PCA analysis and prediction on return of SNP500 dataset

Xu Han

Department of physics, The Hong Kong University of Science and Technology
xhanah@connect.ust.hk

INTRODUCTION

Data dimension reduction is very important to analyzing the given data and predicting the unknown result. So far, we have learnt several methods to do data dimension reduction. PCA , a typical data dimension reduction strategy, could be used to deal with the condition that the number of dimension (p) of data is smaller than that of samples (n). When p is largely greater than n, we cannot rely on this method but turn to others in order to obtain a more reasonable predicted result. In this project, I adopted PCA to analyze the SNP500 data in 4 consecutive years. It was found that PCA could not effectively reduce the dimension of this dataset. Therefore, we then directly used the dataset to do the prediction of return with random forest regression and neural network. The performance of both models will be evaluated and compared.

SNP500 DATASET

The SNP500 dataset contains 452 different stocks' closed prices in 1258 consecutive market days. Based on the class information of the companies, we can divide the whole dataset into 10 subsets. The volume of each subset can be seen below in Fig. 1. Then, for each subset, we calculated the return by using the following formula:

$$R = \frac{p_{t+1} - p_t}{p_t}$$

where p_t and p_{t+1} represent the closed price of two consecutive market days respectively. By transferring the closed price to return, it is much easier for us to capture the unusual change in the stock market (see Fig. 2 'MMM' for example). Those unusual changes can be caused by major events of the corresponding companies. Those major events include stock split, dividend and etc. Thus we did some processing to those data with unusual change to obtain a more reasonable dataset. Finally, in each dataset, we constructed a dataset of 1258 samples where different stocks' returns are defined as features.

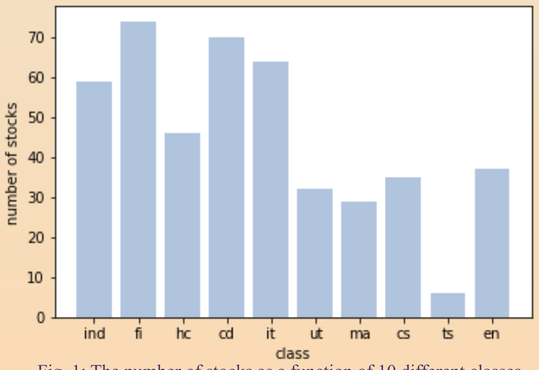

Fig. 1: The number of stocks as a function of 10 different classes.

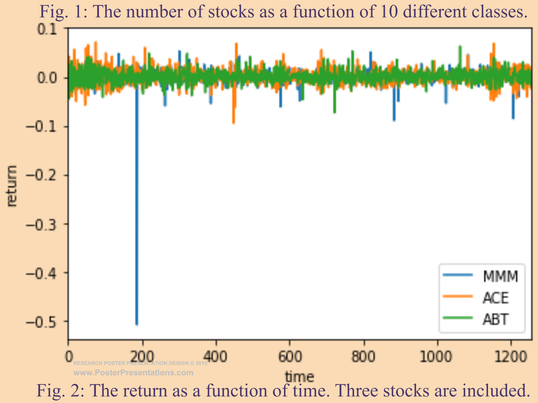

Fig. 2: The return as a function of time. Three stocks are included.

METHODS

Principle component analysis (PCA) invented by Pearson [1] and Hotelling [2], is the most widely used method for dimension reduction with high dimensional Euclidean data.

Give a dataset as $X = [x_1, x_2, \dots, x_n] \in \mathbb{R}^{p\times n}$, where n is the number of samples and p is the dimension, in order to find a k-dimensional affine space in $\mathbb{R}^p$ to best approximate those n samples, the affine space can be parameterized by $\mu + U\beta$ such that $U = [u_1, u_2, \dots, u_n] \in \mathbb{R}^{p\times k}$ consists of

k-columns of an orthonormal basis of the affine space and β is the coordinate under this affine space. The best approximation in terms of Euclidean distance is given by the optimization problem,

$$\min_{\beta,\mu,U} \sum_{i=1}^{n} ||x_i - (\mu + U\beta_i)||^2$$

where $U^T U = I_p$ and $\sum_{i=1}^{n} \beta_i = 0$. After deduction, μ is found to be the sample mean of all observations and U is the top k left singular vectors of the of $\bar{U}$ in the singular value decomposition of $\hat{X} = X - \mu$. Finally, the coordinate of the affine space can be obtained by $\beta_i = U^T(x_i - \mu)$.

Random forest regression is an ensemble learning method for regression by constructing a multitude of decision trees at training time and outputting the mean prediction of the individual trees [3][4].

Neural network itself is not and algorithm, but rather a framework for many different machine learning algorithms to work together and process complex data inputs [5]. These systems learn to perform tasks by considering examples without any pre-designed programs. Apart from the input and output layers, there are still some hidden layers. In these hidden layers, the connection of connected units are called neurons. With bias and weights and the activation function, each neuron is then given a value with suitable optimizer. In this project, the hidden layer was set to 2 and the activation function was set to relu.

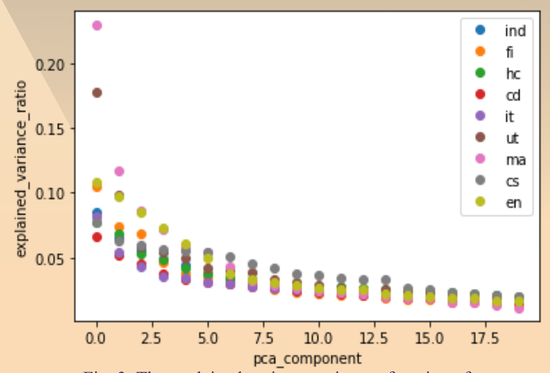

Fig. 3: The explained variance ratio as a function of PCA component.

RESULTS AND DISCUSSIONS

As is shown in Fig. 1, the number of stocks in class 'ts' which is abbreviated for 'Telecommunications Services' is only 6. Therefore, we exclude this class in our PCA analysis. I conducted PCA analysis on 9 different subsets. The result can be seen in Fig. 3. It was discovered that the first PCA component in most classes accounts for 5% to 10% explained variance ratio. Even the largest explained variance ratio of 'Materials' class is only 23%. In this class, the explained variance ratio for the second PCA component is 12%. The comparison can be further studied in Fig. 4, which is a scatter plot of two components. Since the data is centered and scaled when employing PCA, we can see the center of the two components is the origin. And the scale of the first component is a little larger than that of the second component, which can be guaranteed by the explained variance ratio. Furthermore, in order to obtain a 95% total explained variance ratio, we have to include more than 20 components. However, the total number of raw dimension of this class is only 29.

Then, as as to predict the return in a reasonable way, I have to take all these features into consideration. After using random forest regression model to predict the return of the stock in the last column, I can obtain the accuracy of the result, which is represented by mean absolute error (MAE). The cross-validation strategy was applied to improve the accuracy, i.e. reduce the mean absolute error. Apart from this model, I also did the prediction by using the neural network. Apparently, compared to that of random forest regression, the performance of neural network exhibits higher ranges of MAE and even a strange point that it is zero when it comes to 'cs' class. Namely, the random forest regression can provide very stable results in different classes since their MAE are around 0.01.

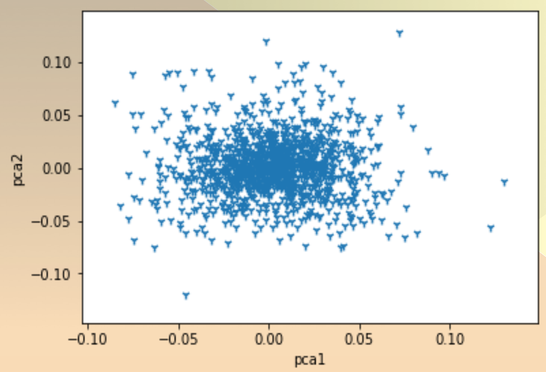

Fig. 4: The second PCA component versus the first PCA component.

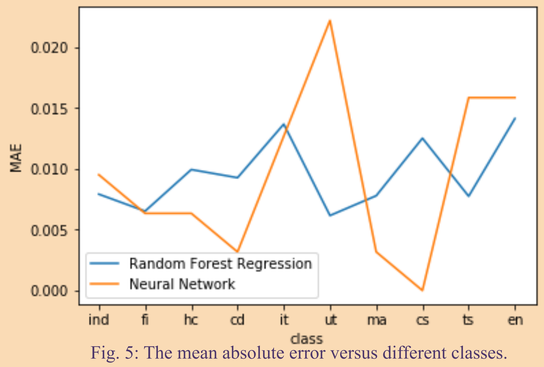

Fig. 5: The mean absolute error versus different classes.

CONCLUSIONS

To summarize, PCA cannot effectively reduce the dimension of our processed return dataset. Since then, we have to use all the given features to do the prediction on the last column's stock's return.

By applying two different models, we find that random forest regression is able to give a more stable result in different classes. It is meaningful to obtain such a good model with finite features. Based on the predicted return, we can adopt some strategies in the stock market to make profits.

REFERENCES

[1] Pearson, K. (1901). Principal components analysis. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 6(2), 559.

[2] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. Journal of educational psychology, 24(6), 417.

[3] Ho, T. K. (1995, August). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition* (Vol. 1, pp. 278-282). IEEE.

[4] Barandiaran, I. (1998). The random subspace method for constructing decision forests. IEEE transactions on pattern analysis and machine intelligence, 20(8).

[5] "Build with AI | DeepAI". DeepAI. Retrieved 2018-10-06.

ACKNOWLEDGEMENT

This work was finished solely by Xu Han.