Visualization and Classification on Finance Data

Zhenghui Chen

Department of Chemical and Biological Engineering
HKUST
zchenef@connect.ust.hk

Abstract

In this project, we perform several dimensionality reduction methods on finance data and visualize the data distribution on first two components. Then we use machine learning algorithms to classify different types of finance data and compare the trade-off between the classification accuracy and data dimension. We achieve this visualization and classification task by python machine learning package scikit-learn [1]. The result shows that Principal Component Analysis (PCA) and Logistic Regression (LR) model achieve the best classification performance.

1 Introduction

Data dimension reduction is the process of transforming high dimensional data into a new representation with low dimension and preserve essential information. It helps to remove redundant noise and break the curse of dimensionality. Besides, it will reduce computational cost for model training and improve the performance of some algorithms. The common application is to visualize the high dimensional data on two-dimensional surface.

In this project, we firstly pre-process and normalize the finance data. Then Principal Component Analysis (PCA), Sparse PCA (SPCA) and Multi-Dimension Scaling (MDS) using Euclidean representations are applied to reduce the dimension of finance data. Manifold learning methods like Isometric Mapping (ISOMAP), Locally linear Embedding (LLE) and T-distributed Stochastic Neighbour Embedding (TSNE) are also utilized. We visualize the data distribution on first two principle components.

Finally, we use supervised learning methods including Random Forest (RF) and Logistic Regression (LR) to classify 9 different types of stocks based on the transformed data. We combine different data reduction with different machined learning methods, evaluate the performance of classification model and explore the relation between classification accuracy and data dimension.

2 Dataset

The finance dataset SNP500 contains 452×1258 matrix, which represents the closed price of stocks from 452 American company in 1258 consecutive workdays. These stocks reflect the economy of the United States in some aspect.

Figure 1 shows the time series of price from 452 different stocks. Stock of different company has its unique curve and even the staring price is not the same.

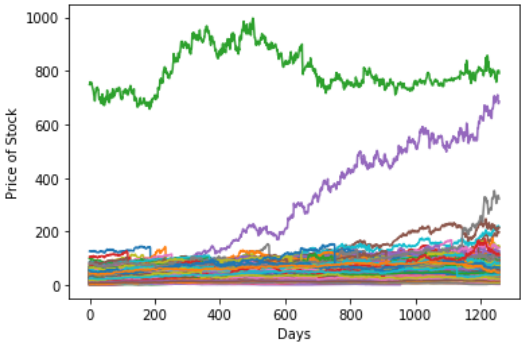


Figure 1: Time Series of Stock Price

From the class information of stock, we can divide the dataset into 10 different classes including Industrials (IND), Financials (FIN), Health Care (HC), Consumer Discretionary (CD), Information Technology (IT), Utilities (UT), Materials (MA), Consumer Staples (CS), Telecommunications Services (TS), Energy (EN). The detailed distribution of stock class can be seen in Figure 2.

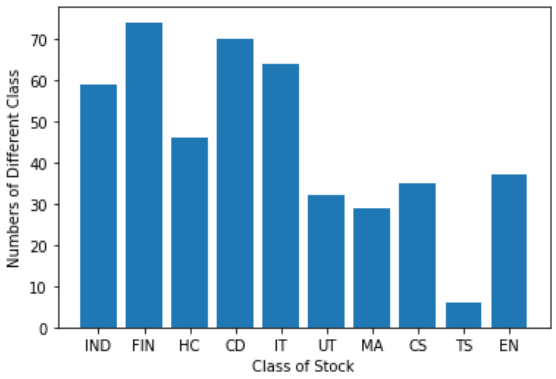


Figure 2: The structure diagram of VGG16

For original time series data, we perform some data pre-processing methods. Firstly, we calculate the increase rate of stock price by the following formula:

$$Rate = \frac{P_{t+1} - P_t}{P_t}$$

where P_t represents the closed price of previous day and P_{t+1} represents the closed price of current day. Figure 3 shows the time series of increase rate from 452 different stocks. As we can see from the figure, there is high increase rate for some stocks at certain day. These changes may be caused by some significant events of the corresponding company.

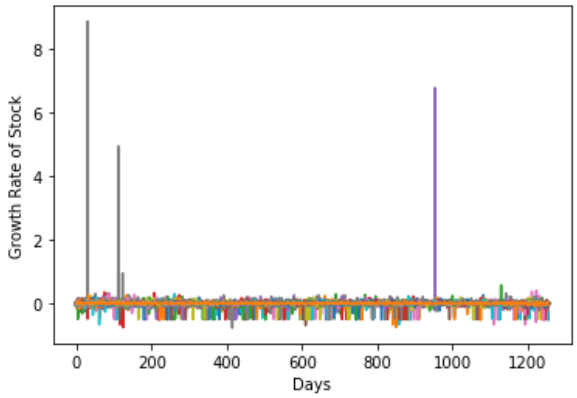


Figure 3: Time Series of Increase Rate

Secondly, we standardize the increase rate by setting the mean value of each feature to zero and the variance of each feature to one. We use this normalized dataset (452×1257 matrix) in the following experiment. Figure 4 shows the increase rate distribution of one stock after data normalization.

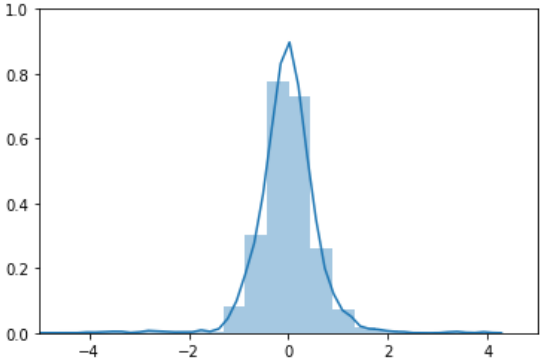


Figure 4: Normalization Result

3 Visualization

3.1 PCA

Principal Component Analysis (PCA) is widely used in data processing and dimensionality reduction. It uses orthogonal transformation to transform a set of possible variable correlation data into a set of linear uncorrelated variables, and the converted variables are called principal components.

We firstly use PCA to convert the dataset into low dimension and maintain 95% information. Figure 5 shows the explained variance ratio of different principal components. The first principal component (PC1) contributes about 6% of explained variance ratio and the second principal component (PC2) contributes about 5% of explained variance ratio. Then we project the data on the space of PC1 and PC2.

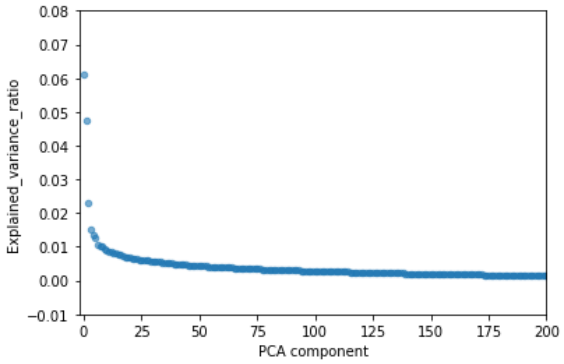


Figure 5: Explained Variance Ratio of Principal Components

Figure 6(a) shows the visualization result and different classes are represented by different colours. From the figure, we notice that some stocks from the same class group together and the cluster separate well with other classes. For example, the blue dots (Energy) and purple dots (Information Technology) are well separated from other class. The brown dots (Utilities) and gray dots (Consumer Staples) are grouped well as clusters. It shows that PCA method has a good performance on data reduction and visualization.

3.2 SPCA

Sparse Principal Component Analysis (SPCA) [2] using the lasso to produce modified principal components with sparse loadings. It can be applied to better interpret the data compared to the PCA, which suffers from linear combination of original variable.

Figure 6(b) shows the visualization result by SPCA. Compared to the PCA result from Figure 6(a), the first principal component of SPCA seems to be sparse. However, these two figures reveal similar distribution of different classes. The blue dots (Energy) and pink dots (Materials) keep away from other class. And the brown cluster (Utilities) and gray cluster (Consumer Staples) can be differentiated in the cloud of points. It indicates that SPCA has comparable performance.

3.3 MDS

Multidimensional scaling (MDS) [3] is a means of visualizing the level of similarity of individual cases of a dataset. It uses the pairwise similarity of samples to construct a low-dimensional space, where the distance of each pair of samples is as consistent as possible while in the original high-dimensional space.

Figure 6(c) shows the visualization result by MDS. The different dots scatter and mix, which is not easy to distinguish between diverse class. The cluster of blue dots and purple dots seem to be distinct in the map.

3.4 ISOMAP

ISOMAP is an extension of MDS [4], which is used for nonlinear dimensionality reduction. It maintains the essential geometric structure of nonlinear data. After data transformation, the geodesic distance between arbitrary point pairs will remain unchanged.

Figure 6(d) shows the visualization result by ISOMAP. The same class dots form the cluster and can be differentiated well from other clusters, especially the blue dots and brown dots.

3.5 LLE

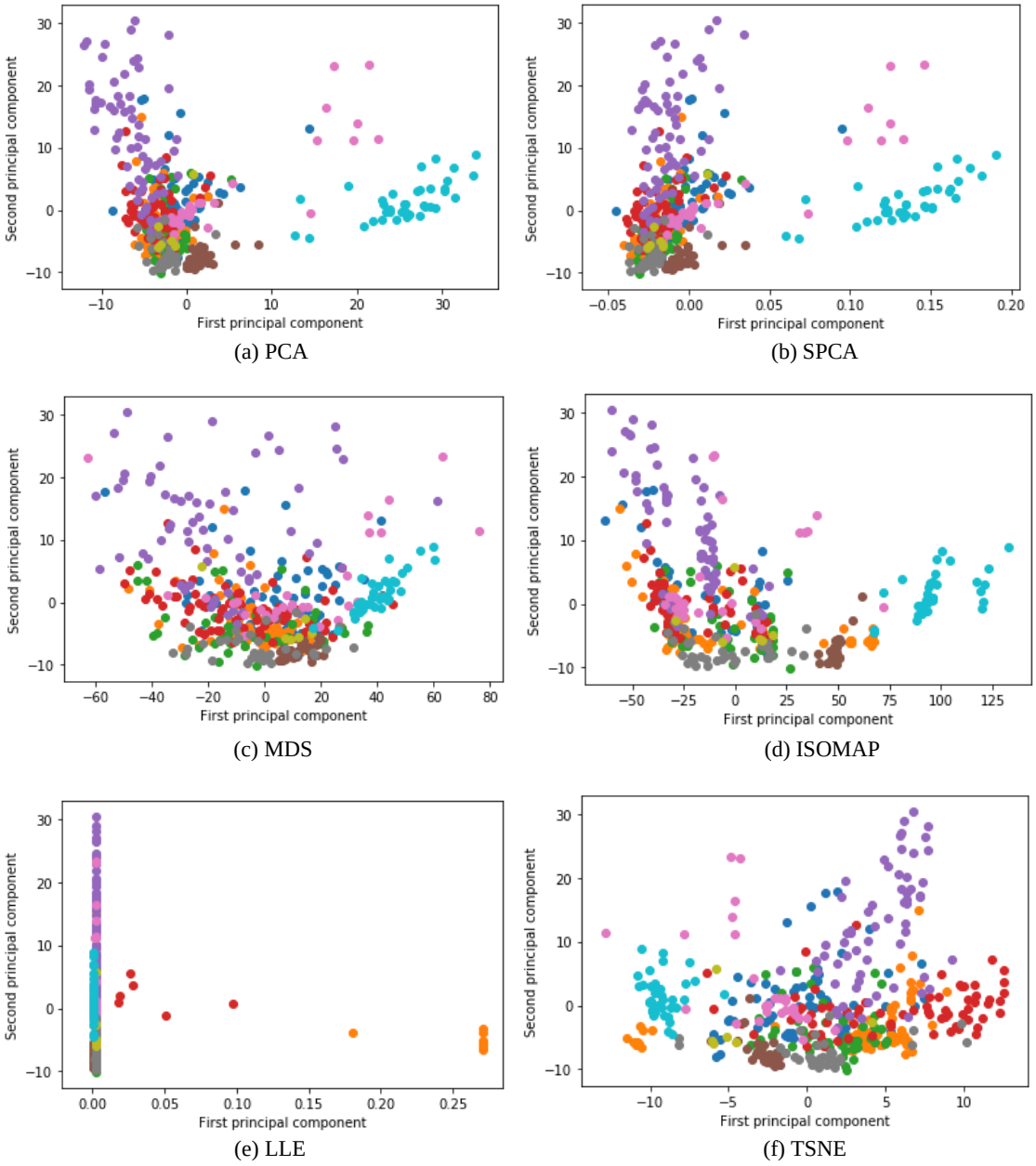
Locally linear Embedding (LLE) [5] is one type of manifold learning method. It focuses on preserving the local linearity of the sample when reduce data dimensionality. It is widely used in image recognition and high dimensional data visualization task.

Figure 6(e) shows the visualization result by LLE. It seems that most of the dots cluster are around zero in x axis, but they have huge difference in the distribution in y axis.

3.6 TSNE

T-distributed Stochastic Neighbour Embedding (TSNE) [6] is derived from SNE algorithm and treat the coordinates in the lower dimension as the t-distribution. It increases the distance between the clusters with large distances and solves the crowding problem.

Figure 6(f) shows the visualization result by TSNE. Different classes can be easily identified through the location of dots in the map, such as gray dots and brown dots.



(a) PCA (b) SPCA

(c) MDS (d) ISOMAP

(e) LLE (f) TSNE

Figure 6: Project Data on PC1 and PC2

4 Classification

In this section, we perform two types of classification methods including Logistic Regression (LR) and Random Forest (RF). Due to the limited class size of Telecommunications Services (TS), we remove this class from normalized dataset. Therefore, we only need to classify 9 different types of based on dataset represented by 446×1257 matrix.

The dataset is split into training set with 312×1257 matrix and test set with 134×1257 matrix. We use original data and transformed data (by PCA, ISOMAP and LLE) to build the classification model, compare the performance between data dimension and overall classification accuracy.

For model based on transformed data, we choose the best test accuracy from a series of different data dimensions. During training phase, we use the embedding model to fit and transform the training data, then train the classifier with the corresponding class labels. During test phase, the test data are transformed by same embedding and predicted by the model.

The results are shown in Table 1. It shows that PCA method can improve the performance of LR model and maintain similar accuracy on RF model with less data dimension. However, IOSMAP and LLE perform not well on LR model because the linear classifier is not suitable with nonlinear features generated by manifold learning. They also have lower accuracy on RF model which may be caused by the loss of information during data reduction.

Table 1: Different Model Performance

Method	Test Accuracy (%)	Dimension
LR	90.3	1257
RF	84.3	1257
PCA+LR	94.8	152
ISOMAP+LR	70.9	52
LLE+ LR	46.3	62
PCA+RF	82.8	102
ISOMAP+RF	70.1	152
LLE+ RF	76.1	182

4.1 Logistic Regression

Logistic Regression (LR) is a generalized linear classification algorithm. Sigmoid function is used to strengthen nonlinear factors and deal with classification problem easily. From Table 1, we can see that the LR model based on untransformed data with 1257 dimension achieves 90.3% test accuracy. Then we apply PCA, ISOMAP and LLE methods to reduce the features from 2 to 292 respectively, evaluate the performance of LR model based on transformed data.

Figure 7(a) shows the classification results of PCA and LR. The model can achieve good training accuracy and test accuracy when the number of principal components reaches 75. And the accuracy nearly stops increasing when the dimension of data higher than 80. The model based on transformed data with 152 features achieve the best test accuracy of 94.8%, which exceeds the model based on original data. It indicates that PCA reserves most important features and decreases redundant information for classification.

Figure 7(c) shows the classification results of LLE and LR. The performance of model is quite bad both in training and test accuracy. The training accuracy is less than 90% even when the dimensions of data reaches 192. The features have no contribution to the test accuracy since the dimension is greater than 75.

Figure 7(e) shows the classification results of ISOMAP and LR. Similarly, the model has poor performance. The training accuracy increases with the increasing number of data features, but the test accuracy even drops in the last half of the curve. This is because linear classifier is not suitable for manifold learning methods.

4.2 Random Forest

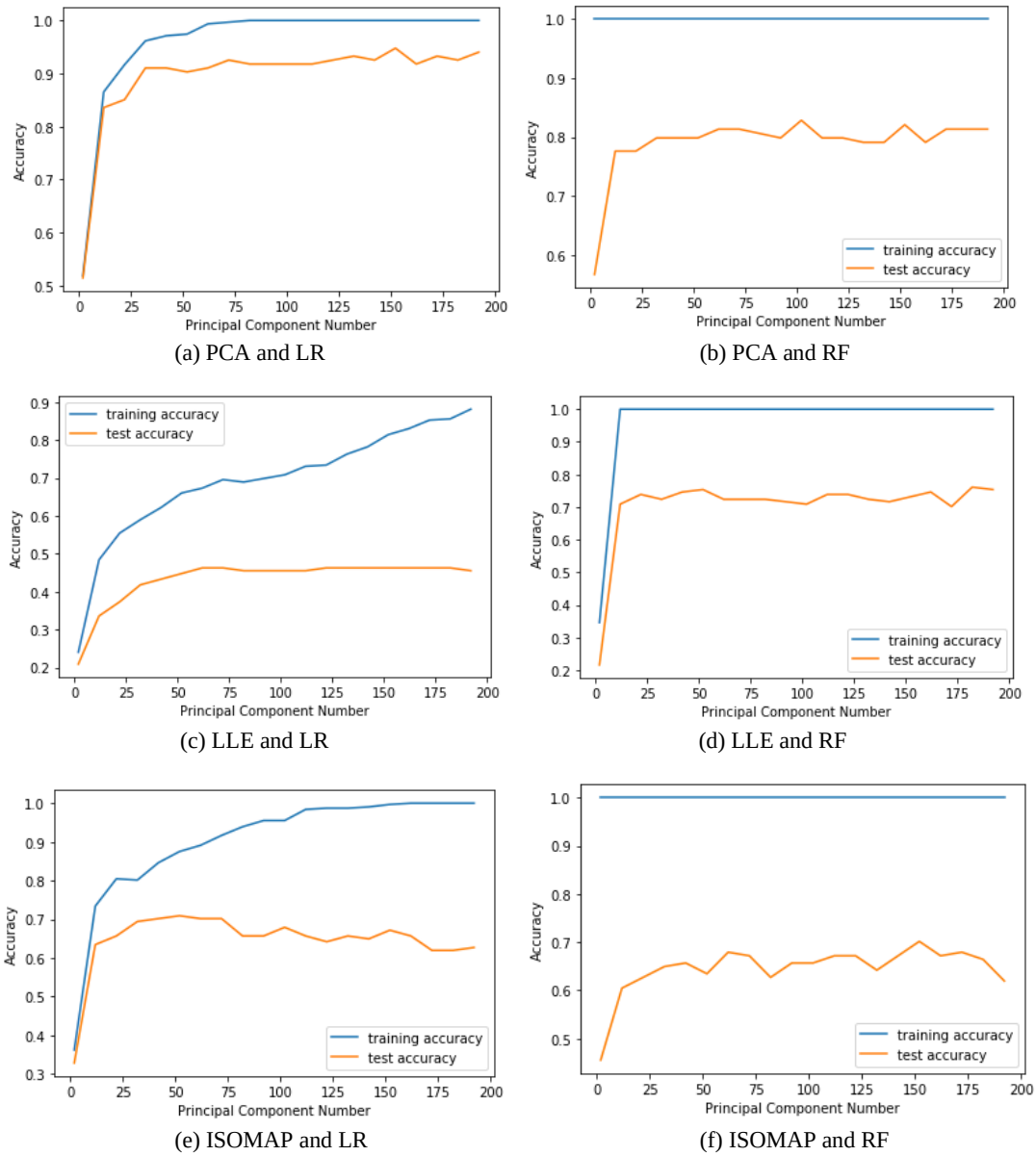
Random Forests (RF) is an ensemble learning method for classification, which helps to solve the overfitting problem of single decision tree. This method constructs a multitude of decision trees and averages the outcomes of these decision trees for prediction.

From Table 1, we can see that the RF model based on untransformed data with 1257 dimension achieves 84.3% test accuracy. Then we apply PCA, ISOMAP and LLE methods to transform data into low dimension, build the RF model based on transformed data.

Figure 7(b) shows the classification results of PCA and RF. The model based on transformed data has similar performance compared with the model based on original data. The model achieves best accuracy of 82.8% when the number of principal components reaches 102. The accuracy nearly stops increasing when the dimension higher than 40, which means we can use less features of the data to build the classification model while maintain good performance.

Figure 7(d) shows the classification results of LLE and RF. The best accuracy drops about 8% compared to the RF model based on 1257-dimensional data. This phenomenon is predictable and acceptable, because the loss of features during data reduction lead to inaccurate prediction.

Figure 7(f) shows the classification results of ISOMAP and RF. The high training accuracy and the low test accuracy may be caused by the overfitting problem and the loss of information during dimension reduction process.



(a) PCA and LR

(b) PCA and RF

(c) LLE and LR

(d) LLE and RF

(e) ISOMAP and LR

(f) ISOMAP and RF

Figure 7: Classification Results with Different Embedding

5 Conclusion

For data visualization part, we perform PCA, SPCA, MDS, ISOMAP, LLE and TSNE methods to reduce the features of data, project the data distribution on low dimension embedding. For methods including PCA, SPCA, ISOMAP and TSNE, stocks with same type group closely with each other and form different clusters of stock class, which can be easily distinguished from other class.

For data classification part, we achieve reasonable test accuracy by different combinations of data reduction and classification algorithms. The PCA method helps to keep critical features for different class and reduce the surplus dimensions, which improves the classification speed and maintains similar accuracy. The combination of PCA and LR model achieves the best result. The ISOMAP and LLE methods are likely to lose some class information. And the LR classifier is not appropriate for manifold learning methods. Maybe classifier like nonlinear support vector machine (SVM) will be more suitable with nonlinear dimensionality reduction method.

In this project, we demonstrate that the data reduction method can contribute to the improvement of classification performance. It will reduce the redundant noise and preserve the important features for classification. However, we should consider and design the different combination of data reduction and classification algorithm carefully. We also need to well evaluate the model to avoid the bad combination.

6 References

[1] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR, 2011.

[2] Structured Sparse Principal Component Analysis, R. Jenatton, G. Obozinski, F. Bach, 2009.

[3] Modern Multidimensional Scaling, I. Borg, P. Groenen, Springer Series in Statistics, 1997.

[4] A global geometric framework for nonlinear dimensionality reduction, J.B. Tenenbaum, V. De Silva et al., Science, 2000.

[5] Nonlinear dimensionality reduction by locally linear embedding, S. Roweis, L. Saul, Science, 2000.

[6] Visualizing High-Dimensional Data Using t-SNE, L.J.P. van der Maaten, G. Hinton, Journal of Machine Learning Research, 2008