2ZC-01

機械学習による一般紙金融情報欄の元画像を用いた市場安定度分類

脇田拓弥 †　　奥田隆史 †

愛知県立大学情報科学部情報科学科 †

1　はじめに

人々はデータ全体を眺めて総合的に社会情勢を把握しているのではないだろうか．我が国の 2020 年における新聞（一般・スポーツ紙）の発行部数は約 3509 万部である．そのうち，一般紙の発行部数は約 3245 万部である [1]．総人口約 3 割が一般紙を通して，ローカル，グローバルの出来事から社会情勢の変化を認識していると推察できる．なお新聞科学研究所の調査 [2] では，購読者は景気や物価の動向に対する情報感度が高い傾向があることが報告されている．

本研究では，全体を眺めて社会情勢変化を認識可能なものとして，**一般紙金融情報欄**（以後，**日間株式欄**）に着目する．日間株式欄は通常，前日が休日ではない火曜日から土曜日までの一般紙における経済面に掲載されている．

本研究の目的は，図 1 のように機械学習の 1 つである畳み込みニューラルネットワーク（Convolutional Neural Network: 以後，CNN）を用いて，日間株式欄の画像から株式市場の安定度分類（以後，市場安定度分類）を行う．なお，日間株式欄画像のサイズ，そのアスペクト比（以後，縦横比）を考慮して，CNN へ入力する．

本研究において，CNN を用いる理由は 2 つある．1 つは人が日間株式欄から直感的に市場安定度を把握可能なことである．そのため人の視覚に基づいて考えられた CNN は，市場安定度を分類可能と推測できるからである．もう 1 つは，日間株式欄画像を 2・3 次元形状で CNN に入力可能だからである [3]．

以下，第 2 節では市場安定度の分類方法ついて，第 3

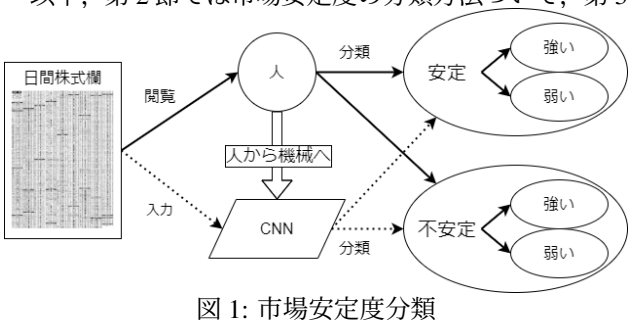

図 1: 市場安定度分類

2　市場安定度の分類方法

本節では，市場安定度の分類方法ついて説明する．本研究では，日間株式欄画像を分類分析する期間を**分類分析期間**と表現する．分類分析期間の日間株式欄画像へ意味づけを行うための基準データ取得期間を，**取得期間**と呼ぶ．なお取得期間には分類分析区間を含める．

2 値分類は，取得期間の数値的株価情報（以後，株価情報）から各日毎の昨年・年初来高値（安値） の更新株式数（以後，更新数）を抽出する．抽出した全ての昨年・年初来高値（安値）更新数における相対平均を分類基準値とする．そして更新数が分類基準値未満を安定，分類基準値以上を不安定とする．

4 値分類は，2 値分類において安定・不安定に属する更新数の相対平均を安定分類基準値・不安定分類基準値とする．そして，更新数が安定分類基準値未満を強安定，安定分類基準値以上かつ分類基準値未満を弱安定，分類基準値以上かつ不安定分類値基準を弱不安定，不安定基準値以上を強不安定とする．

取得期間の株価情報から決定した各分類基準値を利用して，分類分析期間の日間株式欄画像からデータセット（学習・検証データ）を作成する．なお，取得・分類分析期間は第 4 節（数値例）で示す．

3　CNN 構成と評価方法

本研究では，基本的なアーキテクチャの CNN を Keras を用いて実装する [4]．日間株式欄画像のサイズ（縦，横，チャンネル数）は，(256，256，1)，(512，512，1)，(1024，1024，1）である．この順に **A**，**B**，**C** とする．

日間株式欄画像は，始めに畳み込み層を通過する．畳み込み層（Convolution Layer; Conv）は，フィルターを入力画像に適用することにより，特徴量を抽出する．さらに特徴量から特徴マップ（2 次元形状）を作成する．その後，活性化関数を通過する．活性化関数には，正規化線形関数（以後，Relu 関数）を用いる．Relu 関数は，入力値が正であるなら同値を通過させ，負であるなら 0 にして通過させるものである．勾配消失問題の解消，計算速度向上の効果がある．Relu 関数通過後は，最大プーリング層を通過する．最大プーリング層は，フィルターサイズにおける最大値を取り出す．計算負荷，メモリ使用量，パラメータ数の削減の効果がある．なお，畳み込み層，活性化関数，最大プーリング層を一層とカウントする．上層の一連の流れの後には，特徴マップを 1 次元形状に変換し，全結合層（Fully Connected Layer; FC）を通過する．全結合層では，特徴量に基づいて，分類を行う．

各入力サイズに対応する CNN の構成を表 1，そのパラメータ数を表 2 に示す．畳み込み層を Conv（フィルターサイズ）-（チャンネル数），全結合層を FC-（ニューロン数）と表記する．A は，第 1 層から第 7 層まで畳み込み層，活性化関数，最大プーリング層の繰返しとする．第 8 層から第 9 層まで全結合層，Relu 関数の繰返しとする．B は，第 1 層から第 8 層まで畳み込み層，活性化関数，最大プーリング層の繰返しとする．第 9 層から第 10 層まで全結合層，Relu 関数の繰返しとする．C は，第 1 層から第 9 層まで畳み込み層，活性化関数，最大プーリング層の繰返しとする．第 10 層から第 11 層まで全結合層，Relu 関数の繰返しとする．なお，A（第 10 層），B（第 11 層），C（第 12 層）の出力層は，全結合層，Sigmoid 関数（2 値分類）または，Softmax 関数（4 値分類）とする．

最適化アルゴリズムは，適応モーメント推定（adaptive moment estimation; Adam）とする．なお，過学習は，ドロップアウト層の設置，学習率変更，荷重減衰設定などで，抑制する [3][5]．

評価方法は，CNN が正しく分類しているかを確認す

Stability Classification of Stock Market with Original Images of General Newspaper's Financial Information Columns by Using Machine Learning

†Takuya WAKITA，Takashi OKUDA

†Department of Information Science and Technology，Faculty of Information Science and Technology，Aichi Prefectural University

る学習時正解率（Training-Accuracy）と検証時正解率（Validation-Accuracy）を用いる．

表 1: CNN の構成

Layer	CNN (Input Size)		
	A (256, 256, 1)	B (512, 512, 1)	C (1024, 1024, 1)
1	Conv3-2	Conv3-2	Conv3-2
2	Conv3-4	Conv3-4	Conv3-4
3	Conv3-8	Conv3-8	Conv3-8
4	Conv3-16	Conv3-16	Conv3-16
5	Conv3-32	Conv3-32	Conv3-32
6	Conv3-64	Conv3-64	Conv3-64
7	Conv3-128	Conv3-128	Conv3-128
8	FC-128	Conv3-256	Conv3-256
9	FC-128	FC-128	Conv3-512
10	FC-1（FC-4）	FC-128	FC-128
11	None	FC-1（FC-4）	FC-128
12	None	None	FC-1（FC-4）

表 2: CNN のパラメータ数

CNN		A	B	C
Number of Parameter	TwoClassification	180,857	541,561	1,852,793
	Four Classification	181,244	541,948	1,853,180

4 数値例

本研究では，東京証券取引所第 1 部（以後 東証 1 部）における**取得期間** 2016 年 1 月 1 日から 2020 年 12 月 2 日までの株価情報，**分類分析期間** 2020 年 2 月 29 日から同年 12 月 3 日までの朝日新聞朝刊，東証 1 部の日間株式欄画像，合計 161 枚を用いる [6]．第 2 節（市場安定度の分類方法）に基づく，各分類基準値と更新数のヒストグラムを図 2 に示す．分類基準値（黄）は 129.9 個，安全分類基準値（青）は 72.8 個，不安全分類基準値（赤）は 245.8 個となる．さらに各分類基準値を利用して，日間株式欄画像に市場安定度を意味付け，データセットを作成する．

図 2 からデータセットの偏りが推察できるため，日間株式欄画像の回転，反転，ぼかしなどの一般的なデータ拡張を行う [7]．

本研究では，データセットを k 個に分割し，そのうち 1 つを検証データとする．残りの $k-1$ 個を学習データとして正解率評価を行う層化 k 分割交差検証を用いる．層化 5 分割交差検証とし，学習回数（Number of Epoch）は 100 回とする．CNN の各分類結果（学習曲線）を図 3, 4 に示す．各 CNN の構成による正解率を A_y，A_n，B_n，B_y，C_n，C_y と表す．また添字は縦横比考慮あり（y），なし（n）とする． なお，破線は学習時正解率，実線は検証時正解率を表す．

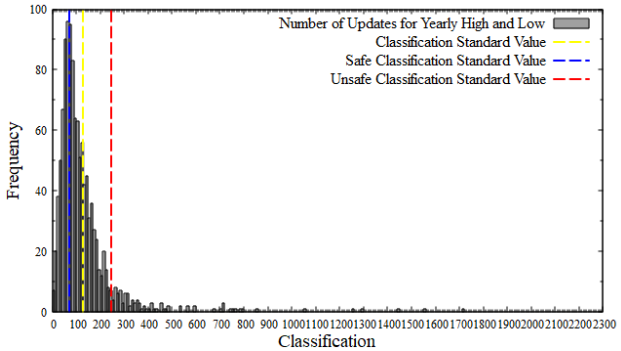

図 2: 昨年・年初来高値（安値）更新数のヒストグラム

2 値分類

2 値分類の正解率における学習曲線を図 3 に示す．日間株式欄画像サイズ，縦横比に関わらず，A，B，C の分類結果は，学習・検証時ともに 95%以上の正解率となった．また学習時正解率と検証時正解率がほぼ同じように推移しているため，未知データに対応できる汎化性能も高いといえる．

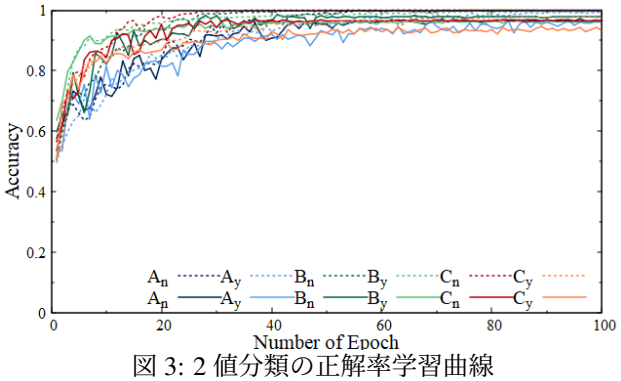

図 3: 2 値分類の正解率学習曲線

4 値分類

4 値分類の正解率における学習曲線を図 4 に示す．日間株式欄画像サイズが C，B，A の順に学習・検証時ともに正解率が高かった．これは日間株式欄画像サイズが大きく，CNN の深層化により，細かな特徴量を得られたからだと推察できる．また縦横比に関わらず，日間株式欄画像サイズが同じ場合は，ほぼ同じような結果となった．検証時正解率は 60%以上であり，それなりの結果であった．

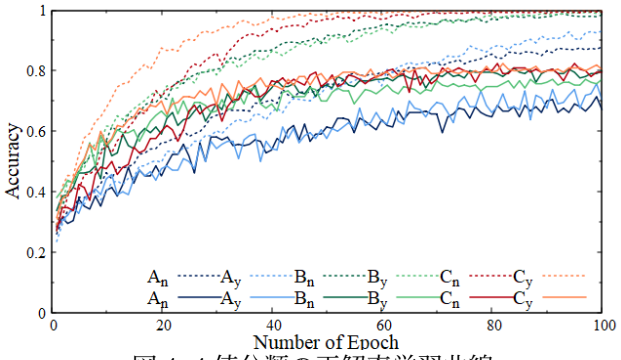

図 4: 4 値分類の正解率学習曲線

5 おわりに

本研究では，CNN を用いて，日間株式欄画像から市場安定度分類を行った．検証時正解率が 2 値分類では 95%以上，4 値分類では 65%以上であった．さらに 4 値分類では，入力サイズと CNN の深層化により精度が変化した．以上から，紙面の日間株式欄と同程度の日間株式欄画像サイズを与えることで，細かい特徴量を抽出可能であると言える．また人が日間株式欄から直感的に市場安定度を分類することを，CNN を用いることで機械が客観的に分類可能であることがわかった．今後の課題は，CNN の画像認識箇所の特定，4 値分類の検証時正解率のさらなる向上などである．

参考文献

[1] 日本新聞協会，“新聞の発行部数と世帯数の推移｜調査データ｜日本新聞協会”，https://www.pressnet.or.jp/data/circulation/circulation01.php，最終閲覧日: 2021 年 1 月 5 日．

[2] 新聞科学研究所，“将来への経済的不安、新聞で少なく”，https://np-labo.com/archives/episode/201904kiji-03，最終閲覧日: 2020 年 12 月 25 日．

[3] オーレリアン・ジェロン，『scikit-learn と TensorFlow による実践機械学習』，オライリー・ジャパン，2018．

[4] Keras，“Keras: the Python deep learning API”，https://keras.io/，最終閲覧日: 2020 年 12 月 25 日．

[5] 斎藤康毅，『ゼロから作る Deep Learning ― Python で学ぶディープラーニングの理論と実装』，オライリー・ジャパン，2016．

[6] 東京証券取引所，“東証上場会社情報サービス”，https://www2.tse.or.jp/tseHpFront/JJK010010Action.do?Show=Show，最終閲覧日: 2020 年 12 月 25 日．

[7] 足立悠，『機械学習のための「前処理」入門』，リックテレコム，2019．