IT 企業の株価データを基にした K-MeansClustering 結果の考察

毎田　定弘†

†産業技術大学院大学 〒140-0011 東京都品川区東大井 1-10-40
E-mail:　†a14z7ym@aiit.ac.jp

あらまし　本レポートは、東証に上場している情報通信業に分類される IT 企業の株価データを基に K-MeansClustering を実施し、クラスタリング結果による分類の特徴を考察する。IT 企業の選定は「企業の成熟度」を重視し、東証一部・二部から 42 社、マザーズ・ジャスダックから 45 社の計 87 社を無作為に抽出した。また、本レポートは 2016 年 2 月 4 日を基点とし、株式市場の営業日 30 日間のリターンインデックスを基に分析を行った。本レポートは手法として、Python の numpy、pandas、scikit-learn 等のデータ分析用ライブラリを用いてクラスタリング等を実施した。

キーワード　株価データ，ビッグデータ，クラスタリング，K-Means，テクニカル指標

1. 問題・目的

1.1. 問題

企業に勤めるものにとって、自社の業績、または興味ある企業の業績が、市場でどう評価されているのかは興味あるものである。それを測る指標として、株価データがある。株価データは企業の業績と密接な関係があるといわれている。それは投資家たちが将来の業績に対して投資・投機を行うからである。また、本授業名でもある「ビッグデータ」にとって、株価データは 2 つの意味で利点がある。まず、無料で誰でも入手することができること。そして、提供的なデータであるため分析しやすいことである。そのため、本レポートは企業の業績を示す指標となる“リアル”なデータである株価データを用いることにする。

1.2. 目的

本レポートの目的は、企業の特性によって株価変動に規則的な変動があるかを調べることである。現在、東証に上場している情報通信業に分類される IT 企業の株価データを対象とした。今回は、東証上場銘柄一覧(参考文献[1])より、企業の成熟度という観点で、東証一部・二部から 42 社、マザーズ・ジャスダックから 45 社を無作為に抽出し、計 87 社を調査対象とした。株価データは 2016 年 2 月 4 日を基点とし、そこからそれ以前の営業日 30 日間を対象とした。また、規則的な変動を捉えるためにクラスタリングによるグルーピングを実施する。

2. 方法

Python の pandas ライブラリを用いて、株価データをスクレイピングし、scikit-learn ライブラリを使ってクラスタリングを実施した。

2.1. データを取得する

調査対象企業の 30 日分の株価データをインターネットより取得する。その際、注意点として、対象期間中に株式分割によって 1 株の価値が変わる危険性がある。株式分割の実施の前後で株価を連続的にとらえるために、「調整後終値」を算出しているサイトからスクレイピングを実施した。

2.2. データ整形と前処理

(i)データ整形

2.1 で記した通り、本レポートでは、「調整後終値」を株価データとして扱うことにする。また、株価の上限をリターンと呼ぶなら、金融の世界におけるリターンとは通常、ある日を起算日とした資産価格のパーセント変化を指す (参考文献[2]) 。そのため、今回は 2015 年 12 月 25 日の株価を基準に 1 とし、資産の価値がどう変化するかを測る。また、今回、テクニカル指標としてリターンインデックスを採用する(参考文献[3])。これは、投資単位を表す値を持つ時系列なデータである。

(ii)前処理

取得した株価データは何らかの形で欠損値 NaN(Not a Number)が入り込む。今回、欠損値の扱いは、前後の日の株価の平均値に置き換えることで対処した。

2.3. モデルの選択・構築

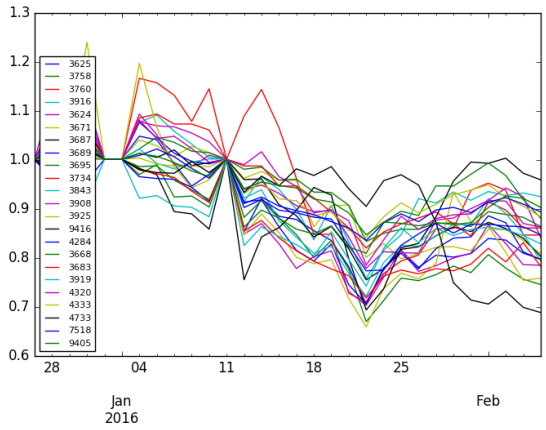
モデルの選択には、scikit-learn algorithm cheat-sheet(参考文献[4])を参考にした。今回、事前にクラスターの数が分かっていない。本講義でも説明があったように、その場合、キャノピークラスタリングを用いるべきであり、また、cheat-sheet によると、MeansShift や VBGMM のアルゴリズムを用いるべきだが、今回は K-Means のアルゴリズムを用いることにす

る。それは、私事だが、K-Means に慣れていて理解がしやすかったからである。そのため、クラスタリングによるカテゴリー数は、クラスタリング実行結果を見て判断した。

3. 結果

K-Means によるクラスタリングは実行結果からカテゴリー数 6 が最適だと判断した。結果は次のとおりである。4 桁の数字は銘柄である。縦軸は、リターンインデックス(1.0 は 2015 年 12 月 25 日の株価を基準としている。)、横軸は日時(2015 年 12 月 25 日～2016 年 2 月 4 日)である。また、日経平均との移動相関を見るために、相関係数(小数点第 2 位までを絶対値表示)を記載している。

●クラスター0 番　（計 23 社）

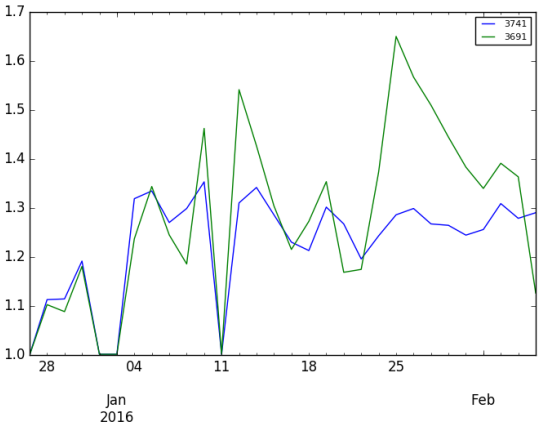


主な企業：

コード	銘柄名	市場	相関関係
3624	アクセルマーク	マザーズ	0.05
3625	テックファームホールディングス	JASDAQ	0.06
3668	コロプラ	東証一部	0.15
4284	ソルクシーズ	東証二部	0.01

特徴：
変動は比較的穏やか。株価は下落傾向である。日経との相関は各企業間でばらつきが大きく、関連性は不明。市場の内訳は、マザーズ 8 社、JASDAQ4 社、東証一部 8 社、東証二部 3 社である。

●クラスター1 番　（計 2 社）

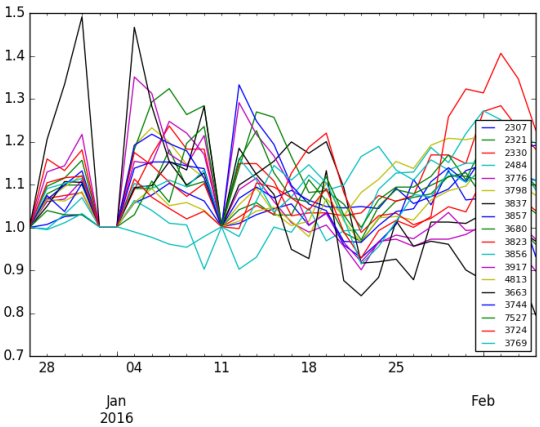


主な企業：

コード	銘柄名	市場	相関係数
3691	リアルワールド	マザーズ	0.21
3741	セック	JASDAQ	0.54

特徴：
株価の変動が激しい。株価は総じてやや上昇。日経との相関はみられる。市場の内訳は、マザーズ 1 社、JASDAQ1 社である。

●クラスター2 番　（計 18 社）

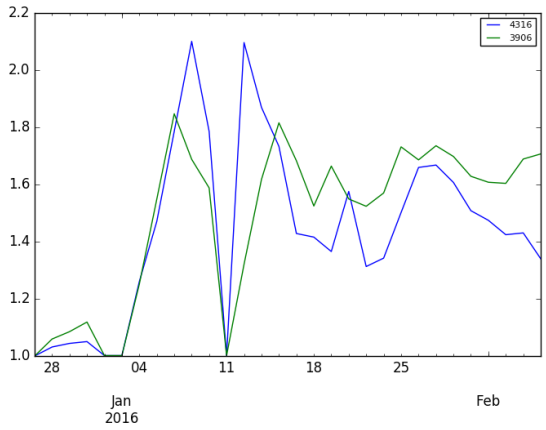


主な企業：

コード	銘柄名	市場	相関関係
2307	クロスキャット	JASDAQ	0.31
3823	アクロディア	マザーズ	0.10
3724	ベリサーブ	東証一部	0.28
7527	システムソフト	東証二部	0.08

特徴：変動は比較的激しい。株価はほぼ一定に保っている。日経との相関は各企業間でばらつきが大きく、関連性は不明。市場の内訳はマザーズ 6 社、JASDAQ7 社、東証一部 2 社、東証二部 3 社である。

●クラスター3 番 (計 2 社)

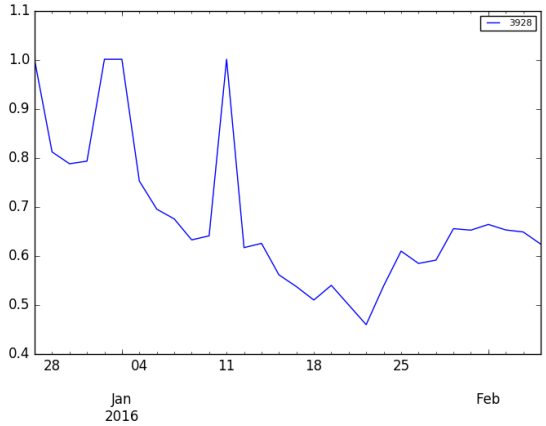


主な企業：

コード	銘柄名	市場	相関係数
3906	ALBERT	マザーズ	0.250
4316	ビーマップ	JASDAQ	0.282

特徴：
株価の変動がかなり激しい。アップダウンを繰り返してはいるが、株価は大きく上昇。1 月 2 週～3 週の値動きが大きい。日経との相関はやや見られる。市場の内訳はマザーズ 1 社、JASDAQ1 社である。

●クラスター4 番 (計 1 社)

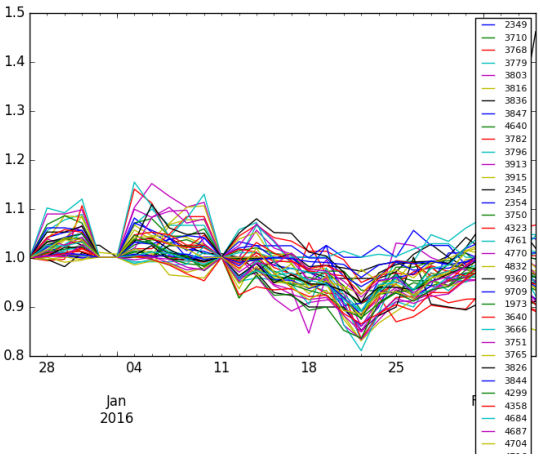


主な企業：

コード	銘柄名	市場	相関係数
3928	マイネット	マザーズ	0.174

特徴：
株価の変動が激しい。かなりの下落傾向。日経との相関はあまり見られない。市場の内訳はマザーズ 1 社である。

●クラスター5 番 (計 41 社)



主な企業：

コード	銘柄名	市場	相関係数
1973	NEC ネッツエスアイ	東証一部	0.092
2345	システム・テクノロジ・ーアイ	東証二部	0.001
2349	エヌアイデイ	JASDAQ	0.15
3782	ディー・ディー・エス	マザーズ	0.08

特徴：株価の変動が穏やか。株価は一定に保たれている。日経との相関は各企業間でばらつきが大きく、関連性は不明。市場の内訳はマザーズ 4 社、JASDAQ9 社、東証一部 19 社、東証二部 9 社である。

4. 考察

4.1. 結果の考察

結果から見て取れることは、マザーズ、ジャスダックは株価の変動が激しいものが多い（クラスター1,3,4 番）。逆に、東証一部、東証二部は株価の変動は比較的落ち着いているものが多い(クラスター0,2,5 番)。
これは、市場に属する企業規模のポジションが、ジャスダック（グロース）→マザーズ→ジャスダック（スタンダード）→東証二部→東証一部とされていることに由来すると考えられる。企業の業績が大きくなるほど、株価も robust（値の変化が小さなもの）になる傾向があると言える。

各企業と日経平均との相関も見てみたが、今回の調

査では、考察に値する結果は得られなかった。それは、各クラスター内での相関関係にかなりばらつきがあるからである。ここからいえることは、日経平均だけを投資の意思決定に用いるべきではないことが言える。

また、今回 30 日間の株価調査で、株価を大きく下げた企業は 1 社(クラスター4 番)、下落傾向は 23 社(クラスター1 番)、値を上げたのは 4 社（クラスタ 1,3 番）だった。ただ、株価の変動が大きい企業は予測が難しい。その企業のニュースを調べても取り立てて大きなニュースも出てこない。話題の商品の付属の部品を扱っていることや、マネーゲームとなっていること等が考えられるが、ビッグデータとしてこれらを考慮して分析するのは至難の業だと考えられる。

4.2. 分析方法の考察

今回、情報通信業に分類される企業を扱ったが、サンプル数が少ない。そして、扱う期間も短い。また、カテゴリが分からない状態で K-MeansClustering を実施するのは、もっと大きなビッグデータを扱う場合、困難である。そのため、キャノピークラスタリング、MeansShift や VBGMM 等のアルゴリズムで、もっと多くの企業を扱い、5 日、25 日、75 日移動平均線等のテクニカル指標を比較すること等が精度の高い分析結果となってくる。

4.3. 今後の展望

今後の展望として、日経平均以外の指標との関連、またはそれらを組み合わせて分析することは非常に有益だと考えられる。

また、決定木アルゴリズム等の教師あり学習で、今回求めたリターンインデックスを学習させ、今後の予測を立てることも可能だと考えられる。

また、Hadoop 分散処理系との比較を行うことも非常に有益な情報を得られると考えられる。

文　献

[1] 日本取引所グループ：東証上場銘柄一覧 http://www.jpx.co.jp/markets/statistics-equities/misc/01.html

[2] Qiita : scikit-learn を使ってみる (1) - K 平均法によるクラスタリング http://qiita.com/ynakayama/items/2efd2578fab73760c6e0

[3] Qiita : pandas による金融データの分析とその可視化(2) http://qiita.com/ynakayama/items/1801d374224d6914a382

[4] scikit-learn : Choosing the right estimator http://scikit-learn.org/stable/tutorial/machine_learning_map/