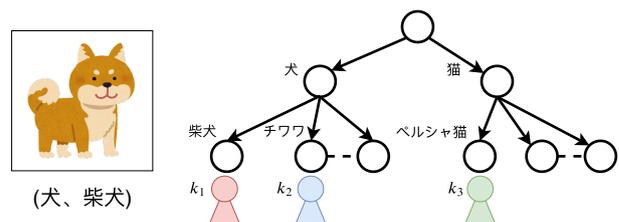


図 1 階層分類タスクの例：三人のワーカーが一つの画像に映っている動物を分類する。画像に映っているのは柴犬であるため、真の階層ラベルは(犬, 柴犬)である。ワーカー k_1, k_2, k_3 はそれぞれ階層ラベル(犬, 柴犬),(犬, チワワ),(猫, ベルシャ猫)を付与した。

合する手法がいくつも提案されている [3-5]。ところが、階層分類のように複数のラベルを付与する場合、各ワーカーにサンプルを提示した上で各ラベルが付与されるかどうかを問い合わせる必要があるため、回答数が相対的に少なくなる。特に、意欲や自信の低いワーカーが回答を行わない場合 [6, 7] や、金銭や時間的な制約で問い合わせができない場合 [8]、十分な回答数を得ることが難しい。その結果として、回答を統合しても真のラベル z の推定が困難になる。このように、得られる回答数が相対的に少ない状態を回答がスパースであるという。実際に回答が全体の 0.6%しか得られない例があり [9]、スパースな回答に対処する手法が必要とされている。

マルチラベル分類タスクにおいて回答のスパース性に対処した Li ら [7] の回答統合手法は、階層分類タスクではうまく動かない。Li らの手法では、回答におけるラベルの共起性を利用して、欠損した回答を補完することで、回答のスパース性に対処する。

1 : <https://www.mturk.com/>

しかし、階層ラベルは単純なマルチラベルとは異なる共起性を持つ。階層ラベルにおいて、あるラベルが付与されると、その上位概念のラベルは付与されるが、逆は成り立たない。そのため、階層ラベル特有の共起性をモデル化できていないという課題がLiらの手法には残っている。

階層分類タスクでは、ラベルの階層構造を活用した回答統合を行うことで、ラベルの推定精度を高めることができる。例えば図1において多数決で回答統合を行うことを考える。葉ラベルのみに着目すると、各ワーカーの回答は異なるため、回答を統合することはできない。ところが、一つ上の階層に着目して、葉ラベル間の相違度を考慮すると、二人のワーカーは「犬」と答えたため、「ペルシャ猫」より「柴犬」か「チワワ」が確からしいことが分かる。大谷ら [10] は、階層構造の上位にあるラベルから下位にあるラベルの順にワーカーに質問した結果、1つのサンプルに対して全ての候補ラベルについての回答が行われるという仮定の下、葉ラベル間の相違度を考慮した回答統合手法を提案した。しかし、回答がスパースなときは、葉ラベルについての回答が行われるとは限らないため、そのようなワーカーの能力のモデル化による回答統合は難しい。

本論文では、階層分類タスクにおいて、回答のスパース性に頑健な回答統合手法を提案する。提案モデルでは、真のラベル \mathbf{z} の事前情報であるラベルの階層構造を潜在空間へ埋め込むことで、サンプルの表現ベクトルの事前分布を得る。ノイズが乗ったサンプルの表現ベクトルから、ワーカーの回答が生成されると仮定することで、事前分布の制約によって回答の補完ができる。その結果、葉ラベルに対する回答が無いスパースな状況でも、ラベル間の相違度を考慮したワーカーの能力のモデル化と回答の統合が可能になる。

本研究の貢献は以下の点にまとめられる。

- (1) 階層分類タスクに対する回答統合において、欠損が存在する場合、Liらの手法による補完アプローチがうまく動かないことを見出した。
- (2) ラベルの階層構造を利用した、ワーカーへのサンプルとラベルの割り当てに依存しない回答の生成プロセスをモデル化し、それに基づく回答統合手法を提案した。
- (3) 人工データを用いた実験を実施し、ラベルの階層構造を考慮しない手法や多数決に比べて性能向上することを示した。

2 関連研究

2.1 階層分類

階層分類は、ラベルについて事前に定義された階層構造に沿って各サンプルを分類する問題であり、様々な実用上の問題に利用されている [1]。例えば、IPC と呼ばれる特許に関する階層的なカテゴリ分けを用いて、特許文書を分類する問題がある。階層分類が用いられる場面は、文書分類に限らず、医療画像や遺伝子情報、音楽など多岐にわたる。多くの手法が提案されており、代表的な手法である [11] は、親ノードごとに多クラス分類を学習しそれを繰り返すことによって階層分類を実現する。

2.2 階層分類タスクの回答統合

階層分類に用いる教師データを収集するためにクラウドソーシングを利用する場合、専門家と同等の品質を保つために回答統合が必要となる。大谷ら [10] は、クラウドソーシングにおける階層分類タスクに注目した回答統合手法を提案している。階層分類タスクをマルチクラス分類が連なったタスクと捉え、GLAD モデル [4] を拡張した手法を提案した。彼らの手法は、ワーカーは全てのラベルに対する回答を行なっていることを前提としている。そのため、先に述べたマルチラベル分類タスクで一般的に用いられる、ワーカーが全てのラベルに対して回答を行わない場合を取り扱うことができない。

2.3 マルチラベル分類タスクにおける回答統合

マルチラベル分類は1つのサンプルに複数のラベルがつけられることを許す分類であり、マルチラベル分類タスクにおける回答統合ではラベルの共起関係の利用が重要になる。ラベルの共起関係を利用することで、付与されていないラベルであってもワーカーが見落とした可能性を考慮でき回答統合性能を向上させることができる。

マルチラベル分類タスクにおける回答統合では、通常のカテゴリ分類タスクのために提案された回答統合手法の DS [3] や GLAD [4] をラベルの組合せに対して適用する手法が提案されている [12,13]。また、ワーカーごとの回答行列に低ランク性を仮定し、行列分解をベースとしたアプローチでラベルの共起関係を捉えた回答統合を行う手法も提案されている [14]。

マルチラベル分類タスクにおいては、回答のスパース性が重要な課題となる。Li と Jiang [7] は、回答のスパース性に対処するために、ワーカーの回答を欠損値入りのテンソルとみなし、テンソル分解による補完を行なったのちに、多数決を行う手法を提案した。一方で、本論文が取り扱う課題では、回答のテンソルが低ランク性を持っておらず、彼らの手法がラベルの階層構造が利用できないことから、よい性能を期待できない。

3 問題設定

N 個のサンプルに対して、 M 人のワーカーが L 個の候補ラベルの中から複数のラベルを付与する。サンプルを $i \in \{1, \dots, N\}$ 、ラベルを $j \in \{1, \dots, L\}$ 、ワーカーを $k \in \{1, \dots, M\}$ で表す。各ワーカー k の回答を行列 $\mathbf{Y}^k \in \{+1, -1, 0\}^{N \times L}$ で表現する。 $Y_{ij}^k = +1(-1)$ はワーカー k がサンプル i にラベル j を付けた（付けなかった）ことを意味する。 $Y_{ij}^k = 0$ は回答無し、つまり欠損を意味する。同様に真のラベルを行列 $\mathbf{Z} \in \{+1, -1\}^{N \times L}$ で表現する。 $Z_{ij} = +1(-1)$ はサンプル i にはラベル j が付く（付かない）ことを意味する。本論文の階層分類タスクでは、各サンプル i の真のラベル \mathbf{z}_i は二つの条件を満たす。(1) 各階層のラベルのいずれか一つのみが付く。(2) もしラベル j が付くなら、そのサンプルにはラベル j の階層構造上での先祖ラベル全てが付く。本論文では二つの条件を合わせて単一パス条件と呼ぶ。

本論文では、欠損を含むワーカーの回答 $\mathbf{Y} = \{\mathbf{Y}^k\}_{k=1}^M$ と、ラベルの階層構造を表す根付きの有向木 \mathcal{T} が与えられた下で、真の

ラベル \mathbf{Z} を推定する問題に取り組む。

4 提案手法

階層分類タスクにおける品質管理において、葉ラベル間の相違度を考慮した回答の統合とワーカーの能力のモデル化は有用である。しかし、ワーカーへのサンプルとラベルの割り当てを限定しない設定では、必ずしも葉ラベルに対する回答が与えられるわけではない。そこで提案手法では、木構造上での葉ラベル間の距離を保存するようなサンプル表現の事前分布を導入することで、この問題に対処する。また提案手法では、ワーカーの回答をサンプル表現と線形分類器の重みに分解することで、推論時に回答の補完を行い、回答のスパース性にも対処する。本章では、先に回答の生成プロセスのモデルについて概要を語る。その後、具体的な確率モデルと推論方法を説明する。

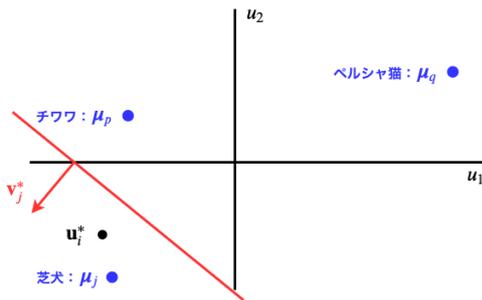


図2 ユークリッド空間を潜在空間としたときのサンプル表現。青い点はプロトタイプ、赤い線は決定境界を示す。

4.1 回答の生成プロセスのモデル

提案モデルにおいて、サンプル i の候補ラベル j に対する真のラベル Z_{ij} は、潜在空間における真のサンプル表現 $\mathbf{u}_i^* \in \mathbb{R}^D$ を、重み $\mathbf{v}_j^* \in \mathbb{R}^{D+1}$ を持つ候補ラベル j の線形分類器で分類することで生成される (図2)。真のサンプル表現 \mathbf{u}_i^* は、常に正しい回答を行う理想のワーカーが知覚しているサンプル i の表現だと解釈できる。また、葉ラベル j が付与されるサンプル \mathbf{u}_i^* は潜在空間で葉ラベル j を表現するベクトル $\boldsymbol{\mu}_j \in \mathbb{R}^D$ (= プロトタイプ) を中心として分布する。そして、プロトタイプ $\boldsymbol{\mu}$ は階層構造上での葉ラベル間の相違度をユークリッド距離として保存するように分布する。例えば、サンプル i は柴犬が写っている画像だとする (図1)。理想のワーカーは画像を見て、柴犬だと分類できるような視覚特徴を得る。この特徴が柴犬のプロトタイプ $\boldsymbol{\mu}$ である。そして、同じ犬というカテゴリに属するため、理想のワーカーが発見する柴犬とチワワの特徴は類似している。

提案モデルにおいて、サンプル i の候補ラベル j に対するワーカー k の回答 Y_{ij}^k は、潜在空間におけるワーカー固有のサンプル表現 $\mathbf{u}_i^k \in \mathbb{R}^D$ を、重み \mathbf{v}_j^* を持つ候補ラベル j の線形分類器で分類することで生成される。ただし、ワーカー固有のサンプル表現 \mathbf{u}_i^k は真のサンプル表現 \mathbf{u}_i^* にワーカー固有のノイズが乗ることで生じる。このノイズは葉ラベル間の相違度を考慮したワーカーの能力を表す。小さなノイズであれば、ワーカーのサンプル表現 \mathbf{u}_i^k は真の葉ラベル j のプロトタイプ $\boldsymbol{\mu}_j$ から大きく

は離れない。そして、プロトタイプは葉ラベル間の相違度を保存するように分布している。従って、真の葉ラベル j から遠く離れた別の葉ラベルがサンプル i に付与される可能性は低くなる。逆にノイズが大きければ、そのような可能性は高くなる。

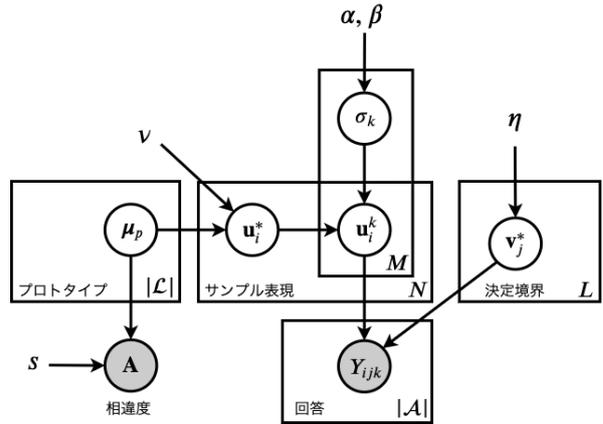


図3 回答プロセスのグラフィカルモデル

4.2 提案モデルと推論

提案モデルのグラフィカルモデルを図3に示す。提案モデルでは全てのパラメータに事前分布を置いて、パラメータのMAP解を求めることで真のラベルを推定する。全ての観測変数 $\{\mathbf{Y}, \mathbf{A}\}$ と潜在変数 $\boldsymbol{\theta} = \{\mathbf{U}, \mathbf{U}^*, \mathbf{V}^*, \boldsymbol{\sigma}, \boldsymbol{\mu}\}$ の同時分布は式(1)で定義される。

$$p(\mathbf{A}, \mathbf{Y}, \boldsymbol{\theta}) = p(\mathbf{Y}|\mathbf{U}, \mathbf{V}^*)p(\mathbf{U}|\mathbf{U}^*, \boldsymbol{\sigma})p(\mathbf{U}^*|\boldsymbol{\mu})p(\mathbf{A}|\boldsymbol{\mu})p(\boldsymbol{\mu})p(\boldsymbol{\sigma})p(\mathbf{V}^*) \quad (1)$$

回答 Y_{ij}^k の分布は入力 \mathbf{u}_i^k と重み \mathbf{v}_j^* を持つロジスティック回帰モデルで定義される:

$$p(\mathbf{Y}|\mathbf{U}, \mathbf{V}^*) = \prod_{k=1}^M \prod_{i=1}^N \prod_{j=1}^L p(Y_{ij}^k | \mathbf{u}_i^k, \mathbf{v}_j^*)^{I_{ijk}},$$

$$p(Y_{ij}^k | \mathbf{u}_i^k, \mathbf{v}_j^*) = a_{ijk}^{I(Y_{ij}^k=1)} (1 - a_{ijk})^{I(Y_{ij}^k=-1)},$$

$$a_{ijk} = \frac{1}{1 + \exp(-\langle \mathbf{u}_i^k, \mathbf{v}_j^* \rangle)}.$$

回答 \mathbf{Y}^k はサンプル表現 $\mathbf{U}^k = \{\mathbf{u}_i^k\}_{i=1}^N$ とワーカー間で共通の重み $\mathbf{V}^* = \{\mathbf{v}_j^*\}_{j=1}^L$ に分解されるので、推論時に回答 \mathbf{Y}^k の欠損を補完することができる。スパースな回答に対して過学習を起こさないように、重み \mathbf{V}^* の事前分布として平均が $\mathbf{0}$ のガウス分布 $p(\mathbf{V}^*) = \prod_{j=1}^L \mathcal{N}(\mathbf{v}_j^* | \mathbf{0}, \eta \mathbf{I})$ を置く。

ワーカーのサンプル表現 \mathbf{u}_i^k の分布は、真のサンプル表現 \mathbf{u}_i^* を平均、 σ_k^2 を分散とするガウス分布で定義される:

$$p(\mathbf{U}|\mathbf{U}^*, \boldsymbol{\sigma}) = \prod_{k=1}^M \prod_{i=1}^N \mathcal{N}(\mathbf{u}_i^k | \mathbf{u}_i^*, \sigma_k^2 \mathbf{I}) \quad (2)$$

分散パラメータ σ_k は葉ラベル間の相違度を考慮したワーカー k の能力を表現する。分散パラメータ σ_k が小さいと、ワーカー

k は正しく回答しやすい. 分散パラメータの事前分布としてガンマ分布 $p(\boldsymbol{\sigma}) = \prod_{k=1}^M \text{Gam}(\sigma_k^2 | \alpha, \beta)$ を置く.

真のサンプル表現 \mathbf{u}_i^* の事前分布は葉ラベル $p \in \mathcal{L}$ のプロトタイプ $\boldsymbol{\mu}_p$ を平均とした混合ガウス分布で定義される:

$$p(\mathbf{U}^* | \boldsymbol{\mu}) = \prod_{i=1}^N p(\mathbf{u}_i^* | \boldsymbol{\mu}), \quad (3)$$

$$p(\mathbf{u}_i^* | \boldsymbol{\mu}) = \frac{1}{|\mathcal{L}|} \sum_{p \in \mathcal{L}} \mathcal{N}(\mathbf{u}_i^* | \boldsymbol{\mu}_p, \nu \mathbf{I}).$$

この事前分布によって, 真のサンプル表現 \mathbf{u}_i^* が真の葉ラベル p のプロトタイプ $\boldsymbol{\mu}_p$ の近くに分布するように正規化される. プロトタイプの事前分布 $p(\boldsymbol{\mu})$ として一様分布を置く.

葉ラベル p と q の相違度 A_{pq} の分布は, プロトタイプ $\boldsymbol{\mu}_p$ と $\boldsymbol{\mu}_q$ のユークリッド距離 \tilde{A}_{pq} を平均とするガウス分布で定義される:

$$p(\mathbf{A} | \boldsymbol{\mu}) = \prod_{p \in \mathcal{L}} \prod_{q \in \mathcal{L}} \mathcal{N}(A_{pq} | \tilde{A}_{pq}, s) \quad (4)$$

$$\tilde{A}_{pq} = \|\boldsymbol{\mu}_p - \boldsymbol{\mu}_q\|_2$$

この分布により, 葉ラベル間の相違度をユークリッド距離として保存するように, プロトタイプ $\boldsymbol{\mu}$ が正規化される. 今回, 有向木 \mathcal{T} 上での葉ノード間のパスの長さを, 相違度 A_{pq} として用いた.

最急降下法を用いてパラメータ $\boldsymbol{\theta}$ の MAP 解を求め, 推定された分類器の重み $\tilde{\mathbf{v}}_j^*$ で, 推定された真のサンプル表現 $\tilde{\mathbf{u}}_i^*$ を線形分類することで, 推定ラベル \tilde{Z}_{ij} を得る:

$$\tilde{Z}_{ij} = \begin{cases} +1 & ((\tilde{\mathbf{u}}_i^*, \tilde{\mathbf{v}}_j^*) \geq 0) \\ -1 & ((\tilde{\mathbf{u}}_i^*, \tilde{\mathbf{v}}_j^*) < 0) \end{cases} \quad (5)$$

5 実験

本実験では, 人工データを用いてスパースな回答を生成し, 既存手法と提案手法を比べることで, 各手法の頑健性を評価する. 回答率と真のラベルの推定精度の関係を調べるために, 真のラベル \mathbf{Z} から人工的に回答 \mathbf{Y} を生成して, 回答統合後の推定ラベル $\tilde{\mathbf{Z}}$ を真のラベル \mathbf{Z} と比較し, 回答統合の性能を評価した.

5.1 実験設定

5.1.1 データセット

今回の実験では, 提案モデルを使って生成した 10 人分の人工的な回答 \mathbf{Y} を利用する. まず, 階層分類用のラベル付きデータセット ImageCLEF07A [15] の真のラベル \mathbf{Z} と葉ラベル間の相違度 \mathbf{A} を用いて, 同時分布 $p(\mathbf{A}, \mathbf{Z}, \mathbf{U}^*, \mathbf{V}^*, \boldsymbol{\mu}) = p(\mathbf{Z} | \mathbf{U}^*, \mathbf{V}^*) p(\mathbf{U}^* | \boldsymbol{\mu}) p(\mathbf{V}^*) p(\mathbf{A} | \boldsymbol{\mu}) p(\boldsymbol{\mu})$ の MAP 解を求めることで, 真のサンプル表現 \mathbf{U}^* と線形分類器の重み \mathbf{V}^* を学習する. 提案モデルの回答 Y_{ij}^k の分布と同様に, 真のラベル Z_{ij} の分布 $p(\mathbf{Z} | \mathbf{U}^*, \mathbf{V}^*)$ は入力 \mathbf{u}_i^* と重み \mathbf{v}_j^* を持つロジスティック回帰モデルである. それ以外の分布は提案モデルと全く同じである. その後, ワーカー k ごとに分散パラメータ σ^k を設定して, 式 (2) を使ってサンプル表現 \mathbf{U}^k を生成して, 重み \mathbf{V}^* を用いて

線形分類を行うことで, 全てが埋まった回答 $\tilde{\mathbf{Y}}^k$ を生成する. その後, 回答率に合わせて回答テンソル $\tilde{\mathbf{Y}}$ の要素をランダムに選択して, それ以外の要素を欠損状態 ($Y_{ij}^k = 0$) にして, スパースな回答 \mathbf{Y} を生成する. 実験では, 回答率を 10% から 100% まで変化させて, 回答率と推定精度の関係を調べた.

5.1.2 比較手法

比較手法として以下の 3 つの手法を用意した. (1)MV:各ラベルごとに回答の多数決で統合する.(2)CRISA_S:回答におけるラベルの共起性を利用して, 欠損した回答の推定を行い, 二値化してから, 多数決で回答統合する [7].(3)CRISA_V:CRISA_S と同様に欠損した回答の推定を行い, その結果を二値化せずに重み付き多数決によって回答統合する.

5.1.3 評価指標

各手法で推定したラベルを以下の 3 つの指標で評価した. (1)leaf macro-F1:各ラベルごとに F1 スコアを計算し, 葉ラベルについてのみ平均をとる. 葉ラベルのみの推定の正しさが確認できる.(2)all macro-F1:各ラベルごとに F1 スコアを計算し, 全てのラベルについて平均をとる. 全てのラベルの正しさが確認できる.(3)hierarchical macro-F1:推定ラベルの先祖ラベルの和集合を, 新たな推定ラベルとして, ラベル全体に対する F1 スコアを評価する [16]. 推定ラベルと真のラベルの階層上の距離を考慮した, 推定ラベルの正しさが確認できる.

5.2 実験結果

実験結果を表 1, 表 2, 表 3 に示す. 提案手法は回答率が 20% 以上のとき, 既存手法より高い精度を達成した. 一方, 回答率が 10% のときは提案手法は多数決より低い性能となった. これは, ロジスティック回帰モデルがスパースな回答 \mathbf{Y} に過剰適合し, 真のサンプル表現 \mathbf{U}^* を正しく推定できないためだと考えられる.

一方, CRISA_S, CRISA_V は常に多数決よりも低い性能となった. これは CRISA によって補完した回答 \mathbf{Y}' が, 人工的に生成した真の回答 $\tilde{\mathbf{Y}}$ と異なり, 回答統合に悪影響を与えたためだと考えられる. CRISA はマルチラベル分類タスクの設定において, 回答におけるラベルの共起性を利用して, 欠損した回答を推定する手法である. しかし, 階層ラベルは単純なマルチラベルとは異なる共起性を持つ. 階層ラベルでは, ラベル p が付与されるなら, ラベル p の階層構造上での先祖ラベル q が付与されるが, その逆は成り立たない. CRISA はこのような方向を持つ共起性を考慮しないため, 欠損した回答の推定に失敗したと考えられる.

leaf macro-F1 のスコア (表 1) と all macro-F1 のスコア (表 2) を比べると, leaf macro-F1 は多数決から, より大きく精度が向上している. 今回使ったデータセットでは, 全体のラベル数に対して葉ラベルの割合は約 65% 程度であることから, ランダムに回答を欠損させた場合, 葉ラベルがよりスパースになりやすい. 葉ラベルの補完が高い精度でできたため, 提案手法の性能は大きく向上したと考えられる.

表 1 10 人のワーカーの回答が回答率に従ってランダムに観測されたとして、回答統合して精度を比較した。回答を 5 回サンプリングしたときの leaf macro-F1 の平均と標準偏差を示す。

回答率	10%	20%	40%	60%	80%	100%
MV	0.527±0.009	0.698±0.010	0.863±0.019	0.938±0.012	0.968±0.003	0.971±0.004
CRIA _S	0.000±0.000	0.016±0.003	0.287±0.009	0.627±0.048	0.871±0.011	0.971±0.004
CRIA _V	0.020±0.005	0.171±0.025	0.678±0.024	0.890±0.024	0.961±0.003	0.971±0.004
提案手法	0.523±0.023	0.771±0.026	0.926±0.015	0.977±0.008	0.992±0.003	0.996±0.003

表 2 all macro-F1 の平均と標準偏差を示す。

回答率	10%	20%	40%	60%	80%	100%
MV	0.605±0.005	0.776±0.006	0.908±0.012	0.959±0.008	0.979±0.002	0.981±0.003
CRIA _S	0.014±0.001	0.111±0.007	0.519±0.008	0.755±0.032	0.915±0.007	0.981±0.003
CRIA _V	0.071±0.004	0.335±0.015	0.785±0.017	0.928±0.016	0.974±0.002	0.981±0.003
提案手法	0.630±0.017	0.843±0.017	0.952±0.010	0.985±0.005	0.995±0.002	0.997±0.002

表 3 hierarchical macro-F1 の平均と標準偏差を示す。

回答率	10%	20%	40%	60%	80%	100%
MV	0.788±0.005	0.882±0.005	0.954±0.006	0.981±0.004	0.990±0.001	0.991±0.001
CRIA _S	0.179±0.011	0.487±0.009	0.835±0.006	0.913±0.009	0.965±0.003	0.991±0.001
CRIA _V	0.336±0.008	0.685±0.008	0.918±0.006	0.970±0.006	0.989±0.001	0.991±0.001
提案手法	0.764±0.009	0.923±0.004	0.977±0.005	0.993±0.002	0.997±0.001	0.999±0.001

6 結 論

本論文では、階層分類のためのラベル付きデータセットをクラウドソーシングで収集する際に必要となる回答統合について、欠損が多く含まれる回答に対しても頑健で高精度な回答統合を行う手法を提案した。提案手法では、真のラベルの事前情報であるラベルの階層構造を、潜在空間へ埋め込むことで、サンプルの表現ベクトルの事前分布を得る。ノイズが乗ったサンプルの表現ベクトルから、ワーカーの回答が生成されると仮定することで、事前分布の制約によって回答の補完ができるため、葉ラベルに対する回答が無いスパースな状況でも、ラベル間の相違度を考慮したワーカーの能力のモデル化と回答の統合が可能になる。人工データを用いた実験において、提案手法は既存手法および多数決に比べて有効であることを確認した。

文 献

- [1] C. N. Silla, A. A. Freitas, A survey of hierarchical classification across different application domains, *Data Mining and Knowledge Discovery* 22 (1-2) (2011) 31–72.
- [2] F. Daniel *et al.*, Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions, *ACM Comput. Surv.* 51 (1) (2018) 7:1–7:40.
- [3] A. P. Dawid, A. M. Skene, Maximum likelihood estimation of observer error-rates using the em algorithm, *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 28 (1) (1979) 20–28.
- [4] J. Whitehill *et al.*, Whose vote should count more: Optimal integration of labels from labelers of unknown expertise, in: *Advances in Neural Information Processing Systems*, 2009, pp. 2035–2043.
- [5] P. Welinder *et al.*, The multidimensional wisdom of crowds, in: *Advances in Neural Information Processing Systems*, 2010, pp. 2424–2432.
- [6] S.-Y. Li *et al.*, Multi-label learning from crowds, *IEEE Transactions on Knowledge and Data Engineering* 31 (2019) 1369–1382.
- [7] S.-Y. Li, Y. Jiang, Multi-label crowdsourcing learning with incomplete annotations, in: *Pacific Rim International Conference on Artificial Intelligence*, Springer, 2018, pp. 232–245.
- [8] Y. Yan, S.-J. Huang, Cost-effective active learning for hierarchical multi-label classification., in: *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, 2018, pp. 2962–2968.
- [9] H. J. Jung, M. Lease, Improving quality of crowdsourced labels via probabilistic matrix factorization, in: *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.
- [10] N. Otani, Y. Baba, H. Kashima, Quality control for crowdsourced hierarchical classification, in: *2015 IEEE International Conference on Data Mining*, IEEE, 2015, pp. 937–942.
- [11] S. Dumais, H. Chen, Hierarchical classification of web content, in: *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2000, p. 256–263.
- [12] L. Duan *et al.*, Separate or joint? estimation of multiple labels from crowdsourced annotations, *Expert Systems with Applications* 41 (13) (2014) 5723–5732.
- [13] K. Yoshimura, Y. Baba, H. Kashima, Quality control for crowdsourced multi-label classification using raket, in: *International Conference on Neural Information Processing*, Springer, 2017, pp. 64–73.
- [14] J. Tu *et al.*, Multi-label answer aggregation based on joint matrix factorization, in: *2018 IEEE International Conference on Data Mining*, IEEE, 2018, pp. 517–526.
- [15] I. Dimitrovski *et al.*, Hierarchical annotation of medical images, *Pattern Recognition* 44 (2011) 2436–2449.
- [16] S. Kiritchenko, S. Matwin, A. F. Famili, Functional annotation of genes using hierarchical text categorization, in: *Proceedings of the ACL workshop on linking biological literature, ontologies and databases: mining biological semantics*, 2005.