模倣学習を用いた階層的商品分類

三田　雅人
奈良先端科学技術大学院大学
mita.masato.mz2@is.naist.jp

村上　浩司
楽天技術研究所ニューヨーク
koji.murakami@rakuten.com

1　はじめに

文書分類は自然言語処理において最も重要なタスクの一つである．既存の文書分類手法は，非階層型（Flat Model）と階層型（Tree Model）の大きく二つに分けることができる．Flat Model は，カテゴリツリーが与えられたとしてもその階層構造を用いず，予め与えられた末端カテゴリ候補群の中から最適なカテゴリに分類する多クラス分類問題として定式化されることが多い [3]．Tree Model は，典型的には与えられた木構造のカテゴリを最上層から最下層へと逐次的に局所的な分類を繰り返すことで最適なカテゴリに分類する手法である．分類対象のカテゴリ数が膨大なときには大きな 1 つのモデルで分類するよりも，階層構造を利用してツリー上の各階層で小さなモデルを作成して分類するほうが全体として計算・容量効率を向上させることができる [6]．しかしながら、各カテゴリにおいての分類は局所的であり、繰り返すことで誤り伝搬が起きるために精度が低下しがちという欠点が知られている [1]．誤り伝搬は分類対象の文書の種類に関係なく，Tree Model で分類する場合に起こる問題である．この結果，文書が正しいカテゴリに分類されずカテゴリツリー上でも関連の薄いカテゴリに分類されることがしばしばある．

一般的な文書分類を考えた場合，正しく分類された割合が精度の評価には重要であるが，誤分類の質を考慮する必要が生じる場合がある．特に Amazon[1] や楽天 [2] など，オンラインで数多くの種類の商品を取り扱う場合である．階層構造を使った商品分類においては関連の薄いカテゴリに誤分類した場合と，例えば兄弟カテゴリに誤分類された場合では商品を取り扱う店舗にとっては大きく意味合いが異なる．店舗の目的は商品販売による利益の最大化および損失の最小化であることから，ある商品が関連の薄いカテゴリに分類されると購入されにくいが，兄弟カテゴリのような関連性があるカテゴリに分類されると，正しいカテゴリに分類された場合には及ばないまでも，ユーザによって購入される確率が高くなる．そのため階層的商品分類では，より高精度で商品分類ができ，誤分類の場合にもできるだけ正解カテゴリに近いカテゴリへの分類が可能なモデルが望ましい．

本稿の貢献は次の通りである．

- 模倣学習手法の一つである Dataset Aggregation（DAGGER）を既存の Tree Model に用いた階層的商品分類モデルを提案した．
- 商品データを用いた評価実験により，提案手法は誤り伝搬を低減し，分類精度の向上および Average Revenue Loss （ARL）の低下を実現した．
- ARL という観点から，既存の Flat Model と Tree Model に対する分析を行った．

本稿の構成は以下の通りである．２節で模倣学習，特に DAGGER の説明をし，３節で DAGGER の階層的商品分類への適用法について述べる．さらに，４節で実際の商品データを用いた階層的商品分類の評価実験を示す．

2　模倣学習

模倣学習はエキスパートの行動を観察し，その行動を真似ることで，エキスパートと同様の行動ができるように方策を学習する手法である．ここで，エキスパートとは理想的な行動をとる主体を指す．模倣学習を NLP 分野に適用した例は多くはないが，最も関連の深い研究として，坪井ら [5] がある．坪井らは，模倣学習手法の一つである Dataset Aggregation（DAGGER）アルゴリズム [4] を系列ラベリングと Shift-reduce 依存構造解析の 2 つの構造予測問題における決定的解析に適用し，誤り伝搬を回避できることを示した．本稿においても，模倣学習手法として DAGGER を用いる．

方策を現在の状態 $s \in S$ から次の行動 $a \in A$ への写像 $\pi : S \rightarrow A$ と定義し，π が訪れた状態を s_π と書く．一般的には，方策は分類器によって近似される．

[1] http://www.amazon.co.jp
[2] http://www.rakuten.co.jp

Algorithm 1 DAGGER

初期化: $D \leftarrow \emptyset, \pi_1$ は任意の方策
for $k = 1, 2, \cdots, K$ **do**
  $\pi_k \leftarrow \beta_k \pi^* + (1 - \beta_k)\hat{\pi_k}$
  π_k を実行し $D_k = \phi(s_{\pi_k}), \pi^*(s_{\pi_k}))$ を収集
  データを集約: $D \leftarrow D \cup D_k$
  D を用いて π_{k+1} を学習
end for
検証用データで性能の良い π_k を選択

表 1: DAGGER アルゴリズム

正しい行動を返すオラクルを π^* と書く．DAGGER は反復方策と呼ばれるアルゴリズムで，各反復においてオラクルとそれまでに学習済みのすべての方策が訪れた状態の下で新たに方策を学習する [4]．DAGGER の最初の反復では，既存の教師あり学習と同じようにオラクルが訪れた状態を訓練データとして方策を学習する．次の反復以降では，オラクルが訪れた状態に加えて学習した方策が訪れた状態も訓練データに加えて学習する．DAGGER をアルゴリズム 1 に示す．ただし，β_i は状態遷移を行う際に使用するオラクルと訓練している方策との混合方策の混合率で $N \rightarrow \infty$ のとき $\frac{1}{N}\sum_k \beta_k \rightarrow 0$ を満たす数列である．混合率は $\beta_1 = 1, \beta_k = 0(k > 1)$, つまり初回はオラクル π^* を用い，2 回目以降は現在の方策を用いる方法が経験的に良いことが報告されている [4]．方策が実際に訪れるであろう状態をサンプリングしその下で経験誤差を最小化するため，方策が選ぶ行動がオラクルと異なる，つまり過去に誤りがあった場合にもそれ以降の行動は誤らないことが期待できる．

3 提案手法

階層的商品分類において Tree Model に DAGGER を適用する方法について述べる．階層的商品分類は，与えられた商品 x に対して，最適なラベル $y \in Y$ を予測する問題である．ここで，y はあらかじめ定義されたカテゴリラベルの集合であり，階層構造で組織化されているとする．Tree Model の場合には，階層を最上層から最下層まで分類器を逐次適用することで最終的に末端カテゴリ y_T を予測する．各カテゴリの親カテゴリを辿っていけば，最上層カテゴリから現在着目するカテゴリまでのパスを一意に得ることができる．Tree Model では，階層を逐次的に分類していく際，閾値より高い枝は保持し，それ以外は削除していく枝切り法と呼ばれる手法もある．しかし本稿では，模倣学習が誤り伝搬の回避に有効であることを確認するために，

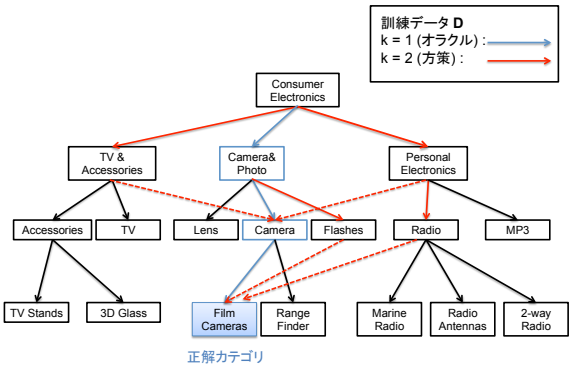

図 1: 方策が訪れる状態のサンプリング例

最も高い確率値（スコア）を持つ枝のみ保持することにする．NLP タスクに模倣学習を適用するにはエキスパート役であるオラクルを設計する必要がある．階層的商品分類におけるオラクルは，状態とは独立に正解カテゴリパスを参照し，現在着目する階層 t の正解カテゴリを返せばよい．つまり，$\pi^*(s_t) = y_t$ である．DAGGER の最初の反復では，オラクル π^* を用いるため，通常の教師あり学習と同じように正解カテゴリパスだけで学習を行う．つまり，オラクルが訪れた状態分布の下で経験誤差の最小化を行う．2 回目の反復以降では，オラクルが訪れた状態に加えて学習した方策が訪れた状態も訓練データに加えて学習する（図 1）．これにより最初の反復では，図 1 の Consumer Electronics から正解カテゴリである Film Cameras までのパスは一通りしか学習事例として存在しなかったが，反復方策によりあらかじめ方策が訪れる状態をサンプリングすることで，過去の予測が誤ることを考慮された複数のパスが学習事例に加わる．ここで，図 1 の点線は，オラクルによって軌道修正されたパスを表す．

4 実験

階層的商品分類において模倣学習の有効性を調査するために，以下の実験を行った．

4.1 実験設定

実験のためのデータセットとして，Rakuten Commerce (RDC)[3] の商品データを用いた．カテゴリは全 6 階層の木構造として構築されており，各商品に対して 1 つのカテゴリが登録されている．商品が所属するの

[3] http://www.rakuten.com

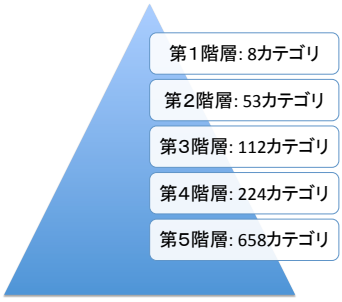

図 2: 訓練データの階層構造

データセット	商品数	末端カテゴリ数
訓練セット	20,000	658
開発セット	2,000	321
評価セット	2,000	303

表 2: 実験で使用したデータセット

はすべて末端カテゴリであり，末端カテゴリは最短 2 階層から最長 6 階層まで合計約 9,000 存在する．本実験では誤り伝搬の低減による精度向上を目的とするため，階層が浅すぎると誤り伝搬の低減が確認しにくいという理由から長さが等しく 5 の階層のみ扱う．これにより対象カテゴリ数は約 1,600 になる．実験で使用した階層的商品分類データセットの詳細を表 2 に示す．また，訓練データの階層構造は図 2 にようになった．

本実験では，商品タイトルのみを用いる．素性には，単語表層 1, 2-gram，また各階層のカテゴリ予測履歴の素性として全予測履歴にあたる 1, 2, 3, 4-gram を用いた．方策の実装にはロジスティック回帰を用い，開発セットを用いて π_k を選択した．また，Flat Model と DAGGER を適用していない Tree Model をベースラインとした．

4.2 評価尺度

実験の評価尺度として，階層的文書分類タスクでも一般的に用いられる Micro F1 に加え，Average Revenue Loss（ARL，収益損失率）[2] を用いた．分類対象が商品であるため，一般的な文書分類タスクで使われている評価尺度だけではシステムの性能を測るのは不十分だと考えるためである．店舗にとっての目的は利益の最大化（損失の最小化）である．ARL は（1）商品が本来所属すべき最適なカテゴリに分類されたとき，ユーザはその商品を支障なく見つけることができるため，店舗はその利益を充分に得ることができる．(2) 商品が誤ったカテゴリに分類された場合，ユーザはその

Model	Micro F1	ARL
Tree Model (ベースライン 1)	70.90	17.85
Tree Model w/ DAGGER (提案手法)	**74.75**	**15.90**
Flat Model (ベースライン 2)	74.30	17.18

表 3: RDC データでのテスト性能比較

商品を見つけて購入することが難しくなることから，店舗は正解カテゴリと誤分類カテゴリとの階層上の距離に比例する損失コストを被るという 2 つの仮説に基づいおり，次式で表される．

$$\frac{1}{m} \sum_{(x,y,y') \in D} v(x) \cdot L_{yy'} \quad (1)$$

ここで，m は全商品数，y, y' はそれぞれ正解カテゴリ，誤分類カテゴリを示している．$v(x)$ は商品 x に対する潜在的な利益で，各商品 x に対する重みを表しているが，本実験では全て等しく 1 とする．$L_{yy'}$ は損失率で，y と y' との階層上の距離に比例する単調増加関数である．

4.3 実験結果

表 3 に実験結果を示す．提案手法は DAGGER を用いない Tree Model であるベースライン 1 に比べて F 値で 3.85 ポイント向上した．表 4 は実際にシステムが出力した中で誤り伝搬を回避できた例である．模倣学習を用いない Tree Model は第 4 階層のカテゴリ予測を Window Treatments ではなく Artwork に誤ってしまい，そのため誤り伝搬が起き第 5 階層でもまた予測を失敗してしまうのがわかる．一方，模倣学習を用いた Tree Model では同様に第 4 階層で予測を誤っているが，続く第 5 階層では誤り伝搬の回避に成功し，正解カテゴリを予測していることが確認できる．これは，反復方策を行なうことで Artwork → Curtains という方策が実際に訪れるであろうパスをサンプリングした下で学習しているためであると考えられる．

また，提案手法は ARL においてもベースライン 1 に比べて 1.95 ポイント低減することができた．提案手法は，各階層でオラクルによって軌道修正されるように設計しているため，模倣学習を用いないベースライン 1 に比べると正解カテゴリと予測カテゴリの平均距離がより短くなるようなモデルであると言える．

4.4 考察

提案手法と既存の Flat Model であるベースライン 2 との比較では，F 値はわずかに上回ったもののほとんど同程度の結果になった．しかし，興味深いことに ARL

Tree Model の予測列	Home & Outdoor → Furniture, Décor & Storage → Décor & Artwork → **Artwork** → Paintings
Tree Model w/ DAGGER の予測列	Home & Outdoor → Furniture, Décor & Storage → Décor & Artwork → **Artwork** → Curtains
正解カテゴリ列	Home & Outdoor → Furniture, Décor & Storage → Décor & Artwork → Window Treatments → Curtains

表 4: 誤り伝搬を回避した例

では 1.28 ポイントの向上が見られた．これはベースライン 2 と提案手法の両方で正しく分類された商品数はほとんど変わらないが，誤分類したときの正解カテゴリとの平均距離がベースライン 2 に比べて提案手法のほうが短いことが言える．また，模倣学習を適用していない Tree Model は Flat Model と比べて F 値では 3 ポイント以上低いにも関わらず，ARL ではわずか 0.7 ポイント程度の差となった．この結果に対する一つの考察としては，Tree Model では階層構造を手がかりに最上層から逐次的に分類することで末端カテゴリを予測するモデルのため，模倣学習の適用とは無関係に階層距離は自ずと近くなるであろうと考えられる．一方で Flat Model の場合，階層構造を分類の際の制約として利用せずに，全末端カテゴリの中から一度に最適なカテゴリを選択するモデルであるため，正解カテゴリと予測カテゴリの距離が階層構造上近くにある保証はない．

一般的に階層的分類タスクにおいては，階層構造を利用しない Flat Model を用いた方が精度が高い場合もあれば，逆に Tree Model のように階層構造を利用して決定的に分類していくほうが精度が高い場合もある．これらは，分類対象の末端カテゴリ数やカテゴリツリーの構造に依存して各手法の向き不向きがあるためである．しかし階層的商品分類では分類対象が商品のため，誤分類の場合にも，できるだけ正解カテゴリに近いカテゴリへの分類を期待されるという特徴がある．そのため店舗の収益損失率という観点からは，Flat Model より Tree Model のような階層構造を利用するモデルの方が望ましい．

5 おわりに

本稿では，模倣学習手法の一つである DAGGER を既存の Tree Model に用いた階層的商品分類モデルを提案した．実際の商品データを用いた実験により，提案手法は誤り伝搬を低減し，より高精度で商品分類ができ，誤分類の場合にもより正解カテゴリに近いカテゴリへの分類が可能なモデルであること示した．さらに，収益損失率という観点から既存の Flat Model と Tree Model に対する分析を行った．今後の課題としては，方策が訪れる状態分布のサンプリングの工夫が挙げられる．また評価実験では長さが等しく中規模なデータセットを用いたが，実応用を見据えて長さが異なる大規模なデータを用いて検証する必要がある．

謝辞

本研究は，楽天技術研究所ニューヨークでのインターンシップの成果である．本研究に対してご指導頂いた楽天技術研究所の諸氏に深謝する．

参考文献

[1] Bennett, P. N. and Nguyen, N. 2009. *Refined Experts: Improving Classification in Large Taxonomies.* In Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'09), pages 11–18.

[2] Chen, J and Warren D. 2013. *Cost-sensitive Learning for Large-scale Hierarchical Classification of Commercial Products.* In Proceedings of 22nd ACM International Conference on Information and Knowledge Management, pages 1351–1360.

[3] Kozareva, Z. 2015. *Everyone Likes Shopping! Multi-class Product Categorization for e-Commerce.* The 2015 Annual Conference of the North Americal Chapter for the ACL, pages 1329–1333.

[4] Ross, S. Geoffrey, J. G. and Bagnell, D. 2011. *A reduction of imitation learning and structured prediction to no-regret online learning.* In Proceedings of 14th International Conference on Artificial Intelligence and Statistics, pages 627–635.

[5] 坪井 祐太. 2013. 模倣学習による決定的解析での誤り伝搬の回避. 言語処理学会第 19 回年次大会.

[6] Wang, X.-L., Zhao, H., and Lu, B.-L.. 2011. *Enhance Top-down Method with Meta-Classification for Very Large-scale Hierarchical Classification.* In Proceedings of 5th International Joint Conference on Natural Language Processing, pages 1089–1097.