企業の業界分類予測における共変量シフト問題の抑制

増田　太郎† 　石原祥太郎† 　吉田　勇太††

† 株式会社日本経済新聞社　〒 100–8066 東京都千代田区大手町 1-3-7
†† 株式会社ブレインパッド　〒 108–0071 東京都港区白金台 3-2-10 白金台ビル
E-mail: †{taro.masuda,shotaro.ishihara}@nex.nikkei.com, ††yuta.yoshida@brainpad.co.jp

あらまし　本研究では，企業の属性情報や概要説明文のテキスト情報を利用して業界分類を機械学習で予測する際に，学習時と予測時のデータの性質の違いによる影響を抑制する手法を提案する．一般的に学習データとして利用可能なラベル付きデータと，本番環境で予測すべきラベル無しデータとが特徴量の性質の違い（共変量シフト）を有していることは珍しくなく，学習データのドメインに過学習せずに，本来の目的であるドメインでの汎化性能を高めることが望ましい．本研究では，(i) 属性情報の人為的な欠損，(ii) 学習データと予測データの間の分布の違いを考慮した特徴選択，(iii) 入力テキストのトークン長の調整 という 3 つの手法を利用し，共変量シフトの影響を低減させた．実験を通じて，上場企業に対して 58.5 %，非上場企業に対して 56.1 %の正答率を達成し，上述の 3 つの対策を講じない場合と比較して約 2 ポイントほどの性能向上を確認した．

キーワード　企業情報，業種分類，共変量シフト，Adversarial Validation，自然言語処理

1 はじめに

企業において，他社の情報を正確に検索・把握する需要は高い．例えば経営企画部において，業務提携や M&A の相手企業を検討する材料にしたり，競合企業の動向を把握したりするために他社の正確な情報が必要になる．その中でも特に，他社企業が属する業界分類を正確に把握することができれば，上述の目的においてより効率的な企業検索が可能になるため便利である．

例として，著者らの一部が所属する日本経済新聞社では，上場企業について業界分類の情報を豊富に有しており，正解ラベルとして使用可能なデータが 4 万社ほど存在する．しかしながら，国内外には上場企業以外にも多くの企業が存在しており，国税庁調べ [1]では日本だけでも利益計上法人の数が年間 2 万社ほど増加する傾向にある．そのため，業界分類ラベルを人手で整備し続けることは時間的・金銭的コストの面から難しい．また，正解ラベル付きの企業と，新たにラベル付けを行いたい新興企業とが，必ずしも同様のデータ特性を有しているとは限らない．本研究では，既存のラベル付きデータを用いて業界分類を機械学習で自動的に予測することを目的とし，具体的には手持ちのラベル付き上場企業データと，予測対象となる非上場企業データとの間に存在する性質の違いに対応するためのアプローチについて提案する．

データセットシフトと呼ばれる学習用と評価用のデータセットの性質の違いに着目した研究は，近年盛んに取り組まれている [1] [2] [3]．本研究ではデータセットシフトの一種である，入力となる特徴量の性質が変化する共変量シフト [4] に焦点を当てる．共変量シフトに対応する主要な手法の一つに，2 つのデータセットに潜在的に共通する特徴量を抽出する方法がある [5]．Pan ら [6] は，学習用と評価用のデータセットを分類するモデルを作り，その正答率が閾値よりも低くなるまで重要度の高い特徴量を削除していく手法を提案した．

企業の業種ラベルを予測する問題は古くから研究されており，近年は機械学習に基づく方法が注目を集めている [7] [8]．一方で筆者らの知る限り，企業の業種ラベル予測においてデータセットシフトや共変量シフトの話題に焦点を当てた研究は多くない．

本研究では上場企業データと非上場企業データの間に存在する共変量シフトの影響について，両者に共通する特徴量を抽出するという方針で取り組む．最初に第 2 章で予備実験を通じて，本研究で扱う上場企業データと非上場企業データの間の特徴の違いを分析し，第 3 章で共変量シフトに対応するための提案手法を説明する．第 4 章で実験結果を報告し，第 5 章では結論を述べる．

2 予備実験

本章では，手持ちのラベル付き上場企業データと，予測対象となる非上場企業データとの特徴の違いを明らかにするための予備実験について説明する．利用できる情報としては，企業の属性情報と，概要・製品に関する説明文のテキスト情報の 2 種類がある．業種分類の正解付きデータとして，FactSet 企業情報データ [2]の中から事前に人手でラベル付けした上場企業データを用いる．ラベルは，日本経済新聞社の担当者が事業内容や会社の公式ホームページなどを確認しながら，分類の定義と照らし合わせて 1 社ずつ付与した．業種分類は本研究に限らず有用な情報であるため，本研究が始動する前から 4 万件ほど付与済みのものが用意されていた．このデータの中には複数ラベルが付与されている企業も存在したが，その場合は主要な業種

1：“令和元年度分「会社標本調査」調査結果について.” https://www.nta.go.jp/information/release/kokuzeicho/2020/kaisha_hyohon/index.htm

2：データセットの詳細については付録 1 を参照.

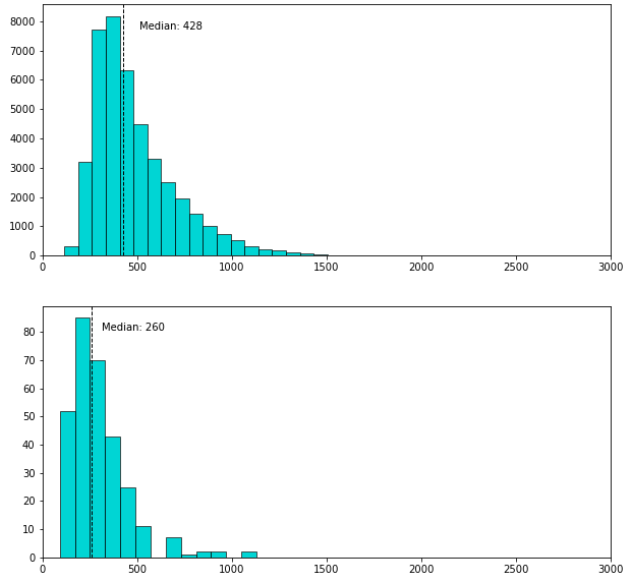

図 1　学習/検証データ (上) と評価データ (下) それぞれの単語カウントのヒストグラム．横軸は単語数，縦軸は度数を表す．破線は中央値を表し，その値は図中に示した．

コード 1 つのみを正解とするシングルラベル問題に帰着させた．それに加えて，本番での予測対象に相当する非上場企業についても 289 件ほどラベル付きデータを追加で用意し，上場企業との性質の差異を確認するために使う．

まず，属性情報の分布の違いについて調べた結果について説明する．全ての上場企業において入力特徴となる業界カテゴリ属性の値が埋まっているのに対し，評価データとなる非上場企業については約 55 %程度においてその値が欠損しているという事実が判明した．ゆえに，たとえ機械学習での予測において欠損データを補完せずに入力できたとしても，学習時点でその属性値に過度に依存したモデルを作ることは好ましくない．なお，この業界カテゴリは FactSet 社独自の企業情報データソースで定義された分類情報に基づいているため，本研究での予測対象となる日本経済新聞社で別途定義された業種分類に単純に紐づけることは難しい．

次に，テキスト情報の分布の違いについても調査結果を説明する．今回実験で使用する各企業のテキストの単語数を集計してヒストグラムを表示させたものが図 1 に示してある通り，上場企業データに対して非上場企業データの単語数は全体傾向として短いものとなっている．また，テキストデータの中身を探索的に確認することで得られた事実として，その企業の事業内容や製品など，分類に重要な情報はテキストの初めの方に現れやすかった．一方，テキストの後半には創業者の名前や本社の所在地など，予測にあまり重要でない説明が並ぶことも傾向として分かった．

3　提 案 手 法

本章では，今回の目的である業界分類コードの自動付与を達成するために，我々が提案する共変量シフトを緩和するための手法について説明する．

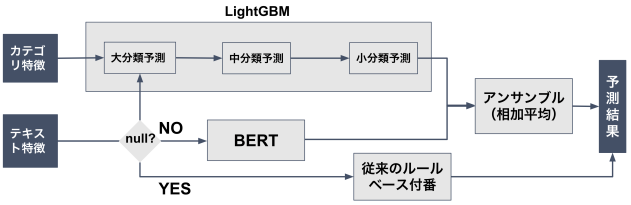

図 2　実験のフローチャート

3.1　属性情報の人為的な欠損

第 2 章で確認した通り，学習データと予測データとの間には属性情報の欠損率に大きな差があった．そこで本手法では，学習データと評価データ双方について，一定の比率で属性情報の値を人為的にランダムに欠損させることで，より実際の予測対象となるデータの特性に近づくようにした．

3.2　Adversarial Validation を利用した特徴選択

学習データと評価データとの共変量シフトの影響を低減させる方法として，Adversarial Validation [6] とよばれる手法が知られている．具体的にはまず，与えられた企業データが学習データ・評価データのいずれであるかという 2 クラス分類モデルを学習させる．その後，学習データと評価データの特徴分布の乖離を軽減するために，重要度の値が高い特徴量を除外する．

3.3　テキストのトークン長の調整

BERT などのテキスト系列を入力する自然言語処理モデルにおいて，Devlin ら [9] は論文中で入力トークンの長さを 512 程度とし，短いトークン列に対しては 0 で埋めて同じ長さに揃えている．しかし，本研究ではトークン長を 128 と短くすることを提案する．その理由は 2 つある．

1 つは，本番での予測対象と近い性質を持つ評価データは，全体的な傾向として文のトークン長が 512 よりも短く，長い文の入力を前提とするモデリングは適さないと考えたためである．

もう 1 つの理由は，第 2 章で確認した通りテキストの後段には予測にあまり重要でない説明が並んでいるため，特に学習データにしばしば含まれる 128 語を超える説明文を冗長とみなして切り捨てる方が良いと考えたためである．

4　実　　験

本章では，提案した手法を使って日本経済新聞社が抱える企業データについて実際に業種分類モデルを適用した場合の実験について報告する．ここでの目的は，後述するように「小分類」と呼ばれる最も細かい粒度である 524 クラスの分類問題について調査する．本実験のフローチャートを図 2 に示す．

4.1　機械学習モデル

本研究では機械学習のモデルとして，LightGBM [10] と BERT [9] の 2 つを使用して予測する．それぞれの使い方について以下で詳しく述べる．

4.1.1 LightGBM

使用できるデータの中には，国名や分野，FactSet 社 [3]が定義する業界カテゴリ属性などのカテゴリ変数および，英語のテキストデータ 2 種類（企業概要および製品情報）が含まれている．そこで，まず表形式のデータに対して高い性能を出しやすい勾配ブースティング木モデルである LightGBM を採用した．カテゴリ変数については，それぞれ単純なラベルエンコーディングにより数値化した特徴量を用いた．

LightGBM での予測は，図 2 の通り 3 つのステージに分かれている．3 つのステージではそれぞれ「大分類」「中分類」「小分類」と呼ばれる業種を予測していくが，日本経済新聞社が定義するこれらの業種分類についての詳細は付録 1 を参照されたい．まず，カテゴリ特徴の全てと，企業概要テキストの TF-IDF [11] と，製品情報テキストの TF-IDF を含む計 696 次元の特徴量を用いて，最終的に予測したい「小分類コード」ではなく，より粗い粒度の分類である「大分類コード（15 種類）」を予測する．その理由は，小分類コードの予測においても，より粗い粒度の（推定された）分類情報が大きな手がかりになると考えたためである．次に，前述の 696 次元の特徴量に，大分類コードの予測結果（15 次元）を加えた計 711 次元の特徴量を用いて「中分類コード（68 種類）」を予測する LightGBM モデルを新たに作る．

最後に，大・中分類コードの予測結果の一部や，企業概要テキストの TF-IDF，製品情報テキストの TF-IDF といった 20 次元の特徴量と，追加のテキスト特徴量として「SWEM-average [12]」,「fastText [13]」,「BM25 [14]」,「Universal Sentence Encoder [15]」をそれぞれ特異値分解（SVD; Singular Value Decomposition）により 50 次元ずつに圧縮したものを採用して，合計 220 次元の特徴量を用いて小分類コード（524 次元）を予測する LightGBM モデルを新たに作る．SWEM-average において必要となる単語埋め込みベクトルとしては，Google ニュースデータセットで事前学習された word2vec（`GoogleNews-vectors-negative300.bin`）を用いた．

4.1.2 BERT

前節で用いた LightGBM において入力する特徴量は，いずれも単語ごとの特徴表現となっており，テキストの系列的な特徴や文脈を考慮した表現に欠けている．また，一般に勾配ブースティング決定木モデルとニューラルネットワークとのアンサンブルが高い効果を発揮しやすいことが経験的に知られている [16]．そこでもう 1 つのモデルとして，BERT モデルを今回のテキストデータにファインチューニングさせたものを採用した．入力トークンは,「企業概要テキスト」「製品情報テキスト」の 2 つを単にスペース区切りで連結し，全体をトークン ID に変換することで用意した．事前学習済みモデルとしては，Hugging Face 社が提供する Transformers [17] の `bert-base-uncased` を採用した．BERT においては LightGBM と異なり，属性情報を入力に使わないため，大分類・中分類の予測は挟まずに初めから小分類を直接予測させる．

3：`https://www.factset.com/`

4.1.3 アンサンブル

前節までに説明した LightGBM と BERT それぞれのモデルが出力した予測値を利用して，最終的な予測としてアンサンブルした予測値を採用する．具体的には，それぞれのモデル単体で検証した性能がほぼ同等であったため，今回は単純な相加平均を採用した．後の実験結果から明らかになる通り，アンサンブルにより正答率自体の改善ができるだけでなく，共変量シフトの問題が軽減される効果もある．

4.2 実験設定

ここでは実験の評価方法について述べる．まず，ラベル付き上場企業データ 42,752 件に対して，学習データと評価データに分割するため，全体で 1 企業しか存在しない小分類コードについては学習・予測の対象から外した．その上で，データ全体を 7:3 の比率で，層化抽出により学習データと評価データに分割した．さらに，その学習データについて，5-fold 層化クロスバリデーションを用いて 5 つのモデルを作成し，初めに分割した評価データを利用して予測の評価を行った．単体モデルの最終的な予測は，5-fold により生成された 5 つのモデルそれぞれの予測値に対して相加平均を取ることにより算出した．評価指標としては，件数ベースでの正答率を上げたいという要件を加味し，単純な正答率（Accuracy）を採用した．また，日本経済新聞社内の有識者らの協力により，本番相当の非上場企業についてもラベル付きのデータが 289 件ほど新たに得られたため，これらについても評価データ（非上場）として正答率を算出した．理想的には，非上場企業についても上場企業と同等程度のラベル付きデータを用意することが望ましいが，業種分類の知見を有する者によるラベル付けのコストが高く，今回は簡易的なデータサイズでの評価にとどまっている．使用したモデルのパラメタは付録 2 に示した．

次に，比較手法について説明する．同様のデータセットに対する関連研究は存在しないため，参考値として，従来日本経済新聞社のサービスにて導入されていたルールベースに基づく業界コード付与の正答率と比較することとした．ルールベースに基づく手法の詳細については付録 3 に示した．

4.3 実験結果

まず，提案手法で述べた Adversarial Validation の結果について説明する．BM25 特徴量の 1 次元目が他の特徴量よりも重要度が遥かに高かったため，共変量シフトへの悪影響が特に強いと判断して手動で除外した．したがって，LightGBM において最終的に使用する特徴量の次元数は 219 とした．

次に，正答率を算出した結果を表 1 に示す．この結果から，次に述べるいくつかの事実が分かった．

現状のサービスに導入されているルールベースの手法を上回る性能を LightGBM で達成できたが，検証データと評価データの正答率に 20 ポイント以上の差があり，共変量シフトの影響を強く受けていた．それに対して，Light GBM 単体で提案手法 (i),(ii) を導入した場合と比較すると，いずれも正答率の乖離が改善したことがわかった．BERT においても，提案手法

表 1　各モデルの正答率の結果.「検証」は検証データ（上場企業）の正答率,「評価」は評価データ（非上場企業）の正答率を表す.

手法名	検証 (%)	評価 (%)
ルールベース	42.08	36.33
LightGBM	58.61	38.06
LightGBM（提案 (i)）	54.57	46.36
LightGBM（提案 (ii)）	55.79	46.71
LightGBM（提案 (i)+(ii)）	54.94	47.06
BERT	50.01	48.24
BERT（提案 (iii)）	50.28	50.59
アンサンブル	59.55	53.98
アンサンブル（提案 (i)）	59.05	55.36
アンサンブル（提案 (ii)）	59.05	55.36
アンサンブル（提案 (iii)）	58.53	55.02
アンサンブル（提案 (i)+(ii)）	59.11	54.67
アンサンブル（提案 (i)+(iii)）	58.27	55.02
アンサンブル（提案 (ii)+(iii)）	58.63	55.02
アンサンブル（提案 (i)+(ii)+(iii)）	58.53	56.06

(iii) を導入することで正答率の乖離が改善することを確認し，その効果を確認できた．アンサンブルした場合においても，提案手法なしの場合と比較して，いずれも正答率の乖離を改善できていることがわかる．

提案手法 (i)–(iii) を全て盛り込んだアンサンブルモデルは，ルールベースに対して約 16〜20 ポイント程度も優れた正答率を達成できた．また，LightGBM あるいは BERT 単体が示す性能と比較して，アンサンブルの効果を確認することができた．さらに，LightGBM 単体で上場企業と非上場企業の間に存在していた 7 ポイント以上もの乖離について，アンサンブルにより 2.5 ポイント程度と大きく低減できたことも分かった．

4.4 考　　察

4.4.1 各手法の貢献度に関する考察

本節では，提案手法がそれぞれ共変量シフトの改善に寄与した貢献度の大きさについて考察する．

最も貢献度が大きかった手法の 1 つは「(i) 属性情報の人為的な欠損」であり，上場企業と非上場企業の正答率の差を 11 ポイント改善することができた．今回の場合は評価データの属性の欠損比率が事前に分かっていたため，この手法が特に有効であったと考えられる．より一般的な問題設定として欠損比率が未知であるデータを予測する場合は，様々に欠損比率を変えて学習させたモデルでのアンサンブルが有効であると考えられる．興味深い事実として，FactSet 社基準で分類された 130 種類ほどの業種ラベル「Industry」列の属性情報を全て利用する場合に比べて，ランダムに人為的に欠損させた場合の方が，検証データに対しては正答率が 3 ポイントほど低下したものの，評価データにおける正答率が 8 ポイント程度向上し，両者の乖離を大きく低減できた．このことから，Industry の値と業界分類との対応関係について，上場企業と非上場企業との間で乖離があったことが明らかになった．同時に，その乖離の問題について，上述の人為的な欠損により効果的に対処できたことも分かった．

「(ii)Adversarial Validation を利用した特徴選択」についても，単体での正答率の乖離の改善幅は 11 ポイントと同程度に大きかった．特徴量の数を削ることと正答率を高く保つこととはトレードオフになりやすいため今回は 1 次元のみの削除を採用したが，今回のデータに限らず，一般的に複数の特徴量の除外を適用することで更なる改善が見込める可能性もある．

「(iii) 入力テキストのトークン長の調整」については，BERT において元々上場企業と非上場企業との正答率の乖離が激しくなかったため，2 ポイント程度の乖離の改善にとどまった．同等程度の乖離の改善法として，LightGBM と BERT のアンサンブルも挙げられる．特に LightGBM にてより顕著であった 7 ポイントもの正答率の乖離について，アンサンブルにより 5 ポイント程度まで改善することができている．BERT 単体のモデルが頑健であり，上場企業と非上場企業との間で乖離が大きくなかったことから，アンサンブルにより共変量シフトの影響を BERT が吸収しながら正答率の底上げも達成できたと考えられる．

以上のことから，提案手法 (i) から (iii) の全てについて有効性が確認でき，全てを導入した場合に正答率の乖離を 2.5 ポイント程度まで改善しつつ正答率も高く保つことができた．

4.4.2 定性的なエラー分析

3.2 節で述べた実験の結果を分類コードごとに集約すると，学習サンプルサイズが小さいにもかかわらず正答率が高かった，あるいは逆に学習サンプルサイズが大きいにもかかわらず正答率が低かったものが見つかった．そこで本節では，これらの違いがどういった性質から生じたものであるかについて定性的に調査した．

上述の調査のために,「学習データサイズが 50 サンプル以上あるにもかかわらず，正答率が最も低かった小分類コード 5 つ」および,「学習データサイズが 50 サンプル未満にもかかわらず，正答率が最も高かった小分類コード 5 つ」それぞれについて，テキストデータを全て人手でチェック・比較し，傾向を観察した．結果として，正答率が低くなる原因として以下の 3 つが見られた．

1 つ目は，その企業が展開する事業が複数・多岐にわたっており，1 つに絞りにくいという点である．逆に言えば，単一事業を展開する会社については正解しやすい傾向にあると言える．

2 つ目は，正答率の低かった分類コードにおいて，人間でも判断が難しいような誤分類が存在するという点である．例えば,「パッケージソフト（一般向け）」という分類と,「パッケージソフト（その他業務向け）」という分類の線引きは，その分類の経緯を良く知らないと人間にとっても難しい．特に，ソフトウェアを提供する IT 関連の企業においては分類コードが多岐にわたっており，それゆえ誤分類が起きやすかったと言える．逆に正答率の高かった分類コードにおいては，判断に迷ってしまうような他の紛らわしい分類コードが傾向として少なかった．

3 つ目は，正解ラベル自体の誤り（Noisy labels）の存在である．例えば,「biologic drugs（バイオ創薬）」事業に関する説明テキストが書かれているにもかかわらず,「ファストフード（ハンバーガー）」が正解であるなど，明らかに誤りと見られるデー

タが一部存在していた．正解ラベル自体が人手作業によるものであるため誤り自体は避けられないが，追加のアノテーションコストを確保して誤りを訂正することにより正答率が改善する可能性がある．

5 結　　論

本研究では，FactSet 社が提供する企業データに対して，業界分類コードを付与するための機械学習アルゴリズムについて提案した．上場企業データと非上場企業データの間に存在した共変量シフトの影響について，(i) 属性情報の人為的な欠損，(ii) Adversarial Validation を利用した特徴選択，(iii) 入力テキストのトークン長の調整 という 3 つの手法により低減させることに成功した．実験を通じて，上場企業の検証データに対して 58.5 %，非上場企業のテストデータに対して 56.1 %の正答率が達成され，上述の 3 つの対策を講じない場合と比較して約 2 ポイントほど正答率が向上した．日本経済新聞社で従来採用していたルールベースの手法の正答率を大きく上回ったため，著者らは提案手法を実サービスへ導入している．

謝　　辞

本研究において，ご提供いただいた企業情報データベースの利用，および研究内容の発表に関して，FactSet Research Systems Inc. からご許諾をいただきました．また，株式会社ブレインパッドの藤田亮様には，本研究のベースラインモデルの構築でご助力をいただきました．最後に，株式会社日本経済新聞社 情報サービスユニット バリューサーチチームの有識者の皆様には，評価用の企業データを用意するために業種のラベル付けにご協力いただきました．以上全ての皆様に，ここに深い感謝を申し上げます．

文　　献

[1] Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence. “Dataset Shift in Machine Learning.” MIT Press (2008).

[2] Jose G. Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V. Chawla, and Francisco Herrera. “A unifying view on dataset shift in classification.” Pattern Recognition, vol. 45, no. 1 (2012): 521–530.

[3] Jie Lu, Anjin Liu, Fan Dong, Feng Gu, João Gama, and Guangquan Zhang. “Learning under Concept Drift: A Review.” IEEE Transactions on Knowledge and Data Engineering, vol. 31, no. 12 (2018): 2346–2363.

[4] Hidetoshi Shimodaira. “Improving Predictive Inference under Covariate Shift by Weighting the Log-Likelihood Function.” Journal of Statistical Planning and Inference, vol. 90, no. 2 (2000): 227–244.

[5] Jingjing Li, Jidong Zhao, and Ke Lu. “Joint Feature Selection and Structure Preservation for Domain Adaptation.” Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16) (2016): 1697–1703.

[6] Jing Pan, Vincent Pham, Mohan Dorairaj, Huigang Chen, and Jeong-Yoon Lee. “Adversarial Validation Approach to Concept Drift Problem in User Targeting Automation Systems at Uber.” ACM AdKDD 2020 (2020).

[7] Sam Wood, Rohit Muthyala, Yi Jin, Yixing Qin, Nilaj Rukadikar, Hua Gao, and Amit Rai. “Automated Industry Classification with Deep Learning.” 2018 IEEE 12th International Conference on Semantic Computing (ICSC 2018) (2018): 64–70.

[8] Andrey Tagarev, Nikola Tulechki, and Svetla Boytcheva. “Comparison of Machine Learning Approaches for Industry Classification Based on Textual Descriptions of Companies.” Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019) (2019): 1169–1175.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019), vol. 1 (Long and Short Papers) (2019): 4171–4186.

[10] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. “LightGBM: A Highly Efficient Gradient Boosting Decision Tree.” 31st Conference on Neural Information Processing Systems (NIPS 2017), vol. 30 (2017).

[11] Juan Ramos. “Using TF-IDF to Determine Word Relevance in Document Queries.” Proceedings of the First Instructional Conference on Machine Learning, vol. 242, no. 1 (2003): 29–48.

[12] Dinghan Shen, Guoyin Wang, Wenlin Wang, Martin Renqiang Min, Qinliang Su, Yizhe Zhang, Chunyuan Li, Ricardo Henao, and Lawrence Carin. “Baseline Needs More Love: On Simple Word-Embedding-Based Models and Associated Pooling Mechanisms.” Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018), vol. 1 (Long Papers) (2018): 440–450.

[13] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. “Enriching Word Vectors with Subword Information.” Transactions of the Association for Computational Linguistics, vol. 5 (2017): 135–146.

[14] Stephen Robertson, and Hugo Zaragoza. “The Probabilistic Relevance Framework: BM25 and Beyond.” Foundations and Trends in Information Retrieval, vol. 3, no. 4 (2009): 333–389.

[15] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Céspedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. “Universal Sentence Encoder for English.” Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018), System Demonstrations (2018): 169–174.

[16] 門脇大輔, 阪田隆司, 保坂桂佑, 平松雄司. “Kaggle で勝つデータ分析の技術.” 技術評論社 (2019).

[17] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. “Transformers: State-of-the-Art Natural Language Processing.” Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020), System Demonstrations (2020): 38–45.

付　　録

1 日本経済新聞社における業種分類について

本章では，日本経済新聞社における業種分類の定義について説明する．日本経済新聞社では，法人向け企業情報データベー

表 A·1　LightGBM のハイパーパラメタ

パラメタ名	大分類・中分類予測における値	小分類予測における値
num_leaves	64	16
max_depth	6	4
feature_fraction	0.8	0.8
bagging_freq	1	1
bagging_fraction	0.7	0.7
min_data_in_leaf	10	10
learning_rate	0.01	0.05
lambda_l1	0.4	0.4
lambda_l2	0.4	0.4
max_bin	100	255

表 A·2　BERT のハイパーパラメタ

パラメタ名	値
lr	1e-5
batch_size	32
max_length	128
max_epochs	100

スである「日経バリューサーチ」[4]というプロダクトを抱えており，業種分類の情報もこのプロダクトから提供している．

日本経済新聞社では，粒度の粗い順に「大分類」「中分類」「小分類」という業種分類をそれぞれ 15 種類，68 種類，550 種類ほど独自に定義している．このうち，本研究では最も細かい分類である「小分類」をなるべく高い正答率で予測することを目指した．その理由は，日経バリューサーチにおける業種分類の用途としては，大分類・中分類だけが分かっていても不十分であるケースがほとんどであったためである．なお，実験の章では小分類を 524 クラスと定義しており，本来の定義である 550 クラスに満たない．その理由は，ラベル付きデータの分割が不可能であった少量のクラスを除外したためである．

日経バリューサーチで利用可能な企業データとして，外部データベースサービスである「FactSet」から取得した企業データが数多くある．FactSet から取得した企業データにおいて，「Industry」という FactSet 社基準で分類された 130 種類ほどの業種ラベルが付与されていることがある．しかし，FactSet で定義されている業種と，日本経済新聞社で定義されている小分類について，一方から他方への単純な対応ルールを高い正答率で作ることは専門家にとっても困難であった．さらには，FactSet で定義されている業種ラベルは必ずしも付与されているとは限らず，本文記載の通り非上場企業のおよそ 55 %について欠損している．そこで本研究では，業種の単純な対応ルールを人手で定義するのではなく，データに内在するパターンを自動で抽出できる機械学習モデルに小分類を自動で予測させた．

2　モデルのハイパーパラメタ

実験において設定した各モデルのハイパーパラメタは表 A·1, A·2 の通りである．

3　ルールベースとの併用

予測対象の企業にテキストデータが一切入っていない場合は，機械学習での予測性能が著しく低下する．そのため，図 2 で示したように，その場合は機械学習モデルの使用を避け，ルールベースに基づく手法に切り替える．ルールベースに基づく手法とは，各企業を「Industry」の属性単位で集約したうえで，各 Industry の属性の中で最頻の小分類コードに割り当てる処理のことを指す．例えば，Industry の属性値が「Electronic Components（電子部品）」となっている企業に対しては，その中で最も頻出である「電子回路基盤・配線板」という小分類であると一律に予測する．しかし，実際には Industry の属性値が「Electronic Components」であっても,「電源・電源装置」「スイッチ・コネクター・ワイヤーハーネス」など異なる小分類が正解であるケースも少なくない．以上の手法について，本文中で「ルールベースに基づく手法」として正答率の比較のために参考値として示した．

以上のルールベースの手法はあくまで実運用のための説明であり，本文に説明した実験においては，全ての企業においてテキストデータが存在していた．したがって，ルールベースと機械学習を併用した場合の実験結果については本稿の対象外となる．

4：https://valuesearch.nikkei.com/