修 士 論 文 概 要 書

Master's Thesis Summary

Date of submission: 01/10/2024 (MM/DD/YYYY)

専攻名（専門分野） Department	経営デザイン専攻	氏 名 Name	山田　晃輝 Koki Yamada	指 導 教 員 Advisor	後藤　正幸　　印 Seal
研究指導名 Research guidance	経営情報学研究	学籍番号 Student ID number	CD 5222F018-2		
研究題目 Title	求人データに対する適切な職種情報付与のためのラベル修正アルゴリズム Label Modification Algorithm for Assigning Appropriate Job Information to Job Description Data				

1. 研究背景と目的

近年人材の流動性が高まりつあり，国内の転職市場も活況を呈している．マイナビによって実施された調査 [1] によれば，2022 年には正社員のうち過去最高の 7.6%が転職を行った．これに伴い，転職活動においてオンライン上の転職サイトの利用が活発になっている．転職サイトでは自身の属性を登録することで，最適な求人の自動推薦や検索などを通じて自身に合致した適切な求人を容易に見つけ出すことができる．自動推薦や検索の際に参照される情報として最も重要なものの一つは，求人に付与される職種ラベルである．すなわち，求職者と求人の適切なマッチングを実現するために，職種ラベルの適切な付与は極めて重要なタスクである．しかし，現状の転職サイトでは，職種ラベルが募集会社や人材エージェントの担当者によって個別に付与されている．そのため，ラベル付与の品質が一定ではなく，一貫性のない職種ラベルが付与されている可能性がある．しかし，すべてのデータを同じ観点を持つ人間によりラベル付けを行うのはコストが高く現実的ではない．以上より，現在の求人データに付与されている職種ラベルのうち，適切ではない再度付与し直すべきものを修正することができれば有用である．

もともとラベルが付与されていないラベルなしデータに対し，学習に効果的なデータに優先的にラベル付けを行うことにより少ないアノテーションコストで精度の高いモデルを構築する手法として能動学習 [2] が知られている．ここで，能動学習におけるラベル付けを行うデータの選択を，ラベルありデータに対してラベルを再付与すべきデータ抽出に利用することが考えられる．ラベルを付与するデータを抽出する基準となる獲得関数として，予測分布のエントロピーが高いデータを選択する方法 (entropy sampling) や，予測確率の最大値が最小のサンプルを選ぶ方法 (least confident) などがある．しかし，能動学習はもともと対象データ (未学習データ) にラベルが付与されていない状況を想定しており，今回のようにすでにラベルが付与されている場合に適用するのは適切ではない．

そこで本研究では，すでに付与されたラベルと真のラベルの同時確率分布を推定することにより不適切なラベル (Noisy Label) を除去できる手法である Confident Learning[3] を援用し，不適切な職種ラベルを抽出し修正する手法を提案する．ここで，研究対象データには大カテゴリと複数付与される小カテゴリの 2 つのラベルが含まれていることを踏まえ，これらの階層性を考慮することでより適切な Noisy Label の抽出を目指す．実験では，求人データの一部に対しラベルを改変したデータを用いて提案手法の有用性を示す．さらに，改変前の実データにも提案手法を適用し，実応用上提案手法が有用であることを示す．

2. 問題設定

2.1. 転職サイトとそのデータの問題点

転職サイトとは，転職を検討しているユーザが求人の閲覧，応募等ができるオンラインサービスのことを指す．転職サイトでは，募集企業よって様々な職種の求人が掲載されており，転職活動における効率的な情報収集に大いに役立つ．例えば，ユーザが現職の属性情報や希望条件等を入力することでユーザにマッチした求人が自動で推薦されるため，従来よりも効率的に自身に合った求人を探すことが可能である．自動推薦時に参照される重要な補助情報として職種ラベルが挙げられる．しかし，求人の転職サイトへの登録自体はそれぞれの企業が個別に行っているため，職種ラベルをはじめとした補助情報付与のクオリティや基準が一定ではなく，利用時にノイズとなる可能性がある．すなわち補助情報が整理されていないことが，適切なジョブマッチング実現の障害になっていると考えられる．

2.2. 分析対象データ

本研究では，大手求人サイト A 上に掲載されている求人データ 60,878 件を対象データとする．本データは 2022 年 11 月 7 日から 11 月 27 日の間にスクレイピングにより収集したものである．本データには，各求人に対しポジション名，職務内容，待遇，応募要件，勤務地などの基本情報と，職種ラベルである大カテゴリと小カテゴリが付与されている．主な内容の具体例を表 1 に示す．

表 1. 分析対象データに含まれる主な内容

項目名	具体例
job_title	コンサルタント (AI：ストラテジスト・プロデューサー・データアーキテクト)
job_description	目的に合致した最適手法での分析実施, 知見獲得，モデル構築支援…
major_category	IT コンサルティング
minor_category	システムコンサルタント, データサイエンティスト, データベースエンジニア

対象データの大きな特徴として，職種カテゴリに，大カテゴリと小カテゴリの 2 種類が存在し，階層関係を有することが挙げられる．具体的には，大カテゴリは “IT コンサルティング” や “営業職” といった大まかな分類が，小カテゴリは “ネットワークコンサルタント” や “法人営業” といった大カテゴリに紐づけられたより細かい分類がそれぞれ該当する． 各求人には，大カテゴリは 1 つが付与され，小カテゴリは，1 つ以上，最大 3 つまでが付与されている．なお，小カテゴリには少なくとも 1 つは大カテゴリに対応した小カテゴリが含まれる．

2.3. 関連研究

2.3.1. 能動学習

能動学習とは，教師なしデータの中からモデルの精度を高められるようなデータに優先的にラベル付けを行うことで，ラベル付けを行うデータ数を削減しつつ高い精度を示すモデルを構築する手法を指す．

また，能動学習においてラベル付けを行うデータを選択する獲得関数について，予測分布のエントロピーが高いデータを選択する方法 (entropy sampling) や，予測確率の最大値が最小のサンプルを選ぶ方法 (least confident) などが知られている．一方で，能動学習は教師なしデータに対しラベル付けを行うものを選択する手法であるため，本研究の対象データのようにすでに付与されたラベルがある場合のデータ選択とは問題設定が異なる．

2.3.2. Confident Learning

Confident Learning(CL) とは，画像や文書などの教師ありデータにおいて，人間の判断ミスに起因してノイズが含まれるラベル (Noisy Label) を，真のラベルと付与されたラベルの同時確率分布を推定することにより抽出する手法を指す．データの真のクラスを y^*，付与されたラベルを $\tilde{y}$ とすると，$y^* = j$，$\tilde{y} = i$ となる確率は式 (1) で表される．これを求めるため，CL では式 (2) で表される同時分布を求める．

$$Q_{\tilde{y}|y^*} = p(\tilde{y} = i|y^* = j) \tag{1}$$

$$Q_{\tilde{y},y^*} = p(\tilde{y} = i, y^* = j) \tag{2}$$

ここで，真のクラス y^* は未知であるため，モデルの出力確率 $p(y|x;\theta)$ が大きくなるクラス y を y^* とみなす．$p(y|x;\theta)$ についてしきい値を設けることで，y^* と $\tilde{y}$ が同時に表れやすい組み合わせを捉えることが可能になる．

CL のステップは以下の通りである．

Confident Learning

Step.1 学習済みモデルを用いて Noisy Label の予測分布を得る．

Step.2 y^* と $\tilde{y}$ の組み合わせについてカウントし，$\boldsymbol{C}_{\tilde{y},y^*}$ にまとめる．

Step.3 $\boldsymbol{C}_{\tilde{y},y^*}$ を正規化し，同時分布 $\hat{\boldsymbol{Q}}_{\tilde{y},y^*}$ を推定する．

Step.4 $\hat{\boldsymbol{Q}}_{\tilde{y},y^*}$ を基に，誤ったラベルを持つデータを抽出する．

3. 提案手法

3.1. 着想

蓮本ら [4] は，EC サイトの購買履歴を用いて，離反予測と customer lifetime value (CLV) の予測という 2 つのタスクを，ニューラルネットワークを用いたマルチタスク学習によって実現する手法を提案した．CLV の予測を行うためには，顧客がいつ離反するのかに加え，どの商品をどれくらい購入し，その取引からの利益はどれくらいなのかを知る必要がある．そのため，蓮本らの手法では，まず比較的推定の難易度が低い離反予測を行い，そこで得られた生存確率 p と潜在変数 z を用いて CLV を予測するという構造のモデルを導入している．

本研究で用いるデータの職種ラベルには階層性がある．ここでは，職種ラベルのうち数が多く誤ったラベルが多く含まれると考えられる小カテゴリについて修正の対象とする．この場合，学習時に小カテゴリの情報のみを予測するよりも，大カテゴリを予測した結果を用いて小カテゴリを予測したほうが精度が高くなると考えられる．このことを踏まえ，提案手法では蓮本らの手法に倣い職種ラベルの階層性を踏まえたニューラルネットワークを用いてモデルを構築する．

3.2. 提案手法の概要

提案手法のステップを以下に示す．

提案手法

Step.1 現在付与されているラベルを用いて，データから階層性を加味したモデルを学習する．

Step.2 Step1 で学習したモデルの出力結果よりラベルの信頼度を Confident Learning により捉え，信頼度が低いデータを抽出する．

Step.3 Step2 で抽出されたデータについて，人間の手で再度ラベル付与を行い，そのラベルに更新する．Step.1 に戻り，予め設定した回数 K に達するまで繰り返す．

3.2.1. Step.1：マルチタスク学習

Step.1 では，初めに対象とする求人データの業務内容 (job_description) について，東北大学による事前学習済み日本語 BERT[5] を用いてベクトルへ変換する．具体的には，入力トークンの系列 (文章) をベクトル化し，CLS トークンに対しての最終層の出力を文章ベクトルとする．これにより，各データに対し，BERT の出力次元数である 768 次元の分散表現を得る．

次に，得られた分散表現に対して階層性を利用したニューラルネットワークを用いてマルチタスク学習を行う．具体的には，小カテゴリが大カテゴリに依存するため，初めに大カテゴリと潜在変数を予測し，その結果を基に小カテゴリの予測を行う構造のモデルとする．提案手法で用いるニューラルネットワークの構造を図 1 に示す．このニューラルネットワークには，64 ノードおよび 32 ノードの共通の隠れ層と，32 ノードの各タスクに特化した隠れ層が含まれる．

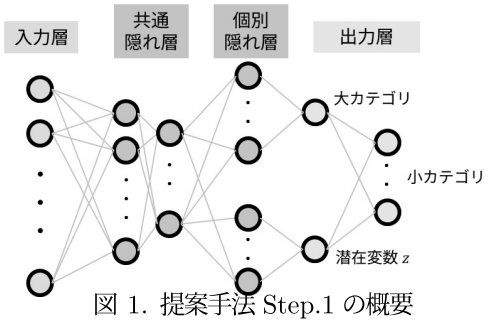

図 1. 提案手法 Step.1 の概要

3.2.2. Step.2：Confident Learning

Step.2 では，Step.1 で学習したモデルから得られる小カテゴリの予測分布と与えられたラベルに Confident Learning を適用し，Confident Learning によって Noisy Label が含まれるデータを抽出する．なお，Confident learning はマルチラベル問題には直接適用できないため，各データで最も確率の高い小カテゴリのラベルを選択して適用する．

3.2.3. Step.3：ラベルの更新

Step.3 では，Step.2 によって抽出されたデータに対し，目視により不適切なラベルを見直し，より適切と思われるラベルに修正する．最終的に，Step1〜3 を一定回数 K に達するまで繰り返すことで，データから Noisy Label を取り除く．

4. 実験

4.1. 実験の概要

本研究では，提案手法の有用性を示すため実験を 2 つ行う．初めに実験 1 では，定量的に提案手法の性能の評価するために，元データの一部のラベルを人工的に誤りラベルを生成して他手法と比較する．次に実験 2 では，実データに提案手法を適用し，定性的に Noisy Label を検出できるのかを確認する．実験 1, 2 に共通して，提案手法の Step.1 の学習における Epoch 数は 256 とし，最適化手法は学習率 0.01 の Adam を用いた．実験 1 では，ランダムにデータを抽出する手法と，能動学習で用いられる予測分布のエントロピーが大きいデータを抽出する手法を比較手法とする．以下に，それぞれの概要を示す．

実験 1：提案手法の性能評価

実験 1 では，人工的に生成した誤りラベルを含むデータを用いて提案手法の性能の評価を行う．実際のデータに含まれる誤ったラベルは未知であるため，すべてのラベルを正しいものと仮定し，全データの 3%について小カテゴリのラベルを置き換える形で誤りのラベルを生成する．ラベルを置き換えたデータをどれだけ抽出できるかで性能を評価する．実験 1 の概要を図 2 に示す．

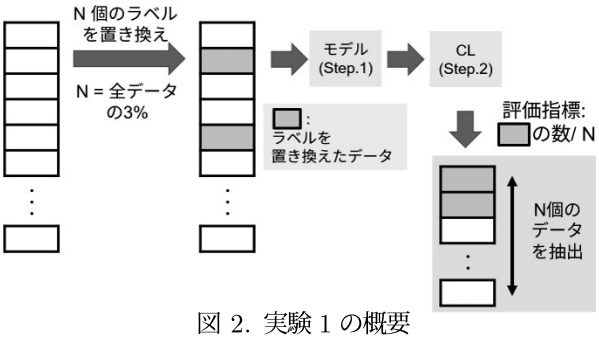

図 2. 実験 1 の概要

実験 1 では，職種ラベルの置き換えについて 2 パターンの実験を行う．実験 1-1 では，ラベルをランダムに置き換える．本研究で対象とするデータでは，例えば「データサイエンティスト」と「データベースエンジニア」など，ラベルの意味が近いものについて，その解釈の違いに起因した Noisy Label ラベルが多いと考えられる．そのため，実験 1-2 では，ラベル間の類似度を考慮してラベルの置き換えを行う．具体的には，同じラベルに所属するデータのベクトル平均をそのラベルの潜在表現とし，その cos 類似度で類似度を計算する．類似度の計算方法を式 (3)，(4) に示す．ここで，$\boldsymbol{V}_l$ はラベル l の潜在ベクトル，$\boldsymbol{v}_d$ は求人 d の潜在ベクトル，D_l はラベル l に所属する求人データ集合，$\mathrm{sim}(l_1, l_2)$ は，ラベル l_1 と l_2 の類似度をそれぞれ表す．

$$\boldsymbol{V}_l = \frac{1}{|D_l|} \sum_{d \in D_l} \boldsymbol{v}_d \tag{3}$$

$$\mathrm{sim}(l_1, l_2) = \frac{\boldsymbol{V}_{l_1} \cdot \boldsymbol{V}_{l_2}}{\|\boldsymbol{V}_{l_1}\| \|\boldsymbol{V}_{l_2}\|} \tag{4}$$

あるラベル l について，式 (4) で得られた類似度が高いもの上位 20 個の中から式 (5) に示す確率で置き換える．ここで，$P(l')$ はラベル l がラベル l' に置き換わる確率，L_{top20} は l' との類似度が高い上位 20 ラベルの集合をそれぞれ表す．

$$P(l') = \frac{\mathrm{sim}(l, l')}{\sum_{m \in L_{\mathrm{top20}}} \mathrm{sim}(l, m)} \tag{5}$$

実験 2：実データを用いた有効性評価

実験 2 では，ラベルの置き換えを行っていない実際の元データに提案手法を適用することで，実応用上 Noisy Label を検出できるのかを確認する．具体的には，抽出されたデータに対し，適切と思われるラベルをデータの中身から判断し，提案手法により Noisy Label を抽出できるのかを確認する．簡単のため，実験 2 では提案手法の Step.1〜3 は 1 回のみ適用する．

4.2. 実験結果

実験 1-1 の結果を表 2，実験 1-2 の結果を表 3 にそれぞれ示す．表 2 および表 3 より，実験 1-1，1-2 いずれにおいても比較手法と比べ提案手法が最も高い精度を示したことがわかる．

表 2. 実験 1-1 における置き換えられたデータの抽出精度

提案手法	エントロピー	ランダム
0.109 ± 0.013	0.028 ± 0.005	0.028 ± 0.005

表 3. 実験 1-2 における置き換えられたデータの抽出精度

提案手法	エントロピー	ランダム
0.101 ± 0.006	0.027 ± 0.005	0.030 ± 0.002

次に，実験 2 で提案手法により最も Noisy Label が含まれる可能性が高いと判断された上位 10 求人のうち，2 つについて表 4 に示す．

表 4. 実験 2 によって抽出された求人の例

求人タイトル	職務内容	元ラベル
SAP BASIS コンサルタント経験者採用／PJ 工程設計経験者	製造業を中心とした日系企業の経営・業務の変革を SAP ソリューションを活用してサポートするべく，SAP ソリューションの導入や展開に係る SAP BASIS コンサルタントとしての業務をお任せします．	システムコンサルタント，法人営業
相続財産評価フロント業務★金融機関または税理士法人出身者の募集です	◇親会社の信託銀行から紹介される個人富裕層を中心顧客として，以下の通り，相続人・信託銀行担当者・税理士と連携して，資料収集，契約締結など対外的な業務の全てを行います．	戦略コンサルタント，税理士，個人営業・FP

表 4 の求人について，職種ラベルへの修正を行う．初めに 1 つ目の求人について，職務内容を見ると，SAP のシステム設計及び構築を行うものであり，営業の要素は含まれていないことがわかる．このことから，元のラベルに含まれる「法人営業」は不適切なラベルであるといえる．2 つ目の求人についても職務内容を見ると，個人を対象顧客にした業務であることがわかる．しかし，元のラベルには「戦略コンサルタント」が含まれている．戦略コンサルタントは，企業の経営層に対し経営課題の解決のための戦略を策定する仕事を指すため，この求人に対して「戦略コンサルタント」というラベルを付与することは不適切である．このほかの求人についても中身を確認したところ，上位 10 求人のうち 4 求人には不適切と思われる職種ラベルが付与されていた．そのため，提案手法によって Noisy Label を効率的に修正できることが確認できた．

5. 考察

5.1. 提案手法の有効性

実験 1-1 よりも，1-2 の方が全体的にわずかに精度が低かった．このことから，真のラベルと類似した Noisy Label の抽出は難易度がやや高いことがわかる．しかし，実験 1-1・1-2 いずれにおいても提案手法の精度はランダムなどの比較手法に比べ約 4 倍であった．このことから，求人データから Noisy Label を抽出するタスクに対し，提案手法は十分な性能を発揮することが示唆される．

提案手法では，能動学習の手法とは異なり，すでに付与されたラベルを加味して真のラベルとの同時確率を推定し不適切なラベルが含まれたデータを抽出している．提案手法の有効性は，Noisy Label を抽出するタスクにおいては付与されたラベルを加味することが重要であることを裏付けている．

5.2. 実際のサービスへの活用

従来，求人サイト上の求人ラベルに含まれる Noisy Label を除去するためには，人の手による確認が必要であった．一方で，求人サイト上には膨大な数の求人が存在するため，すべての求人を確認しようとすると膨大な人的コストが発生する．ここで，提案手法では Noisy Label を含む可能性の高い求人を抽出できるため，より効率的にラベルの修正が可能である．そのため，提案手法を用いることでラベルの確認にかかる人的コストを大幅に削減することができる．

6. まとめと今後の課題

本研究では，求人データに対し，不適切な職種ラベルを効率的に修正可能な手法を提案した．提案手法は人工データを用いた実験において比較手法よりも高い性能を示し，実応用上においても有用であることが確認された．

今後の課題としては，求人以外のデータへ適用した際の有効性の検討が挙げられる．

参考文献

[1] 株式会社マイナビ 社長室 HR リサーチ統括部. 転職動向調査 2023 年版（2022 年実績）, 2023. https://career-research.mynavi.jp/wp-content/uploads/2023/03/555cd736526f1ee6b8d2\\1539df2362f1.pdf.

[2] Burr Settles. Active learning literature survey. 2009.

[3] Curtis Northcutt, Lu Jiang, and Isaac Chuang. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, Vol. 70, pp. 1373–1411, 2021.

[4] 蓮本恭輔, 後藤正幸. 多層ニューラルネットワークを用いたマルチタスク学習による顧客購買行動予測. 情報処理学会論文誌, Vol. 63, No. 6, pp. 1276–1286, 2022.

[5] 東北大学乾・鈴木研究室. Pretrained Japanese BERT models, 2023. https://github.com/cl-tohoku/bert-japanese.