VAE を用いた視覚的株価予測 AI

チーム慶工　　綱島秀樹　中間康文　枇々木裕太

1. 序論

近年, 本国では少子高齢化が進行し, 資産運用の重要性が高まってきている．機関投資家が顧客の資産を運用する際，従来では基本的な定量的手法としてマルチファクターモデルが広く使用されてきた．マルチファクターモデルとは，ポートフォリオの収益率を，株価指標や経済指標などの複数のファクターを用いて説明するモデルであり, 特に, マーケットポートフォリオのリターン・小型・割安ファクターを考慮した Fama-French の 3 ファクターモデルが有名である．しかし，昨今の機械学習及び AI ブームが金融の分野にも浸透しつつあり，機械学習及び AI によって高い運用パフォーマンスが期待できるといった論文も投稿されている．特に，近年の深層学習(ディープラーニング)の発展は著しく，本研究では，ディープラーニングを株価予測 AI の作成に応用する．ディープラーニングには様々な分野があるが，その中でも, 特に画像処理に注目し, 本研究では画像の生成モデルの 1 つである Variational Autoencoder(VAE)を用いた株価予測 AI の作成を目指す．VAE によって，本コンペの投資期間(2019 年 2 月)における各企業の株価の予想チャートを画像生成し，株価の騰落を予想する．チャートのパターンを参考に取引をしている投資家は一定数存在するため，VAE を用いてそのパターンを上手く学習できれば，投資期間において上昇する企業株を精度良く判断できることが期待される．また, それに加えて Twitter の感情分析も組み合わせてポートフォリオを組む.

1.1 全体像

本研究では，東証 1 部・2 部，マザーズ，JASDAQ に上場している日本企業 3622 社を対象に 5～10 社のポートフォリオを作成する．しかし，VAE によって 3622 社の株価の予想チャートを生成するのは時間的制約上厳しいため，まず 2 章の方法で 3622 社から 30 社まで企業数を絞り，3 章で VAE を用いて，投資期間において株価が上昇すると判断された企業をピックアップした．最後に 4 章で, 3 章までに選ばれた企業に関して Twitter の感情分析をし，そのスコアをもとに最終的に投資する企業を決定した.

2. 平均・標準偏差を用いた企業の絞り込み

本章では，東証 1 部・2 部，マザーズ，JASDAQ に上場している日本企業 3622 社から 30 社まで企業数を絞りこむ方法について説明する．絞り込む基準として，2016 年～2018 年の 3 年間の 2 月の月次収益率に注目し，その平均÷標準偏差の値が大きい 30 社に絞った．平均÷標準偏差の値が大きい企業は，2 月の月次収益率が安定して高い企業であると言える．また, データを 2 月のみにした理由は, 本コンペの投資期間が 2 月であり，2 月に安定的に高い収益率を生んでいる企業が今年も高い収益率を生むと考えたためである．企業 i の 2016 年 2 月の月次収益率 2017 年 2 月の月次収益率をr^i_{2017}, 2018 年 2 月の月次収益率をr^i_{2018}とすると，企業 i の平均$E_i(r)$と標準偏差$sd_i(r)$の値は以下の式(2-1)～(2-3) によって算出した．ただし，数式の導出の便宜上，標準偏差$sd_i(r)$の導出過程で，分散$Var_i(r)$の導出も行っている．

$$E_i(r) = \frac{1}{3}\left(r^i_{2016} + r^i_{2017} + r^i_{2018}\right) \tag{2-1}$$

$$Var_i(r) = \frac{1}{3}\left\{\left(r^i_{2016} - E_i(r)\right)^2 + \left(r^i_{2017} - E_i(r)\right)^2 + \left(r^i_{2018} - E_i(r)\right)^2\right\} \tag{2-2}$$

$$sd_i(r) = \sqrt{Var_i(r)} \tag{2-3}$$

平均/標準偏差の値をもとに絞り込んだ企業 30 社を以下の表 2-1 に示す.

表 2-1. 平均・標準偏差の値をもとに絞り込んだ 30 社

企業コード	企業名
4310	ドリームインキュベータ
4679	田谷
7169	ニュートン・フィナンシャル・コンサルティング
7521	ムサシ
2608	ボーソー油脂
4151	協和発酵キリン
1799	第一建設工業
1435	ＴＡＴＥＲＵ
3856	Ａｂａｌａｎｃｅ
9424	日本通信
6930	日本アンテナ
4662	フォーカスシステムズ
6920	レーザーテック
6820	アイコム
3360	シップヘルスケアホールディングス
6069	トレンダーズ
2208	ブルボン
1420	サンヨーホームズ
9422	コネクシオ
1780	ヤマウラ
3929	ソーシャルワイヤー
3640	電算
7472	鳥羽洋行
4651	サニックス
4508	田辺三菱製薬
7814	日本創発グループ
2694	ジー・テイスト
2904	一正蒲鉾
3195	ジェネレーションパス
2170	リンクアンドモチベーション

3. Variational Auto Encoder

深層学習における生成モデルの 1 つとして Variational Auto Encoder (VAE) が存在する。VAE とは砂時計型の構造をしており、入力データ x を特定の分布からサンプリングされると仮定する潜在変数 z として潜在空間に埋め込み、埋め込まれた潜在変数 z から再度入力データ x を復元するモデルである。これにより、潜在空間からデータ x'を復元するための潜在変数 z'をサンプリングすることで入力データ集合 D に存在しないデータ x'を作り出すことが可能となるモデルである。

3.1 データの種類

訓練データ、テストデータとして本コンペにて提供していただいた POL コンテスト用データを利用した。株価データとして 2016 年から 2018 年に掛けての 3 年分を利用した。

3.2 前処理、加工方法

前処理の手順としては大きく 3 段階に分けられる。第 1 段階は今回の株価データとしては不適切な ETF を取り除いた。第 2 段階は株価の統合や分割が行われたことによって年間で株価が 2 倍以上変動してしまっている企業の株価を取り除いた。第 3 段階は株価の変動幅の情報を持ったグラフが必要だったため、第 1、第 2 段階の処理を行った上で残った企業の月ごとの株価を図 3-1 のようにろうそくチャートにして画像化を行った。訓練の際には軸とグリッドは必要ないため、軸とグリッドを取り除き、月間で最も下がった株価を下限、最も上がった株価を上限としてグラフの作成を行った。ただし、ZCA 白色化などの前処理は行っていない。

前処理を行った画像のサイズは 32x32、チャネル数は 3 (RGB) としている。訓練データは当月株価画像と次月株価画像のセットであるため、当月株価画像は 1 月から 11 月、次月株価画像は 2 月から 12 月で構成されており、トータル 5500 枚ずつの訓練データとした。検証用データは 800 枚ずつ、テストデータは 922 枚ずつとした。

図 3-1. ろうそくチャート

3.3 VAE の株価予想への対応

VAE は図 3-2 にあるようなデータ x を生成するための潜在変数 z を推論することが主目的であり、学習データから以下の式(3-1)の周辺尤度を最大化することで潜在変数を推定する。

$$p(x;\theta) = \int p(x|z;\theta)p(z;\theta)dz \qquad (3-1)$$

しかし、事後分布$p(x|z;\theta)$は推定が困難なため近似とズ数、epoch はエポック数を表している。

して $q(z|x;\Phi)$という分布を仮定する。最終的に求めるものは Evidence Lower BOund(ELBO)であり、ELBO の最大化が目的である。ELBO の最大化を行うことで事後分布$p(x|z;\theta)$の近似を行うことができる。ELBO の式は以下の式(3-2)の通りである。

$$L(\theta,\Phi;x,z) = \int q(z|x)\log p(x|z)dz - D_{KL}[q(z|x)||p(z)] \qquad (3-2)$$

今回の実験では株価データ x から株価データ x を復元するための潜在空間に仮定する分布を求めるのではなく、当月株価データ x から次月株価データ y を復元するための潜在空間に仮定する分布を求める。それゆえ、求める ELBO は式(3-3)のように変更を行った。証明は Appendix の項目 A に記載する。

$$L(\theta,\Phi;x,y,z) = \int q(z|x)\log p(y|z)dz - D_{KL}[q(z|x)||p(z)] \qquad (3-3)$$

式(3-3)の第 1 項は入力データ x を潜在空間へ潜在変数 z として埋め込み、潜在変数 z から予測データ y を復元するための再構成度合を表している。第 2 項は潜在変数 z と仮定されている潜在変数の事前分布の分布間の距離を表している。実際の計算の際には ELBO の最大化することは困難になるため、式(3-3)を符号反転させ最小化問題として対処をする。また、最終的に式(3-3)第 1 項をさらに符号反転させて再構成度合を再構成誤差とし、再構成誤差と式(3-3)第 2 項の分布間の距離を表すカルバックライブラーダイバージェンス (KL ダイバージェンス) 両者の最小化を行う。

以上により、訓練後の VAE に株価データを入力することで入力された株価の次月の株価画像を出力することが可能となる。

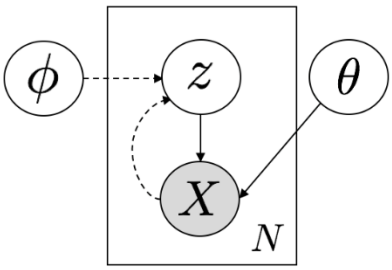

図 3-2. VAE のグラフィカルモデル

3.4 実験結果

実験を行う際の諸元を表 3-1 に示す。今回、潜在変数 z に仮定する事前分布は正規分布を仮定している。また、再構成誤差には Mean Squared Error (MSE) を用いている。

β は KL ダイバージェンスの寄与の割合を、k は潜在空間から潜在変数 z をサンプリングする回数、α は学習率、β1、β2、ε、η は Optimizer のパラメータ (今回の Optimizer は Adam を使用)、w は重みパラメータの初期化に用いる平均 0 の正規分布の標準偏差の値、z_dim は潜在変数 z の次元数、slope は活性化関数 Leaky_ReLU の傾きの値、batchsize はミニバッチサイ

本実験では評価として訓練データ、検証用データ、テストデータでの再構成誤差の数値を用いる。また、本コンペでは最終的に月末の株価が上がったか下がったかを求めることが重要であるため、予測株価の上がったか下がったかの 2 値と実際の株価の上がったか下がったかの 2 値での正誤判定の比較を行う。

3.4.1 再構成誤差

再構成誤差には MSE を用いているため、単純に予測画像と真の画像の画素値の違いが出てくる。再構成誤差
が小さいほど、真の画像を予測できていると考えられる。以下の図 3-3 に訓練データと検証用データの再構成ロスカーブを示す。青い再構成ロスカーブが訓練データ、緑の再構成ロスカーブが検証用データとなっている。

図 3-3 より、70epoch あたりから検証用データの再構成誤差が収束の動きを見せなくなっている。これは当月株価と次月株価の間の関係性が一意ではないため、VAE の十分な汎化が行えていないのではないかと考えられる。

表 3-1. 実験諸元

パラメータ名	パラメータ値
β	1
k	1
α (Optimizer)	0.001
β 1(Optimizer)	0.9
β 2(Optimizer)	0.999
ε (Optimizer)	1e-08
η (Optimizer)	1
w	0.02
z_dim	50
slope	0.2
batchsize	1000
epoch	500

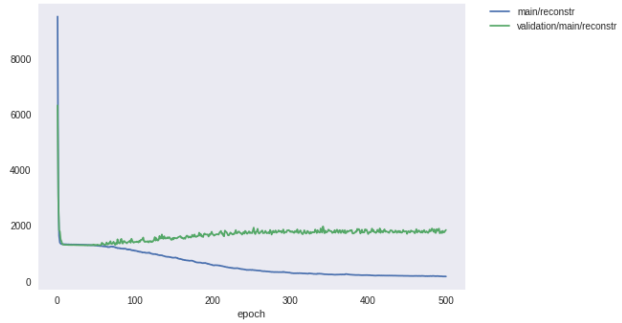

図 3-3. 訓練データと検証用データの再構成ロスカーブ

訓練を 500epoch 終えた時点での訓練データの再構成誤差は 171.14、検証用データは 1841.84、テストデータは 1853.35 である。以下図 3-4 にテストデータを用いて予測した次月株価の画像を示す。図 3-4 は左からテストデータ当月株価画像、次月株価画像、次月予測株価画像となっている。

図 3-4 テストデータ（左二つの画像）と予測株価画像の一例

3.4.2 予測株価月末上下変動の正誤判定

予測画像の株価予想性能を検証するために予測株価月末上下変動の正誤判定を行う。予測が合っていた場合は 1 を加算し、間違っていた場合には 0 を加算する。最終的にテストデータの数で割り、正答率を出す。

しかし、VAE において再構成された画像は株価の変動を数値から作ったものではなく、画像としての相関から再構成されたものであるため、最終的な株価が上がっているか下がっているかを自動的に評価することは困難である。それゆえ、今回は目視での正誤判定を行うことで正答率を求めた。

正答数は 551 枚、テストデータは 922 枚であり、正答率は 59.76%であった。

3.5 考察

§3.4.1、§3.4.2 の実験より考察を行う。

§3.4.1 の図 3-4 から同一の形とまではいかないが、概形は予測できていると考えられる。また、70epoch あたりから検証用データの再構成誤差が収束の動きを見せなくなっていたが、今回は可能な限りシャープな画像を生成するために 500epoch まで訓練を行った。70epoch 時点での検証用データの再構成誤差は 1500 ほどであったが、500epoch でも 1800 ほどであったので、予測には 500epoch での訓練済みモデルを使用した。

§3.4.2 において VAE の次月株価の予測精度は 59.76%となった。VAE 単体としての精度としてはあまり高くはないが、平均/標準偏差の値をもとに絞り込んだ企業 30 社、Twitter においてのネガポジ分析を組み合わせてポートフォリオを作ることでさらに精度が向上することが見込まれる。

また、今回は株価のみから予測した上、不適切なデータ（訓練データの中に月間の株価の値の数が少なすぎる場合など）が入っていたこと、訓練データの数が少なすぎるといったことがあったためこのような精度になったが、上記のことを改善することで更なる VAE 単体での精度の向上が見込まれ、ポテンシャルが高い株価予測モデルだと考えられる。

4. Twitter Sentiment Analysis

今回取り組んでいるテーマは株価予測であるが，株式投資の世界では会社の業績結果の報告以上に株価が上昇するようなケースがよくあり，そのような場合，企業に対する期待や関心といった投資家の感情が大きく関係している．

そこで企業に対する感情情報のリソースとして，Twitter[1]が挙げられる．Twitter は一般に普及している SNS であり，またツイートは 140 字という短文の為トピックが絞られやすく，感情が頻出しやすい傾向にある．

以上のことから，ここでは Twitter の感情情報から特定銘柄の株価を予測することを試みた．具体的には，企業名をキーワードとして含む Tweet のテキスト情報を取得し，既存の感情辞書をもとに感情指数値を算出した. そして, その感情指数値をもとに購入する銘柄を決定した.

4.1 概要

表 2-1 で挙げられた企業 30 社をキーワードとして含む Tweet のテキスト情報を，以下の条件のもとで取得した.

- 投資開始日から，過去 7 日間まで最新の Tweet から取得
- Tweet の取得件数は最大 100 件まで

これは，Tweet を取得するための Twitter API[2]の制限を考慮したものではあるが，企業に対する感情はできる限り最新の Tweet から考慮することが望ましいため，このような条件のもとで Tweet のテキスト情報を取得することにした.

各企業に対する Tweet のテキスト情報を CSV ファイルに保存し，MeCab[3]を用いて形態素解析を行い，既存の感情辞書をもとに感情分析を行った.

今回の株価予測では，5 銘柄から 10 銘柄のポートフォリオを組むという条件があるため，最終的にここで得られた感情指数値をもとにポートフォリオを決定した.

4.2 感情分析

感情分析を行うにあたり，東北大学の乾・鈴木研究室が作成した日本語評価極性辞書[4]を用いた.

感情としてはポジティブ・ネガティブ・ニュートラルの 3 通りが挙げられ，今回は各感情に対するスコアを以下のように設定した.

表 4-1.　各感情に対するスコア

感情	スコア
ポジティブ	1
ネガティブ	−1
ニュートラル	0

そして各 Tweet に対するツイート感情指数値は，以下のように，形態素解析した各語句に対するスコアの和を計算することで算出した．以下では，Nを形態素の個数としている．また $i = 1, 2 \cdots N$ である.

$$\text{ツイート感情指数値} = \sum_{i=1}^{N} \text{語句に対するスコア}_i \quad (4-1)$$

最後に，企業に対する企業感情指数値は，以下のように，企業名をキーワードとして含む Tweet 全てに対するツイート感情指数値の平均を計算することで算出した．以下では，M を企業名をキーワードとして含む Tweet 全ての個数としている. また $i = 1, 2 \cdots M$ である.

$$\text{企業感情指数値} = \sum_{i=1}^{M} \text{ツイート感情指数値}_i \Big/ M \quad (4-2)$$

この企業感情指数値を，各企業について算出した.

4.3 銘柄選択

VAE を用いた株価予測の結果から，表 2-1 で挙げられた 30 社の企業が 11 社にまで絞られた．今回は 5 銘柄から 10 銘柄のポートフォリオを組むため，ここで企業感情指数値を用いて銘柄選択を行う.

30 社の企業感情指数値を見ると，中央値は 0.6 付近であったため，今回は銘柄を選択するにあたって，以下の条件を採用した.

- 企業感情指数値が 0.5 以上
- Tweet の取得件数が最低でも 10 件以上

この条件のもとで選択した銘柄は以下の通りである.

	stock	score	n_tweets	33業種区分
0	ブルボン	0.940000	100	食料品
1	シップヘルスケアホールディングス	0.916667	12	卸売業
2	サンヨーホームズ	0.900000	20	建設業
3	ソーシャルワイヤー	0.869565	23	情報・通信業
4	コネクシオ	0.714286	14	情報・通信業
5	ジェネレーションパス	0.666667	84	小売業
6	レーザーテック	0.580000	100	電気機器

図 4-1.　選択した銘柄

4.4 ポートフォリオ

選択した銘柄でポートフォリオを組むにあたって，以下の条件を採用した.

- 業種で等分
- 業種が被る場合，その業種の中でさらに等分
- 全ての単元株が 100 であったため，予算 1000 万に収まる範囲で適度に調整

---

[1] Twitter, http://twitter.com

[2] Twitter API, http://developer.twitter.com/

[3] MeCab, https://www.mlab.im.dendai.ac.jp/~yamada/ir/MorphologicalAnalyzer/MeCab.html

[4] 日本語評価極性辞書
http://www.cl.ecei.tohoku.ac.jp/index.php?Open%20Resources%2FJapanese%20Sentiment%20Polarity%20Dictionary#t6684569

Appendix

A. VAE の株価予想への対応の式証明

VAE はモデル p(x)の尤度を最大化することを目的としているが、今回はモデル p(y) (y は次月株価データ)の尤度を最大化することが目的である。以下にモデル p(y)に対数を取ったものを書き下す。ただし、仮定の分布として$q(z|x)$を用いる。

$$\begin{aligned}\log p(y) &= \log \int p(y,z)dz \\ &= \log \int \frac{q(z|x)}{q(z|x)} p(y,z)dz \\ &\geqq \int q(z|x) log \frac{p(y,z)}{q(z|x)} dz = L(x,y,z) \end{aligned} \quad (A.1)$$

ここで$L(x,y,z)$は変分下限である。続いてlogp(y)と変分下限のギャップを求める。ただし、3 行目においてz全てに対しての積分であるので$\int q(z|x)dz = 1$となる。4 行目はベイズの定理を用い、またlogp(y)は積分に関係ないので、積分記号の内側に入れる。5 行目は第 2 項 log の中身を分解し、第 1 項と打ち消す。

$$\begin{aligned}&\log p(y) - L(x,y,z) \\ &= \log p(y) - \int q(z|x) log \frac{p(y,z)}{q(z|x)} dz \\ &= \log p(y) \int q(z|x)dz - \int q(z|x) log \frac{p(y,z)}{q(z|x)} dz \\ &= \int q(z|x) \log p(y)\, dz - \int q(z|x) log \frac{p(z|y)}{q(z|x)} p(y) dz \\ &= \int q(z|x)\{-logp(z|y) + \log(q|x)\} dz \\ &= \int q(z|x) log \frac{q(z|x)}{p(z|y)} p(y) dz \\ &= D_{KL}[q(z|x)||p(z|y)] \end{aligned} \quad (A.2)$$

式(A.2)より変分下限は以下のように表せる。ただし、θ と ϕ は分布 p と q が依存するパラメータである。

$$L(x,y,z) = \log(y) - D_{KL}\left[q_\varphi(z|x)||p_\theta(z|y)\right] \quad (A.3)$$

ここで式(A.3)の第 1 項は固定値であるが、第 2 項はカルバックライブラーダイバージェンスであるため最小値が 0 であるため、求めると以下のようになる。ただし、3 行目はベイズの定理を用いている。4 行目の$logp(y)$は z の積分に関与しないので、期待値計算の外に出す。

$$\begin{aligned}&D_{KL}[q(z|x)||p(z|y)] \\ &= \mathbb{E}_{q_\varphi(z|x)}\left[logq_\varphi(z|x) - logp_\theta(z|y)\right] \\ &= \mathbb{E}_{q_\varphi(z|x)}\left[logq_\varphi(z|x) - logp_\theta(y|z) - logp(z) + logp(y)\right] \\ &= \mathbb{E}_{q_\varphi(z|x)}\left[logq_\varphi(z|x) - logp_\theta(y|z) - logp(z)\right] + logp(y) \\ &= \mathbb{E}_{q_\varphi(z|x)}\left[logq_\varphi(z|x) - logp(z)\right] - \mathbb{E}_{q_\varphi(z|x)}[logp_\theta(y|z)] + logp(y) \\ &= D_{KL}\left[q_\varphi(z|x)||p(z)\right] - \mathbb{E}_{q_\varphi(z|x)}[logp_\theta(y|z)] + logp(y) \end{aligned} \quad (A.4)$$

式(A.4)を式(A.3)に代入すると以下のようになる。

$$\begin{aligned}L(x,y,z) &= \log p(y) - D_{KL}\left[q_\varphi(z|x)||p(z)\right] \\ &\quad + \mathbb{E}_{q_\varphi(z|x)}[logp_\theta(y|z)] - logp(y) \\ &= \mathbb{E}_{q_\varphi(z|x)}[logp_\theta(y|z)] - D_{KL}\left[q_\varphi(z|x)||p(z)\right] \end{aligned} \quad (A.5)$$

以上より、§1.3 式(3)の証明を示した。

参考文献

[1] 著者不明(2018)「pandas, Matplotlib（mpl_finance）でローソク足チャートを作成」, <https://note.nkmk.me/python-pandas-matplotlib-candlestick-chart/>2018 年 1 月アクセス

[2] まつけん(2017)「Variational Auto Encoder 徹底解説」,< https://qiita.com/kenmatsu4/items/b029d697e9995d93aa24>2018 年 1 月アクセス

[3] nzw(2016)「Variational Auto Encoder」, <https://nzw0301.github.io/notes/vae.pdf>2018 年 1 月アクセス

D. P. Kingma and M. Welling, “Auto-Encoding Varitional Bayes”, arXiv preprint arXiv: 1312.6114, 2013.