ARG WI2 No.11, 2017
WI2-2017-24

量的不均衡データに対する学習精度改善のための文書かさ増し手法

澤崎 夏希[†,a]　遠藤 聡志[††,b]　當間 愛晃[††]
山田 孝治[††]　赤嶺 有平[††]

† 琉球大学理工学研究科情報工学専攻　†† 琉球大学工学部知能情報コース

a) *k178577@ie.u-ryukyu.ac.jp*　b) *endo@ie.u-ryukyu.ac.jp*

概要　機械学習アルゴリズムが特徴量そのものを学習することで様々な領域での問題解決にブレークスルーが起こっている．テキスト分類の問題領域でも，多くの場合．高い分類精度を達成している．しかし成功例の多くは各正解ラベルのデータ量が均一あるいはそれに近い状態であることが多い．このため．すべての正解ラベルデータを十分量用意するためのコストが機械学習アプローチのボトルネックとなっている．また．ラベル毎のデータ量が不揃いな場合は不均衡データと呼ばれうまく分類できないことが知られている．本研究では，自然言語の不均衡データに対するかさ増し手法を提案する．提案手法を用いた，不均衡データ分類問題の計算実験を行い．分類精度の検証によってその有用性を評価する．

キーワード　かさ増し, 自然言語処理, 不均衡データ, オーバーサンプリング

1　はじめに

現在，様々な問題が機械学習により解決されているが，その多くはデータ数の揃った均衡なデータであり不均衡データについてはまだ解決すべき問題が残っていることが知られている [1]．不均衡データについて網羅的に調査した Haibo らによると不均衡データへの対策としてサンプリング手法やカーネルベースの手法などいくつか効果を発揮している手法は存在するが，どれも特定のデータセットのみを対象としており，一般的な解決には至っていない．ただし，不均衡データの傾向として，データ数を揃えることは学習器本来の性能を発揮しやすく，精度の向上が見込めるとも述べている．そこで本研究では不均衡データに対して，サンプリング手法を用いての精度改善を行う．

自然言語に対して，データを減少させるダウンサンプリング手法は広く知られている．しかし現在，自然言語データに対してのオーバーサンプリング手法は確立されていないことを奥野ら [2] は指摘している．奥野らは多数のラベルを持つ不均衡な日本語のデータを分類する研究で，データの不均衡さから機械学習では十分な精度を得られないと判断し類似度による分類手法を採用している．類似度による分類手法は人手で特徴を設定するため，一つ一つの問題解決に時間がかかり一般化が難しい．一方で，自動的に特徴を獲得する機械学習は不均衡データに対して特徴を獲得出来ない場合がある．

そこで本研究では，自然言語におけるオーバーサンプリング手法を提案する．オーバーサンプリングはデータに対して特定の操作を行うことで，近い特徴を持つ異なるデータを生成する手法で，新規にデータを収集することなく対象のカテゴリのデータ数を増加させることを目的とする．これにより不均衡データ数を揃え，機械学習での分類精度が向上させることができる．自然言語に対するオーバーサンプリング手法の一つとしてかさ増し手法を提案する．かさ増し手法は大きく，単語の入替えと文節の入替えを 2 つの手法を採用した．これは機械的な操作が実装しやすいことと，効果を観察しやすいことを重視したためである．かさ増し手法を適用した際，増加したデータが有用なものであるか確認することが望ましいが，増加したデータ全てを検証するのは現実的ではない．そのため，今回は精度の面からかさ増し手法の有用性を検証する．

評価対象として用いる livedoor ニュースコーパスのうち，データ量が少ないカテゴリである livedoor_HOMME に注目した．このカテゴリについてかさ増しを行い，分類実験を行うことで，かさ増し手法の効果を検証する．ベクトル化には形態素解析した単語ベクトルに，Bag-of-Words に基づいた TF-IDF を適用したもの，Doc2vec[5] 適用したものを用い，分類にはランダムフォレスト，SVM，ナイーブベイズ分類器，cos 類似度を用いて実験を行う．ベクトル化手法の精度を見るため，TF-IDF にはナイーブベイズを，Doc2vec には cos 類似度を用いて分類精度を測る．

本論文では，かさ増し手法のアルゴリズムを紹介し，日本語としての成立度合いを見るため，かさ増し手法前後の文章，頻出単語の変化を見る．次に手法前後で精度変化をみる．最後にかさ増し手法を複数併用した場合の精度変化を見ることで，かさ増し手法の有用性を検証する．

2 かさ増し手法

サンプリング手法手法として大きく，データを少数に合わせるダウンサンプリング，データを多数に合わせるオーバーサンプリングが用いられる．ただし自然言語処理におけるオーバーサンプリングの手法は確立されておらず，人手による所が大きい．これはカテゴリに含まれる特徴が，増加後のデータにも含まれているかという検証が難しいためである．今回の提案手法では以下の 3 つの手法を用いて，自動的にデータ量のかさ増しを行う手法を提案し，精度変化を元に評価する．提案手法は図 1 の通り 3 つである．

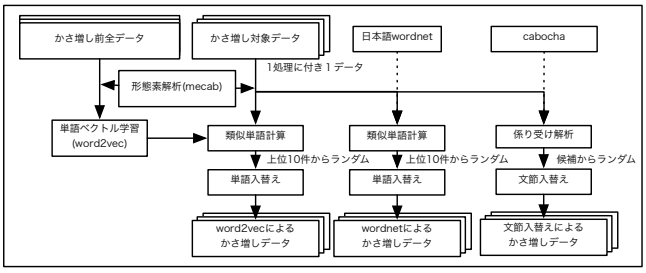

図 1 かさ増し手法

2.1 類似単語入れ替え手法

かさ増しの手法として，類似単語を特定し入れ替える手法を採用した．類似単語を特定するための手法として word2vec[4] による類似度の計算，wordnet による類似単語入替えを用いる．

2.1.1 word2vec を用いた入れ替え

データセット全体に対して word2vec を用いて単語ベクトルを計算し，形容詞を対象に単語の入れ替えを行った．類似度は単語ベクトル同士の cos 類似度で計算され，距離が近い候補を類似単語候補とし，その中からランダムに単語を入れ替えてかさ増しを行っている．

word2vec は自動的に単語ベクトルを計算する．そのメリットとして，類義語が設定されていない単語が頻出するデータにもかさ増し可能であること，かさ増しによるデータの増加量が多いことがある．一方で，周辺単語を元に類似度を計算するため，ある程度の文章量が必要になる．また，単語の使われ方を元に類似度を計算するため，対義語も類似単語として推薦するなど文章の意味を保存したかさ増しを行うのが難しいことがあげられる．

2.1.2 wordnet を用いた入れ替え

wordnet とは人手で設定された単語の構造ネットワークである．今回は，その中から同義語を抽出し入れ替え候補として設定した．英語版 wordnet を翻訳することによって作られているため，数%のノイズが含まれることが公式に指摘されている．

単語に対して wordnet から類義語を取り出し，入れ替え単語として設定する．人手で類義語が定義されているため，多くの場合意味を保存したかさ増しを行うことが可能である．

2.2 文節並び替え手法

文章には入れ替えても意味の変化が無い並列な文節が存在する場合がある．文章を 1 文に切り分け，係り受け解析器 cabocha を用いることで，並列な文節を計算し入れ替えることで同義な文章を生成しかさ増しを行う．文章的に違和感の無いかさ増しが行われる場合が多く，同義文の生成を行えるが，単語の語彙は増えないため，TF-IDF などの単語を元にしたベクトル化ではないと効果を発揮しない可能性がある．

3 livedoor ニュースコーパス

データセットは NHN Japan 株式会社から提供されている livedoor ニュースコーパス [3] を用いる．livedoor ニュースコーパスは livedoor ニュースから可能な限り HTML タグを取り除いて作成したものであり，9 つのニュースカテゴリを含んでいる．livedoor ニュースコーパスは不均衡データであるが，さらに今回は表 1 のようにデータ数を調整した．単一のカテゴリにかさ増し手法を適用することで，精度変化とかさ増し手法の影響が観察しやすくなると考えた．本論文では livedoor_HOMME に対してかさ増しを行い，全てのカテゴリのデータ数を揃えた状態とかさ増しを行う前とでどのような変化が起きたかをみる．

カテゴリ名	調整前	調整後	カテゴリ名	調整前	調整後
dokujo-tsushin	871	770	peachy	843	770
it-life-hack	871	770	smax	871	770
livedoor_HOMME	512	500	sports-watch	901	500
movie-enter	871	770	topic-news	771	770
kaden-channel	865	770			

表 1 調整前後の livedoor ニュースコーパス

4 提案手法によるデータセットの変化分析

提案手法を用いて，調整後の livedoor_HOMME のデータ数を 270 件増やし，全てのカテゴリのデータ数を 770 件に揃えた．どのかさ増し手法でも複数のかさ増し候補が存在するが，今回はそれぞれの候補の中からランダムに選択し増加させている．かさ増し手法で懸念されるのは同じかさ増し結果になってしまい，結果としてデータを複製しただけになってしまうことである．今回のかさ増し手法はどれも 1 文に対して新たな 1 文を生成しているため，1 記事辺りの文章量の多いニュース記事のようなデータに対しては，ランダムでも十分にかさ増しデータの多様性が保たれると考えた．

4.1 提案手法前後の頻出単語の変化

かさ増し手法前後の変化を頻出単語の変化からみる．livedoor_HOMME に対してかさ増し手法を用い，かさ増し手法毎に頻度が高い上位 16 単語を抽出した (表 2)．これは，word2vec では語彙の変化が少なく単体では意味を解釈しにくい単語が上位に来てること，wordnet ではかさ増し前と比較し語彙が増えていること，文節入替えでは語彙に変化が無いこと，を確認するためである．

かさ増し前	word2vec	wordnet	文節入替え
いい	いい	ない	いい
ない	ない	いい	ない
多い	多い	好い	多い
高い	高い	多い	高い
新しい	なく	問題	問題
なく	良い	高い	新しい
問題	やすい	おびただしい	なく
良い	新しい	無い	良い
やすい	多く	新しい	やすい
多く	欲しい	おっきい	多く
欲しい	すごく	なく	欲しい
強い	問題	欲しい	大きく
大きく	強い	良い	すごく
すごく	ほしい	やすい	強い
ほしい	大きく	違い	ほしい
違い	ぽ	多く	違い

表 2 livedoor_HOMME に対するかさ増し毎の高頻度単語

word2vec を用いたかさ増し後の特徴として，かさ増し前に高頻度だった単語が多く受け継がれているが,「ぽ」などの単体では意味の取れない単語も高頻度になっていることがわかる．また，wordnet を用いたかさ増し手法の特徴として，かさ増し前には見られなかった「おびただしい」「無い」「おっきい」のような単語が表れていることわかる．

文節の入替え手法では変化が見られない単語もあるものの順位の変動が起きている．これはかさ増し手法によってデータ量が変化し単語の頻度が変化したためだと考えられる．このことから，単語を中心に見る TF-IDF を用いても文節の入替えによってわずかな変化が見られる可能性がある．

4.2 文章の変化

livedoor_HOMME にかさ増し手法を適用したことにより，文章は図 2 のように変化した．かさ増し手法によって変化した箇所を太字と下線で表す．word2vec を用いたかさ増し手法で日本語として成立していない文章が，wordnet を用いたかさ増し手法では比較的日本語として成立している文章が生成された．文節入替えを用いたかさ増し手法も文章の意味に大きな変化は見られなかった．

【かさ増し前データ】

”黒字社員の特徴も詳しく解説しています。”

”採用面接の際には．何気ない会話の中からビジネスセンスが見抜かれて”

【word2vec によるかさ増し後データ】

”黒字社員の特徴も**ガライヤ**解説しています。”

”採用面接の際には．**しぐさ**会話の中からビジネスセンスが見抜かれて”

【wordnet によるかさ増し後データ】

”黒字社員の特徴も**具 (つぶさ) に**解説しています。”

”採用面接の際には．**何げない**会話の中からビジネスセンスが見抜かれて”

【文節入替えによるかさ増し後データ】

”**詳しく 黒字社員の特徴**も解説しています”

”採用面接の際には．**ビジネスセンスが 何気ない会話の中から**見抜かれて”

図 2 かさ増し前後データ

word2vec を用いたかさ増しで生成されたデータは「詳しく」と「ガライヤ」が交換されている．これは両単語の周辺に出現する単語，例えば「解説」などが同じであるため類義語として選択されている．これは一般的な意味的な類義語ではなく，使われ方の類似性を表している．今回 word2vec によるかさ増し手法を用いたのは，かさ増しするデータが日本語として成立している必要があるのかを調べる目的がある．wor2vec で獲得した使われ方の類似性は機械学習によって自動的に獲得された特徴である．もしこの特徴が有効なら必ずしもデータが日本語として成立している必要はないといえる可能性がある．

一方で wordnet は人手で設定された類義語を元に類似単語を選択している.「詳しく」が「具に」に置換されたように，日本語としての類義語が選択される．word2vec によるかさ増し手法が有効でなく，wordnet によるかさ増しが有効である場合，増加させるデータは日本語として成立している方が望ましい可能性がある．

文節入替えによるかさ増しは並列な文節を入替えているため，他 2 つの手法と異なり語彙が増えない．例では「黒字社員の特徴も」と「詳しく」が入れ替わっている．この手法は語彙を変えず，文章の表現方法を変更させることでカテゴリが持つ特徴に多様性をもたせようとするものである．これは日本語として成立しているが語彙が増えないようなデータがある場合，精度にどう表れるかをみる目的がある．

5 実験

実験はベクトル化に TF-IDF，Doc2vec，分類器に SVM，ランダムフォレストを用いる．また，TF-IDF でベクト

ル化したデータにはナイーブベイズでの分類，Doc2vec でベクトル化したデータには cos 類似度での分類実験も，あわせて行った．

分類を行うデータは表 1 に示した不均衡データ，ダウンサンプリングを行いデータ数を 500 に統一したデータ，word2vec によるかさ増しを行ったデータ，wordnet によるかさ増しを行ったデータ，文節入れ替えによるかさ増しを行ったデータを用いる．それぞれ 6 割のデータを教師データとし，残りの 4 割をテストデータとして精度を測定する．

word2vec によるかさ増し手法を提案手法 1，wordnet によるかさ増し手法を提案手法 2，文節入れ替えによるかさ増し手法を提案手法 3 とし，実験全体の流れを図 3 に示す．表内の NB はナイーブベイズを，cos はコサイン類似度を，RF はランダムフォレストを表す．

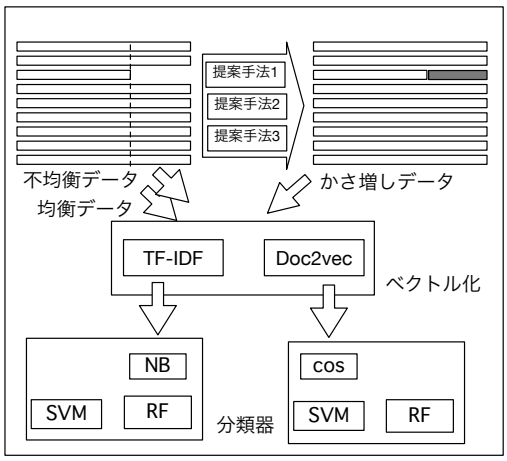

図 3 実験全体図

5.1 実験目的

実験では，かさ増し手法が精度に与える変化をみることを目的とする．不均衡データはこのままでは学習が難しいことを確認するために用い，データ数を 500 に揃えたものは均一化した場合の精度を見るために用いる．提案手法 1 は機械的なかさ増しが精度に与える影響をみるために，提案手法 2 では人手で設定された類義語が与える変化をみるために，提案手法 3 では語彙を増やさずにデータを増やした場合の変化をみるために用いる．

5.2 実験結果

表 3 にかさ増し前のデータ，500 に均一化してダウンサンプリングを行ったデータ，提案手法によりかさ増ししたデータそれぞれ実験の精度の一覧を示す．

まず全体的な精度では TF-IDF が最大 94%，最小 85%，Doc2vec が最大 85%，最小 51%と TF-IDF の方が高くなっている．これは，ニュースカテゴリを構成する特徴として単語の特徴が有効であることを示している．

			全体精度		
ベクトル	実験手法		NB	SVM	RF
TF-IDF	かさ増し前	不均衡データ時	0.85	0.93	0.89
		500 均一データ	0.88	0.90	0.92
	かさ増し後	提案手法 1	0.89	0.94	0.91
		提案手法 2	0.91	0.94	0.92
		提案手法 3	0.86	0.93	0.90
			cos	SVM	RF
Doc2vec	かさ増し前	不均衡データ時	0.51	0.69	0.67
		500 均一データ	0.53	0.73	0.72
	かさ増し後	提案手法 1	0.68	0.84	0.84
		提案手法 2	0.50	0.68	0.69
		提案手法 3	0.67	0.85	0.85

表 3 かさ増し前後の精度平均

次に TF-IDF の数値の変化を見る．不均衡データ対しての精度を基準として見ると，最も効果が有ったのは提案手法 2 の wordnet を用いたかさまし手法であった．wordnet は人手で類義語を設定しているため語彙が大きく増えたことが要因であると考えられる．一方で文節入れ替えによる精度の向上は見られなかった．これは他の単語入れ替え手法と比較し語彙が増えていないことが大きな原因であると考えられる．

Doc2vec でベクトル化した数値の変化を見ると，最も精度を伸ばしたのは提案手法 3 の文節入替えによるかさ増し手法であった．この手法は同じ意味の文章を言い換えることで，表現の幅を広げる事が可能な手法であり，それが反映された結果であるといえる．一方で精度を向上できなかったのは wordnet を用いたかさ増し手法であった．これは，wordnet で類義語として定義された単語が Doc2vec 上では類義語として獲得されず，表現の幅が増えなかったことが原因だと考えられる．

全体として精度が大きく減少したかさ増し手法は無いため，かさ増し手法による悪影響は確認出来なかった．次に精度の詳細をみるため，全体についての適合率と再現率をみた．

5.2.1 全体の再現率・適合率

表 4 によれば適合率と再現率の数値に大きな差はないことがわかるが，TF-IDF にランダムフォレストを用いた場合わずかな変化の差が見られる．適合率が僅かに上昇しているが，再現率が僅かに減少している．これは正しく分類できたデータの割合が増えたが，データ数自体は減少したことを表す．この結果から，かさ増し手法によって，特定のデータの特徴が増加し多様性が失われた可能性があると考えられる．次にかさ増し手法による影響を，livedoor_HOMME の適合率・再現率からみる．

5.3 livedoor_HOMME の適合率・再現率

2 種類のベクトル化手法のどちらも不均衡データは適合率が高く，再現率が比較的低いことがわかる．次にダ

			適合率			再現率		
ベクトル	実験手法		NB	SVM	RF	NB	SVM	RF
TF-IDF	かさ増し前	不均衡データ時	0.87	0.92	0.90	0.85	0.89	0.93
		500 均一データ	0.89	0.92	0.91	0.89	0.92	0.91
	かさ増し後	提案手法 1	0.90	0.93	0.91	0.90	0.93	0.91
		提案手法 2	0.92	0.94	0.92	0.91	0.94	0.92
		提案手法 3	0.88	0.93	0.90	0.86	0.93	0.90
			cos	SVM	RF	cos	SVM	RF
Doc2vec	かさ増し前	不均衡データ時	0.45	0.70	0.69	0.51	0.69	0.69
		500 均一データ	0.48	0.71	0.69	0.53	0.70	0.69
	かさ増し後	提案手法 1	0.63	0.84	0.84	0.68	0.84	0.84
		提案手法 2	0.45	0.68	0.69	0.50	0.68	0.69
		提案手法 3	0.63	0.85	0.85	0.68	0.85	0.85

表 4 全体の再現率・適合率

ウンサンプリングを行った場合は，適合率を維持しながら再現率が上昇し，より多くのデータを正しく分類出来るようになった．これはデータ数を揃えることで，livedoor_HOMME の特徴が表れたためである．そしてかさ増し手法を用いることで，さらに再現率が上昇している．

TF-IDF では提案手法 2 が最も高い再現率を出している事がわかる．これは wordnet によって単語の語彙が増え，より多くのデータを正しく分類出来るようになったことを表している．一方で提案手法 3 がほとんど変化見られなかったのは，語彙が増えず TF-IDF ではほぼ同じデータとして表現されてしまったためである．これは同時にデータを複製するだけでは精度が向上しないことも表している．

Doc2vec で最も精度が向上したのは提案手法 3 であった．これは TFF-IDF では獲得できなかった文章としての表現の幅が獲得された結果であるといえる．一方で最も精度が向上しなかったのは提案手法 2 であった．これは wordnet で生成された類似度が Doc2vec 上で獲得出来なかったためであるといえる．以上のことからかさ増し手法の有効性が示唆された．ただし，ベクトル化とかさ増し手法の組み合わせによって有効性が変化していることには注意が必要である．

6 複数のかさ増し手法の併用

以上の実験より，提案したかさ増し手法を用いることでデータの少ないカテゴリに対して，精度の向上を行える可能性があると考えた．そこでよりデータ数の差がより大きい場合を扱う．かさ増し手法の特徴として，複数を組み合わせて用いることも可能であるため，提案手法 4 として文節入替え+word2vec によるかさ増し手法を，提案手法 5 として文節入替え+wordnet によるかさ増し手法を提案し実験する．併用した際の組み合わせと

dokujo-tsushin	770	peachy	770
it-life-hack	770	smax	770
livedoor-HOMME	70	sports-watch	770
movie-enter	770	topic-news	770
kaden-channel	770		

表 6 評価対象データ数を 70 に設定したデータセット

して，文節入替え手法と単語入替え手法を組み合わせたのは，文節入替え手法によって表現の幅を，単語入替え手法によって語彙を増やすことを目的としている．データの増加方法として，まず文節入替え手法を行いデータ

			適合率			再現率		
ベクトル	実験手法		NB	SVM	RF	NB	SVM	RF
TF-IDF	かさ増し前	不均衡データ時	1.00	0.87	0.97	0.39	0.83	0.53
		500 均一データ	0.93	0.89	0.92	0.72	0.86	0.76
	かさ増し後	提案手法 1	0.94	0.94	0.94	0.78	0.88	0.82
		提案手法 2	0.93	0.93	0.95	0.84	0.92	0.87
		提案手法 3	1.00	0.87	0.98	0.38	0.82	0.54
			cos	SVM	RF	cos	SVM	RF
Doc2vec	かさ増し前	不均衡データ時	0.42	0.70	0.71	0.49	0.38	0.38
		500 均一データ	0.46	0.75	0.68	0.55	0.48	0.48
	かさ増し後	提案手法 1	0.87	0.88	0.85	0.74	0.75	0.75
		提案手法 2	0.45	0.65	0.71	0.61	0.59	0.60
		提案手法 3	0.87	0.87	0.84	0.74	0.76	0.77

表 5 評価対象の再現率・適合率

			適合率			再現率		
ベクトル	実験手法		NB	SVM	RF	NB	SVM	RF
TF-IDF	かさ増し前	不均衡データ時	0.00	1.00	1.00	0.00	0.07	0.05
	かさ増し後	提案手法 4	0.99	0.99	0.98	0.79	0.91	0.88
		提案手法 5	0.96	0.98	0.99	0.99	0.96	0.98
			cos	SVM	RF	cos	SVM	RF
Doc2vec	かさ増し前	不均衡データ時	0.38	1.00	1.00	0.33	0.10	0.13
	かさ増し後	提案手法 4	0.65	0.95	0.96	1.00	0.99	0.98
		提案手法 5	0.68	0.97	0.97	1.00	0.99	1.00

表 8 評価対象データ数を 70 に設定した場合の再現率と適合率

数を 70 増やして 140 とし，その後各単語入替え手法によって 630 件のデータを増やした．

6.1 評価対象データ数を 70 に設定した実験

			全体精度		
ベクトル	実験手法		NB	SVM	RF
TF-IDF	かさ増し前	不均衡データ	0.85	0.89	0.89
	かさ増し後	提案手法 4	0.88	0.91	0.91
		提案手法 5	0.86	0.92	0.92
			cos	SVM	RF
Doc2vec	かさ増し前	不均衡データ	0.53	0.72	0.72
	かさ増し後	提案手法 4	0.56	0.74	0.74
		提案手法 5	0.56	0.74	0.74

表 7 評価対象データ数を 70 に設定した場合の全体精度

複数の手法を併用したかさ増し後の全体精度を評価対象が 500 個である不均衡データ時の結果と比較する．まず TF-IDF の場合を見ると，不均衡データ時は最大が 94%であるのに対し複数のかさ増し手法を併用した場合は 92%の精度になった．かさ増し前のデータ数を 500 から 70 に減らした場合でも近い精度を出す事ができた．

一方で，Doc2vec は不均衡データ時の最大精度が 85%で，極端な場合は最大 74%の精度であった．Doc2vec は文脈を加味してベクトル化を行うため，単語の入替えを中心にした今回の併用方法の場合，文脈の表現が増えなかったことが原因と考えられる．

6.2 データ削減実験時の livedoor_HOMME の精度変化

評価対象のデータ数を 500 に設定した場合と同様に，それぞれの適合率と再現率を見る．まずかさ増し前の値を見ると，TF-IDF，Doc2vec どちらも再現率が低いことがわかる．一方でかさ増し後を見ると，どちらの手法でも適合率，再現率が 90%を越えているが，全体の精度と評価対象の精度の差が大きくなる現象が見られる．これはカテゴリではなく単一のデータの特徴を膨らませた可能性がある．つまり，かさ増しされたデータが似通ってしまったため，分類するためのルールが固定されてしまったといえる．これはデータを複製して増やした場合とほぼ同じ効果であるため，かさ増し手法としては有効ではないことがわかる．よって，かさ増し手法を併用する際，データの多様性を維持する工夫が必要であると考えられる．

7 まとめと今後の課題

今回は不均衡データの精度改善を行うためのかさ増し手法を提案した．手法の有用性を見るため，複数のベクトル化，分類器で実験を行い，全体の精度変化と評価対象の適合率再現率を観察した．結果として，TF-IDF に対しては類似単語の入替えを行うかさ増し手法が有効であり，Doc2vec には文節入替えを行うかさ増し手法が有効であった．これはベクトル化手法の特性に則しており，ベクトル化手法の特徴を把握していれば，ある程度有用なかさ増し手法を推測可能であることを示している．しかし併用した際，かさ増しの多様性が失われてしまうことがわかった．これは特定の品詞や文節以外の特徴を元にかさ増しを行う必要があることを示している．

今後の課題として，オープンテストでかさ増し手法を実験することがまずあげられる．また．特徴を自動的に獲得しかさ増しを行うなど，新たなかさ増し手法の提案が課題としてあげられる．

参考文献

[1] He, Haibo, and Edwardo A. Garcia. ”Learning from imbalanced data.” IEEE Transactions on knowledge and data engineering 21.9 (2009): 1263-1284.

[2] 奥野峻弥, 浅井洋樹, and 山名早人. ”マイクロブログを対象とした 100,000 人レベルでの著者推定手法の提案.” 第 7 回データ工学と情報マネジメントに関するフォーラム (DEIM2015), D8-1 (2015).

[3] livedoor ニュースコーパス `http://www.rondhuit.com/download.html#ldcc`

[4] Mikolov, Tomas, et al. ”Efficient estimation of word representations in vector space.” arXiv preprint arXiv:1301.3781 (2013).

[5] MLA Le, Quoc, and Tomas Mikolov. ”Distributed representations of sentences and documents.” Proceedings of the 31st International Conference on Machine Learning (ICML-14). 2014.