Cluster and Regression Technique for Stock Analysis

HOU LAN 41423018

1. INTRODUCTION

A large collection of stock pursuers flow into stock market, yet few of them survive the unpredictable and high-frequent price fluctuations. To identify subtle changes and make a wise investment, we use cluster techniques to find out top companies among those moving together and regression techniques to predict future stock price for the companies. Specifically, we choose 2390 firms and get daily stock data in the period of one recent stock plunge, (Aug. 1, 2015 – Aug. 31, 2015) from NYSE and NASDAQ stock exchanges. Furthermore, we compare the usefulness of each cluster approach based on silhouette score and apply two regression methods for prediction.

2. DATASET AND METHODOLOGY

2.1 Data Description

Previous researches suggest that stock market is affected partly by news about fundamentals. Therefore, we attempt to choose 2390 firms randomly in one recent period experiencing stock plunge – Aug.1, 2015 and Aug.31, 2015.

Table 1. Sample of Original Stock Price DataFrame

Open	High	Low	Close	Volume	Adj Close	tic
9.81999	9.81999	9.57999	9.63999	8966200	22.69849	AA
9.93999	10.06999	9.67999	9.72999	10216900	22.9104	AA
9.90999	10.14999	9.74999	9.75999	10390100	23.05212	AA
9.75999	9.95999	9.63999	9.90999	9430700	23.4064	AA
9.81999	10.05999	9.35999	9.40999	13834700	22.22545	AA

2.2 Data Preprocessing

Data preprocessing is vital for data mining. In this step, we initiate to include checking stock prices and converting forms of variances, especially stock prices to normalized stock prices returns for clustering and to 5-day simple moving average for identifying the

leading stock in its group.

$$\text{return} = \frac{P_1 - P_0}{P_0} \tag{1}$$

$$\text{SMA}(5) = \frac{P_1 + P_2 + P_3 + P_4 + P_5}{5} \tag{2}$$

P is for price and calculated SMA(5) above is the SMA for the fifth day.

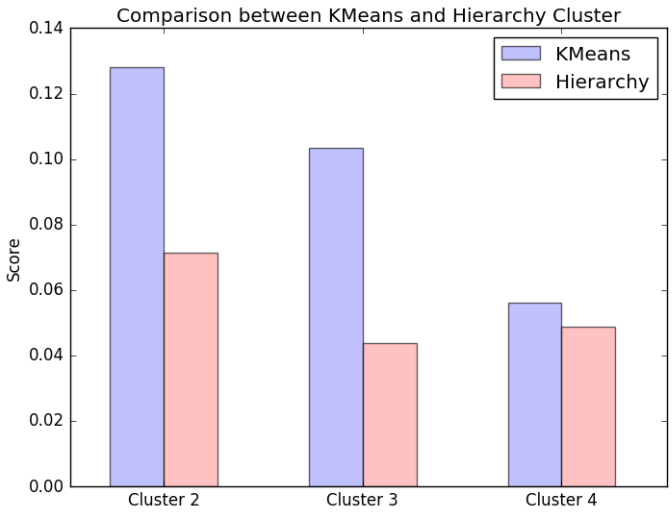
2.3 Cluster

The process for grouping the similar objects together is clustering. The effectiveness of clustering determines results of leading stocks. However, many methods are not appropriate for stock clustering like DBSCAN. Therefore, we choose two typical methods: Partitioning based technique and Hierarchical technique. In the end, we use one popular validation index for clustering algorithms – Silhouette score.

- Partitioning techniques: Partitioning method creates k clusters from a given dataset. This algorithm starts from assigning different objects to different dataset randomly and then at each iteration it reallocate the data objects to another partition for revision. It is not until no further changes that clustering is accomplished. In this method, we focus on K-means as one specific method. Since K should be set in advance, one significant step in K-means is to find out an optimal K. In this paper, we attempt to search elbow point instead.
- Hierarchical techniques: Hierarchical clustering technique creates cluster by data hierarchy and forms a tree based on cluster nodes. In accordance with different directions of decomposition in different hierarchies, this kind of clustering algorithm can be divided into agglomerative and divisive method. In this paper, we initiate hierarchical agglomerative method.
- Validation indexes: Index measure helps to seek out the accuracy of result obtained. Once clustered, it determines the quantity of tuples that are properly labeled and clustered. In this paper, we choose Silhouette score.

  Figure 1 shows that when clustered into two to four, K-Means is better than hierarchy agglomerative method. In this way, we use K-Means for later part of analysis and decide K as three.

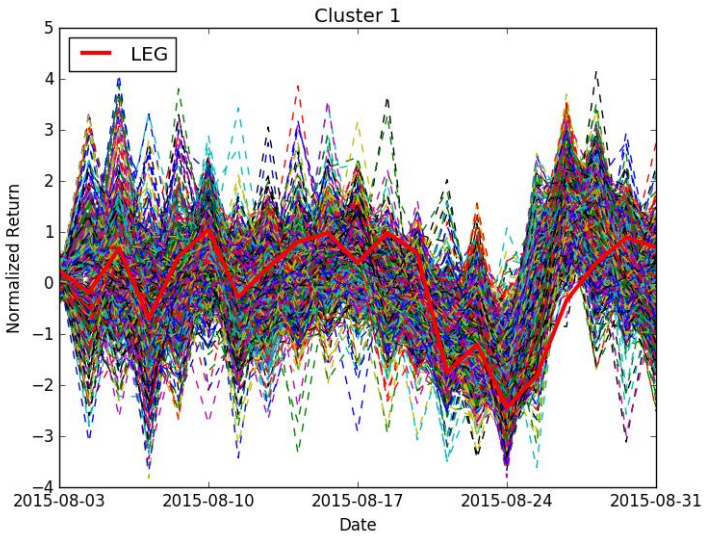
Figure 1. Comparison between K-Means and Hierarchy Cluster



2.4 Correlation

In this section, we use data prepared from all above and then define leading stock as the stock with maximum correlation with future value of other stock. The result of both two clusters of leading stock is similar to Figure 2. In Figure 2, the leading stock is marked with a strong bold line. Since codes of these parts are of great importance, we put them right below.

Figure 2. All stocks in Cluster 1



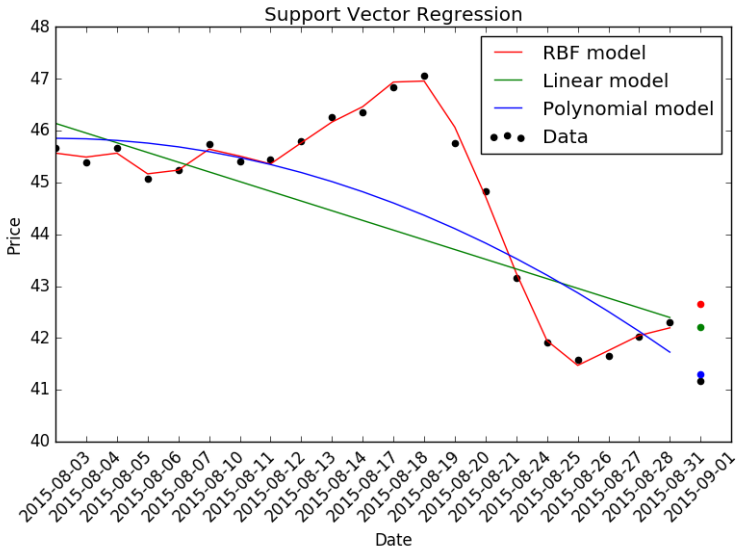

```
def find_leadstock(cluster):
    tic = list(cluster.columns.unique())
    ret_lag = pd.DataFrame()
    fo r i in tic:
```

```
        ret_lag[i+'_lag'] = cluster[i].rolling(window = 5, center =
False).mean()

    df_tmp = ret_lag.join(cluster)
    df_tmp.fillna(0,inplace=True)
    corr_df = df_tmp.corr(method='pearson')
    leader=
np.mean(abs(corr_df.ix[len(cluster.columns):,:(len(cluster.columns)
-1)]).T).idxmax()

    return leader
```

2.5 Regression

Regression is used for predicting an outcome based on a given input. Support Vector Regression is popularly considered to be a reasonable method for stock price prediction. Thus, we apply it in our studying. Figure 3 shows the results of SVR of LEG with different models, the leading stock of Cluster 2. As we can see, though better fitting the curve, RBF model seems not perfect for future price prediction sometimes.

Figure 3. Support Vector Regression of LEG



3. Reflection

In our present work, we use advanced cluster and regression techniques to successfully identify group leader in a cluster, in which their stock prices move together to some extent. Moreover, we apply SVR to predict future stock prices. However, the result of clustering still cannot be satisfying enough.

REFERENCES

Bini, B. S., & Mathew, T. (2016). Clustering and Regression Techniques for Stock Prediction. *Procedia Technology*, *24*, 1248-1255.

* Since most python tutorials were searched through websites, we didn't enumerate all of them.