Title

菊地　翔馬

北海道大学大学院情報科学研究科
kicchi@ist.hokudai.ac.jp

白川　稜

北海道大学大学院情報科学研究科
sira@ist.hokudai.ac.jp

1　はじめに

グラフはネットワークや化学構造式、構文木など広く用いられる重要なデータ構造として知られている。また近年ではグラフコンボリューションネットワーク(GNN)と呼ばれるグラフデータに対する畳み込みネットワークの手法が注目を集め、様々な分野への応用先が発見されるとともに、高いパフォーマンスを発揮している。

そこで本研究では、株価データに対して企業間の類似度を加味したグラフ構造での表現方法を提案し、作成したグラフ上でグラフコンボリューションネットワークの手法を利用することにより、類似企業情報を考慮した株価回帰モデルの構築法を提案する。加えて、作成したグラフ集合において頻出クリークマイニング技術を利用することによるポートフォリオ作成上でのリスク分散手法も提案する。提案手法に対して株価の実データを用いて実験を行い、その精度および有用性を評価する。

2　データ

2.1　種類

データには株価における四本値（始値、高値、安値、終値）と出来高、またそれらの値から算出される代表的なインジケーター（SMA、EMA、SMMA、MOMENTUM、BBAND_UPPER、BBAND_LOWER、MACD、MACD_SIGNAL、RSI、STOCH、STOCHAS_SIGNAL、PSAR）の全17種の値を利用する。

2.2　グラフ表現

ここでは上記のデータに対してグラフ構造での表現方法を提案する。まず初めに、各企業において株価のスケールが異なると類似度の評価が困難になるため、各属性値を企業ごとに標準化し、その値を各企業の特徴ベクトルとする。次に、各企業間の類似度の指標に各企業の特徴ベクトル間の相関係数の値を採用し、全ての企業間での類似度を計算する。企業Xの特徴ベクトルが$(x_1, x_2, \cdots, x_n)$、企業Yの特徴ベクトルが$(y_1, y_2, \cdots, y_n)$の時、企業X, Yの類似度(X, Y)は以下となる。

$$\text{類似度}(X,Y) = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}} \quad (1)$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

ノードを各企業と見なし、一般的に高い相関があるとされる0.8という閾値を基準に、類似度が閾値以上であればノード間にエッジを張るという操作を全ての企業間で行うことで１つの無向グラフを作成する。以上の操作を本研究でのグラフ表現方法とし、これにより作成するグラフを「相関グラフ」（図1）と定義する。

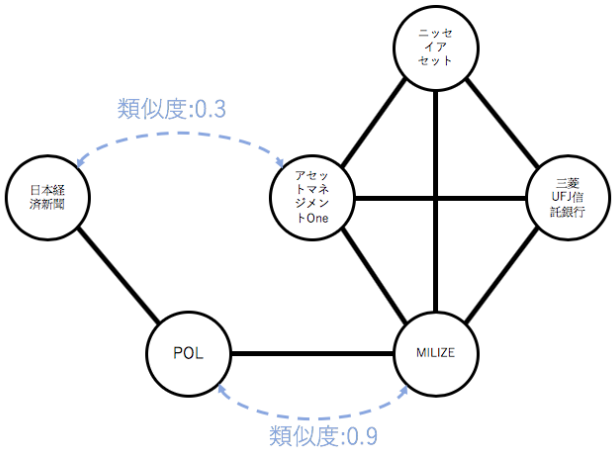

図　1: 相関グラフ

3　モデル

3.1　GNN

hoge

3.2　頻出クリークマイニング

ここではポートフォリオ作成時のリスク分散について考える。株価というものは非常に繊細であり、それまでの株価の値動きからだけでは予測できない場合が存在する。例を挙げるとリーマンショックやITバブルなど、世界全体もしくは一部業界の株価が急激に変動す

ることがある。こういったケースを考慮すると、一部の業界のみでポートフォリオを作成することはリスク分散の観点から見ると良いものではない。複数の業界からポートフォリオを作成する場合の方が、一部業界の株価が暴落した際にも対応しやすい。しかし株価というものは同時にすごく複雑なもので、業界が同じだからといって同じ値動きをするとは限らない、むしろ違う業界の企業間で似た動きをするものも存在する。

この点に関して、本研究ではグラフの頻出クリーク構造という点に着目してリスク分散に対するアプローチを取る。グラフの頻出クリーク構造とは複数グラフに出現するクリーク構造のことを刺し、ここで扱うグラフは類似度を表現するため、値動きが類似している回数が多く見られる企業の集合を表す。本提案では頻出度50%以上の企業集合に関しては、 1つの企業しかポートフォリオに採用しないことでリスク分散を図る。

（注：このリスク分散の提案に関しては計算時間の都合上結果に反映されていない）

4 実験＆結果

4.1 データセット

日経225（2019.02.10時点）の週足の株価データ2016から2018の 3 年分のデータを扱う。また目的変数は、（本コンペティションの問題設定上）説明変数に対応する時刻の三週間後の終値とする。

4.2 実験

hoge

4.3 結果

hoge

5 考察

hoge

6 展望

edgeの重みを考慮したバージョン
回帰にRNNを使う、各企業ごとに(R)NNをフィッティング
計算時間を考慮すると、グラフ全体で特徴ベクトル生成→回帰