分類先の偏りに着目した問合せ文書の自動分類モデルに関する研究

情報数理応用研究

5217C007-1　大窪啓介
指導教員　後藤正幸

A Study on Automatic Document Classification Focusing on Imbalanced Categories of Query Documents

OKUBO Keisuke

1　研究背景・目的

近年，多くの企業では情報システムなどに関する質問やトラブル対応の依頼を従業員がオンラインでサポート部門に送信可能なシステムが導入されるようになった．このような問合せシステム（以下，QA システム）において，Web フォームやメール，チャットなどにより，電子文書形式でやり取りが行われている．これらの問合せ文書に対してはその内容に関わる担当部署に対応してもらう必要があるため，1 件ずつ内容を確認して適切な担当グループへ分類する業務が必須となっている．しかし，問合せは日々大量に寄せられており，これらの分類を人手で処理することは，業務上大きな負担となっている．そのため，問合せ文書を自動的に適切な担当グループへ分類することが可能になれば，企業内の業務効率化にとって有益であると考えられる．

そこで本研究では，某大手企業の社内向け問合せシステムを対象事例とし，問合せ文書の内容から担当グループへ振り分けるための自動分類器を構築することを目的とする．そのため，過去に送信された問合せ文書が分類された担当グループを学習し分類器を構築する．具体的には，問合せ内容に含まれる各単語の頻度ベクトルを説明変数とし，各担当グループを目的変数（以下，カテゴリ）として，Random Forest を用いて分類器の構築を行う．Random Forest は，説明変数が多数であっても性能の良い手法 [1] であり，テキストデータの解析やマーケティングデータの解析に対して有用性が示されている．しかし，本研究で扱う問合せ文書データは，正解である担当グループが特定のカテゴリに大きく偏っており，そのまま分類器を構築すると，所属する問合せ文書数の少ないカテゴリの分類精度が維持できないという問題が生じてしまう．そこで，本研究では，第 1 段階において問合せ件数が多いカテゴリとそれ以外を分類し，第 2 段階以降でそれ以外でまとめたカテゴリについてもさらに分類を行う階層的な分類モデルを提案する．これにより，各段階における学習データの偏りを解消し，かつ分類上，重要な大多数のカテゴリから先に分類することで分類精度の向上が期待できる．最後に，某企業の QA システムに蓄積された問合せ文書の実データに対して提案モデルを適用して検証を行い，その有効性を示すとともに，得られた結果に基づいて考察を行う．

2　準備

2.1　問合せ文書

本節では，本研究で取り扱う問合せ文書及びその対応方法について説明する．問合せ文書は，企業の社内問合せシステムによって，各ユーザから一次窓口である問合せセンターへ随時送られてくる．また，各問合せ文書には，受付番号，ユーザが所属する支店名なども付与されている．この問合せ文書に対する担当部署を表すカテゴリには大きな偏りがあり，単純な方法では所属する問合せ文書数の少ないカテゴリへの分類精度が低くなってしまうという問題が生じる．具体例として，2018 年 3 月に送られてきた問合せデータにおける代表的なカテゴリを表 1 に示す．表 1 より，「企業支援デスク」は全体の約 28%，「PC サポートデスク」は約 23%で，これら 2 つのカテゴリで全体の 50%以上を占めていることが分かる．

表 1．問合せ文書の所属する主なカテゴリの割合

カテゴリ	割合（単位：%）
企業支援デスク	28.47
PC サポートデスク	23.01
アカウントシステムサポート	10.57
その他（15 カテゴリ）	37.94

2.2　関連研究

問合せ文書の分析については，これまでにもいくつかの研究が行われている．例えば，Kang ら [2] は，ユーザからの問合せ文書に対して適切な回答文書を検索するために，問合せのリンク情報と URL 情報を考慮した手法を提案し，高い精度が得られている．また，Torres ら [3] は，ユーザからの問合せに対する適切な応答を提供するにあたって，複数の応答に該当する可能性を持つ問合せに対処するために，Bag of Words を用いて問合せの特徴を表してから分類を行う手法を提案している．これらの研究は，問合せに対して適切な回答を提示することを目的としている．これに対して本研究では，問合せ文書を適切なカテゴリへ分類することを目指す．

2.3　予備実験

本研究の対象事例では，所属するデータ数によってカテゴリに大きな偏りがある．このようなアンバランスなデータに対する一般的な対策として，少数のカテゴリに対して不足データを補完することで，アンバランスを解消するオーバーサンプリングがある．このオーバーサンプリングの代表的な手法として，Synthetic Minority Oversampling Technique[4]（以下，SMOTE）がある．SMOTE とは，サンプルをコピーするのではなく，異なる値のサンプルを新しく生成してサンプルを増やす手法である．そこで，本研究における予備実験として，問合せデータに対して Random Forest を用いて多値分類を行う際に，事前に SMOTE を行う場合と行わない場合に対してそれぞれ実験を行うことで，オーバーサンプリングの効果を検証した．

表 2．予備実験における各評価指標の推移

手法	データ数	適合率	再現率	F 値
Random Forest（SMOTE なし）	35,182	0.5674	0.5763	0.5763
Random Forest（SMOTE あり）	180,324	0.5674	0.5999	0.5832

表 2 より，分類を行う前に SMOTE を実行すると，再現率はわずかに増加するが，適合率と F 値がほとんど改善していないことがわかる．このことから，問合せ文書の偏りが非常に大きいため，オーバーサンプリングでは対応しきれていないことが考えられる．このため，本研究では，あらかじめカテゴリをグループ化し学習することで改善することを考える．

3　提案手法

3.1　着想

本研究では，まず，過去の問合せデータの文書に対して学習を行い，その結果を現在の問合せデータに適用す

ることで，それぞれの問合せに該当するカテゴリの予測を行う．しかし，扱う問合せデータは，全体のうち「企業支援デスク」と「PC サポートデスク」が 50 %以上を占めるなど，一部のカテゴリが多くの割合を占めている．このため，そのまま全体から分類器を学習すると，多数を占めるカテゴリへ分類されるため，データ数の少ないカテゴリの分類精度が維持できないという問題がある．加えて，2.3 節で示したように単純なオーバーサンプリングでは対応できない．そこで本研究では，これらの問題を解決するための階層的な分類モデルを提案する．

3.2 分類器の構成

まず，データ数の多いカテゴリ（多数カテゴリ）を選択し，そのカテゴリとその他で分類を行う（階層 1）．次に，階層 1 でその他のカテゴリへ分類されたデータに対し再度分類を行う（階層 2）．ここで，選択するカテゴリ数によって様々なパターンがあるため，本研究では以下の 4 つのパターンを考える．

- パターン A : 第 1 段階（階層 1）でデータ数が最も多いカテゴリとそれ以外，第 2 段階（階層 2）でそれ以外を分類する方法 (図 1)
- パターン B : 第 1 段階で上位 2 つのカテゴリとそれ以外，第 2 段階でそれ以外を分類する方法
- パターン C : 第 1 段階で上位 3 つのカテゴリとそれ以外，第 2 段階でそれ以外を分類する方法
- パターン D : 1 つのカテゴリとそれ以外とした後で，残りのカテゴリの中で上位 1 つとそれ以外という分類を多段階に繰り返す方法

ここで，階層 2 の分類を行うにあたって，階層 1 においてその他のカテゴリへ分類されたものに，誤って分類されてきたデータが含まれている可能性がある．そこで，階層 1 で選択した件数の多いカテゴリについても階層 2 以降で分類を行う必要がある．本研究では，パターン A からパターン D に対して，階層 2 以降の分類にあたり，前階層での結果を反映した分類器の学習法を提案する．ここで，パターン A の分類の概要図を以下の図 1 に示す．なお，図 1 における 1 から 18 までの数字は，データ数の降順に並べたカテゴリ番号を表している．

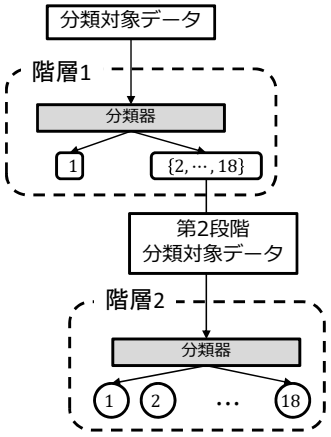

図 1: パターン A における分類の概要図

3.3 分類器の学習

第 2 段階以降の分類は，階層 1 の分類器で「多数カテゴリ以外」と判定されたデータが分類対象となる．したがって，その分類器の学習に用いる学習データの選定にはいくつかのバリエーションが考えられる．本研究では，階層 1 の分類結果を元に階層 2 の分類器を学習する方法として，以下の 2 つの方法を提案する．

提案 1 階層 2 以降における学習の際に，階層 1 で「その他」に分類されたデータを用いて学習する方法

提案 2 階層 2 以降における学習の際に，提案 1 のデータに加え，階層 1 で誤って「多数カテゴリ」に分類されてしまったデータを併せて用いる方法

なお，それぞれの提案の詳細については次節以降で述べる．また，本稿では階層の作り方のパターン A–D と，学習データの選定法 1，2 の組合せによって，提案 A–1，A–2 のように記述する．

3.3.1 提案 1

階層 2 以降の学習を行うにあたって，階層 1 においてその他のカテゴリへ分類された学習データの中に，選択した多数カテゴリの学習データが誤って含まれている可能性がある．前階層での結果を反映した分類器を作成するためには，このような誤って分類された学習データを含めて学習をしなければならない．そのため，階層 2 以降でも全てのカテゴリを分類できる分類器が必要である．そこで，本研究の提案 1 では，階層 2 以降の学習の際に，階層 1 で「その他」に分類されたデータを全て用いて学習を行う．これらのデータには階層 1 で誤って分類された多数カテゴリの学習データも含まれている．ここで具体例として，提案 A–1 における学習と分類の流れについて，図 2 に示す．なお，図 2 における灰色のカテゴリは，階層 1 において学習データを構築した分類器へ入力した後の真のラベルを表している．

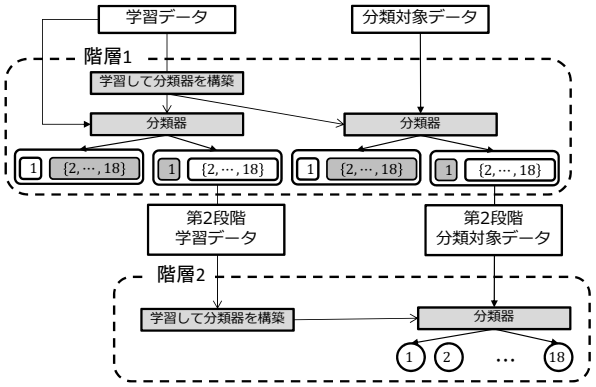

図 2: 提案 A–1 の概要図

3.3.2 提案 2

階層 2 の学習を行うにあたって，階層 1 で「その他」に分類されるデータのみを使うと，誤って「多数カテゴリ」に分類されたデータを階層 2 の学習に使われないことになり，データが不足する可能性がある．そのため，そのまま階層 2 の学習を実行すると，データ数が元のデータ数より少ない状態で学習してしまう．そこで本研究では，提案 A–1 から D–1 に対して，階層 2 以降の学習にあたり，前の階層で正しく学習しきれなかったそれ以外のカテゴリを元に戻した状態で学習を行う手法 (それぞれ，提案 A–2（図 3）–提案 D–2（図 4）) をそれぞれ提案する．

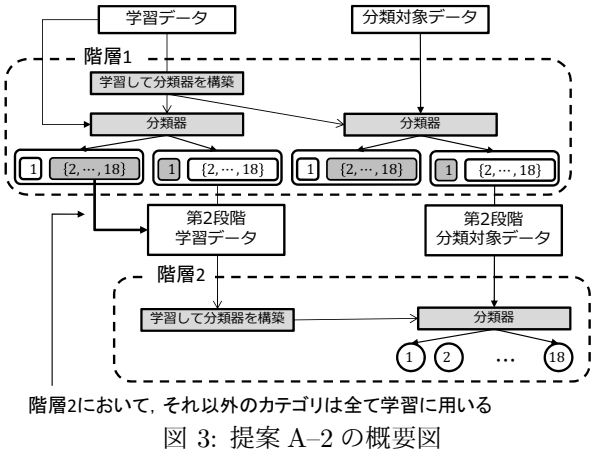

図 3: 提案 A–2 の概要図

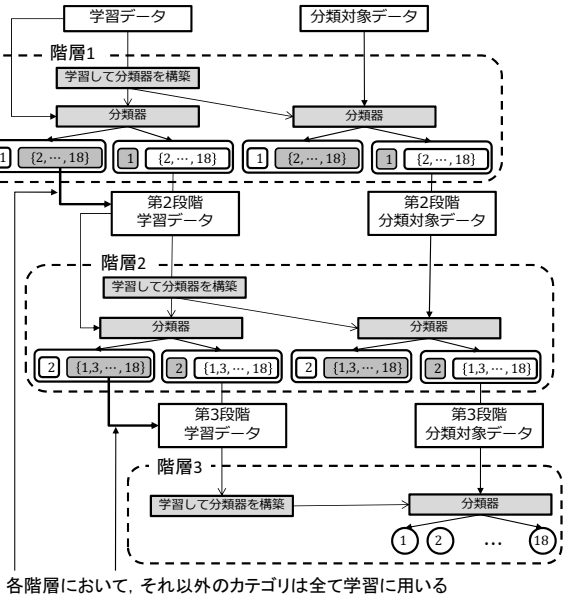

図 4: 提案 D–2 の概要図

4 実証実験

本章では，前節で述べた 4 通りの提案の枠組みを実際の問合せ履歴データに対して適用することで，どの方法が有効かを検証し，その結果を述べる．

4.1 実験条件

本研究では，日本の某一部上場企業における社内問合せシステムの履歴データを用いる．具体的には，実際に送られてきた問合せ文書とその所属するカテゴリのデータを学習し，その結果を元に新規の問合せ文書がどのカテゴリに属するかの分類を行う．本研究では，2017 年分のデータを学習データとし，2018 年 3 月のデータをテストデータとして用いる．ここで，検証用データにおいて，各提案を元に選択した多数カテゴリとそれ以外のカテゴリの元のデータ数を以下に示す．

表 3. 2018 年 3 月分のカテゴリ毎のデータ数

手法	カテゴリ番号	データ数
提案 A–1，提案 A–2（1 対 17）	1	10,131
	2–18	25,869
提案 B–1，提案 B–2（2 対 16）	1,2	19,043
	3–18	16,597
提案 C–1，提案 C–2（3 対 15）	1–3	23,596
	4–18	12,404
提案 D–1，提案 D–2（3 階層）	1	10,131
	2	8,912
	3–18	16,957

ここで，分類手法は Random Forest[1] を用い，Grid Search によりパラメータを決定した．また，比較手法としては，あらかじめカテゴリを分けずにそのまま多値分類を行う手法を用いる．さらに，評価指標については，適合率，再現率，F 値 [5] を用いる．

本研究では，1 つ 1 つのカテゴリに対して正しく分類されているかを求め，それらの平均を全体の評価指標として用いる．ここで，平均を求める際に，全体での平均（ミクロ平均）を用いると多数のカテゴリの結果が強く反映されるため，本研究では全階層により出力された各カテゴリ毎の評価指標（マクロ平均）を用いる．また，本研究では，各文書の単語頻度ベクトルを構成するために，問合せ文書に対して MeCab を用いて形態素解析を行い，名詞と動詞の単語を抽出した．そして，各文書におけるこれらの単語の頻度を元に，単語頻度ベクトルを作成し，これを学習及びテストデータとして用いる．

4.2 実験結果

4.2.1 各提案手法による分類精度の結果

まず，それぞれの手法による分類精度の結果を表 4 に示す．

表 4. 各評価指標の推移

手法		適合率	再現率	F 値
比較手法（18）		0.5674	0.5999	0.5832
2 階層	提案 A–1（1 対 17）	0.5741	0.5888	0.5810
	提案 A–2（1 対 17）	0.6248	0.6121	0.6087
	提案 B–1（2 対 16）	0.5536	0.6765	0.6093
	提案 B–2（2 対 16）	0.6090	0.7317	0.6610
	提案 C–1（3 対 15）	0.5367	0.7219	0.6165
	提案 C–2（3 対 15）	0.5974	**0.7831**	0.6746
3 階層	提案 D–1（1 対 1 対 16）	**0.8840**	0.4914	0.6291
	提案 D–2（1 対 1 対 16）	0.8468	0.5749	**0.6807**

ここで，表 4 における「比較手法」とは，18 個のカテゴリをそのまま多値分類する方法の分類結果を示している．表 4 より，本研究で提案した手法のうち，提案 A–2 ～提案 D–2 の評価指標はいずれも提案 A–1～提案 D–1 及び比較手法の指標を上回っていることが分かる．このことから，本研究の問題に対して，各階層において前の階層で正しく判別できなかった「それ以外のカテゴリ」を元に戻して学習データとする方法が有効であると考えられる．また，これらの結果より，提案手法として挙げたパターンの中では提案 D–2 が最も有効であるといえる．次に，本研究で提案した手法のうち 1 対 17 の適合率については，提案 1，2 共に比較手法と比べて上回っていることが分かる．これは，前者はあらかじめ割合の大きいカテゴリを先に分類することで，それ以外のカテゴリがデータの偏りによる影響を受けることなく分類できたためと考えられる．また，提案 D–1 と D–2 の適合率及び F 値は，比較手法及び提案 A–1 や A–2 と比べて上回っていることが分かる．これは，階層 1 で選択した多数カテゴリが上手く分類できているため，多数カテゴリに含まれるデータがそれ以外のカテゴリの中に誤って分類されることがほぼなかったことが考えられる．さらに，これらの結果より，適合率は提案 D–1 の時が最もよくなっていることが分かる．このことから，1 対 17 を元に階層的に分類していく手法は，本研究で扱った問合せデータのようなカテゴリに偏りのあるデータに対して有効であると考えられる．次に，提案手法のうち，1 対 17 の再現率は比較手法と比べてあまり変わらず，2 対 16 と 3 対 15 の再現率は従来手法より大幅に高くなった．また，これらの結果より，再現率は提案 C–2 の時が最もよくなっていることが分かる．これは，カテゴリ 2 と 3 の問合せ文書には分類に寄与する特徴的な単語が含まれているため，分類が容易であったためだと考えられる．一方で，提案手法について 2 階層の結果を比較すると，1 対 17 の適合率が最も高く，3 対 15 の適合率が最も低いことが分かる．次に，それぞれの提案手法の分類において，各カテゴリがどれだけ正しく分類されたかについて結果を示す．

4.2.2 各カテゴリの実験結果（パターン A）

パターン A における各カテゴリの評価指標の結果を表 5 に示す．

表 5. パターン A における各カテゴリの評価指標の比較

カテゴリ	手法	分類されたデータ数	適合率	再現率	F 値
階層 1	提案 A–1	1,891	0.8398	0.1826	0.2975
	提案 A–2	2,542	0.7492	0.2346	0.3573
階層 2	提案 A–1	34,109	0.5594	0.5825	0.5707
	提案 A–2	33,458	0.6153	0.6408	0.6278

ここで，表 5 における「階層 1」は，階層 1 で分類された多数カテゴリであるカテゴリ 1 に対する評価指標を示している．また「階層 2」は，階層 2 において，それ以外のカテゴリと階層 1 で分類しきれなかったカテゴリ 1 に対する評価指標の平均を示している．表 5 より，多数

カテゴリの再現率は比較手法と比べて大幅に下回っていることから，提案 A–1 及び提案 A–2 では，階層 1 における分類の際にカテゴリ 1 を分類しきれていないことが分かる．この理由としては，選択されていない 17 カテゴリの集合は，統計的性質の異なるカテゴリの寄せ集めであるため特徴のない集合になってしまっており，二値分類器で識別しにくいことが考えられる．また，多数カテゴリにおいて提案 A–1 と A–2 を比較すると，適合率は提案 A–1 が上回っているものの，再現率と F 値は提案 A–2 が上回っていることがわかる．これについては，多数カテゴリは統計的性質の異なるカテゴリの寄せ集めであるため，提案 A–2 によって特徴のない集合もある程度分類できたためだとと考えられる．

4.2.3 各カテゴリの実験結果（パターン B，C）

パターン B とパターン C における各カテゴリの評価指標の結果をそれぞれ表 6，表 7 に示す．

表 6．パターン B における各カテゴリの評価指標の比較

カテゴリ	手法	分類されたデータ数	適合率	再現率	F 値
階層 1	提案 B–1	20,497	0.5549	0.7850	0.6513
	提案 B–2	21,086	0.6104	0.8243	0.7014
階層 2	提案 B–1	15,503	0.5518	0.5723	0.5617
	提案 B–2	14,914	0.6070	0.6009	0.6039

表 7．パターン C における各カテゴリの評価指標の比較

カテゴリ	手法	分類されたデータ数	適合率	再現率	F 値
階層 1	提案 C–1	29,975	0.4835	0.6900	0.5695
	提案 C–2	27,984	0.5319	0.7590	0.6254
階層 2	提案 C–1	6,025	0.8000	0.8374	0.8182
	提案 C–2	8,016	0.8264	0.8673	0.8464

表 6，7 より，パターン B，C のいずれの提案も選択した多数カテゴリにおける再現率が比較手法と比べて大幅に増加していることが分かる．また，それ以外のカテゴリについても，適合率，再現率ともに比較手法と比べて上回っていることが分かる．この結果と F 値より，複数のカテゴリを合わせて選択して分類した場合，1 つ選択する場合と比べて各カテゴリにおける分類が上手くいくことが分かる．これについては，多数カテゴリが全データの大部分を占めていたため，それ以外のカテゴリのみでの分類がデータの偏りによる影響を受けなかったためと考えられる．さらに，表 4 や表 5 と比較してみると，選択するカテゴリ数が増えるにつれて適合率が減少していることが分かる．このことから，選択するカテゴリを増やしていくと上手く分類できないことが考えられる．

4.2.4 各カテゴリの実験結果（パターン D）

パターン D における各カテゴリの評価指標の結果を表 8 に示す．

表 8．パターン D における各カテゴリの評価指標の比較

カテゴリ	手法	分類されたデータ数	適合率	再現率	F 値
階層 1	提案 D–1	1,891	0.8398	0.1826	0.2975
	提案 D–2	2,542	0.7492	0.2346	0.3573
階層 2	提案 D–1	6,281	0.8452	0.5414	0.6580
	提案 D–2	6,653	0.8940	0.5987	0.7171
階層 3	提案 D–1	27,828	0.8979	0.4991	0.6391
	提案 D–2	26,805	0.8443	0.6012	0.7023

表 8 より，提案 D–1，D–2 ともに最初に選択した多数カテゴリと，2 番目に選択した多数カテゴリ，それ以外の適合率はいずれも比較手法と比べて大きく上回っており，上手く分類できていることが分かる．この理由としては，それぞれの多数カテゴリにおける問合せ文書のうち，正確に分類できた文書には分類に寄与する特徴的な単語が含まれているため，分類が容易であったためだと考えられる．また，それ以外の適合率は最初に選択した多数カテゴリと 2 番目に選択した多数カテゴリを上回っていることが分かる．この理由として，多数カテゴリがいずれも上手く分類できているため，多数カテゴリを持つデータがそれ以外のカテゴリを持つデータの中に誤って分類されることがほぼなかったことが考えられる．一方，各カテゴリの再現率は比較手法の再現率を下回っていることが分かる．これは，多数カテゴリを持つ文書の内，分類に寄与する特徴的な単語が含まれていない文書が正確に分類されていないためだと考えられる．

5 考察

まず，表 4, 6, 7 より，パターン B と C についてはどの提案手法についても，F 値が比較手法を上回る結果が得られた．しかし，パターン A とパターン D の結果のように，カテゴリ 1 が含まれていると上手く分類できない結果となった．これについては，階層 1 で選択した多数カテゴリでは，分類に寄与する実用語を含む文書がどれだけ含まれているかが影響していると考えられる．例えば，カテゴリ 2 と 3 には分類に寄与する用語が含まれている文書が大半を占めていたことが予想される．一方で，カテゴリ 1 はどのカテゴリにも該当しないデータが含まれており，そのようなデータは分類に寄与する実用語が含まれていないことが考えられる．このことは，本研究における階層 1 においてカテゴリを選択する際は，分類に寄与する用語について加味することが重要であることを示唆する．次に，表 4〜8 より，提案 A–2〜提案 D–2 の評価指標はいずれも提案 A–1〜提案 D–1 及び比較手法の指標を上回る結果が得られた．これについては，階層 2 においてそれ以外のカテゴリを学習する際に，元のデータ数で学習できたためだと考えられる．このことから，本研究の問題に対して，本研究における階層 2 以降において学習を行う際は，それ以外のカテゴリを全て考慮する手法が有効であると考えられる．さらに，これらの結果により，本研究のような偏りのあるデータに対する階層的手法においては，パターン A〜D のような階層の作り方と，提案 2 のような階層 2 以降の学習データの作り方とでは，後者の方がより有効であると考えられる．

6 まとめと今後の課題

本稿では，特定のカテゴリにデータの偏りがある問題に対して，データの偏りによる影響を少なくするような分類手法の枠組みを提案した．また，提案手法の有効性を検証するために、実際の問合せ文書データを用いて実験を行った．その際に提案手法としていくつかのパターンを用意し，どのパターンが有効かを検証した．その結果，提案手法は本研究で扱った問合せ文書のように，一部のカテゴリが大きい偏りのあるデータに対して効果的であると考えられる．今後の課題としては，4 階層以上の二値分類を行うことや，3 階層の分類において分類する順番を入れ替えて行うことなどが挙げられる．

参考文献

[1] Breiman L, “Random Forests,” *Machine Learning*, Vol.45, No.1, pp.5–32, 2017.

[2] In-Ho Kang, GilChang Kim, “Query type classification for web document retrieval,” *Proceeding SIGIR ’03 Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pp.64–71, 2003.

[3] Rafael Torres, Shota Takeuchi, Hiromichi Kawanami, Tomoko Matsui, Hiroshi Saruwatari and Kiyohiro Shikano, “Inquiry Classification in a Speech-Oriented Guidance System Using Discriminative Learning,” *IPSJ SIG Technical Report*, Vol. 2009-SLP-77 No.13, pp.1–6, 2009.

[4] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall and W Philip Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” *Journal of Artificial Intelligence Research*, 16, pp.321–357, 2002.

[5] 平井有三, “はじめてのパターン認識”, 森北出版, 2012.