Clustering Analysis of Stocks of CSI 300 Index Based on Manifold Learning

Ruiling Liu¹, Hengjin Cai*, Cheng Luo¹,²

¹International School of Software, Wuhan University, Wuhan, China; ²Institute of High Energy Physics, Chinese Academy of Sciences, Beijing, China.
Email: * hjcai@whu.edu.cn

Received November 5th, 2011; revised March 2nd, 2012; accepted March 10th, 2012

ABSTRACT

As an effective way in finding the underlying parameters of a high-dimension space, manifold learning is popular in nonlinear dimensionality reduction which makes high-dimensional data easily to be observed and analyzed. In this paper, Isomap, one of the most famous manifold learning algorithms, is applied to process closing prices of stocks of CSI 300 index from September 2009 to October 2011. Results indicate that Isomap algorithm not only reduces dimensionality of stock data successfully, but also classifies most stocks according to their trends efficiently.

Keywords: Manifold Learning; Isomap; Nonlinear Dimensionality Reduction; Stock Clustering

1. Introduction

Clustering analysis of stocks is necessary when investigating in stocks, Yu and Wang [1] proposed an approach in which kernel principal component analysis is used to reduce the dimensionality of data and k-means clustering method is used to cluster the reduced data so that stocks can be divided into different categories in terms of their financial information. Xu et al. [2] apply unsupervised self-organizing map network, also called SOM, to analyze and cluster stocks. Zhou et al. [3] apply clustering analysis in stock investment. By a synthetic evaluation index system to measure the similarity of stocks, investors can use the evaluation deciding scope and supposing possible variation of stock price. All above studies used financial indicators of the companies, like asset ratio, current ratio, EPS (Earning per Share), ROE (Return on Equity), as foundation of estimation dividing stocks into Blue-chip stocks and junk stocks. The results provided us with a good basis of assessing merits of stocks. Assaleh et al. [4] predict stock price in the Dubai Financial Market from historical price data using the Polynomial Classifiers (PC) and Artificial Neural Networks (ANN). Qin Qin et al. [5] use two different types of prediction models and multiple technical indicators to predict the Shanghai Composite Index returns and price volatility. These methods are proved to be effective by experiments.

In this paper, we will use closing price of stocks and cluster these stocks according to their trends. Stock movement is volatile and uncertain. But from stock charts, difference of trends between different stocks reflects on their price, whether they are rising or dropping and the degree of their movements. That is to say, though stock data is high-dimensional and nonlinear, there are only three underlying parameters. As for how to find these underlying parameters, manifold learning provides a good solution.

Manifold learning is an important field of machine learning which acquires very good achievements in exploring inner law of nonlinear data. It assumes data is even sampled from a low dimensional manifold embedded in a high dimensional Euclidean space. The purpose of manifold learning is to learn this low dimensional manifold from high dimensional data, to get the relevant embedded mapping and finally to reduce the dimension of the data and realize visualization.

Isomap (Isometric Feature Mapping) and LLE (Locally Linear Embedding) are two very famous nonlinear dimensionality reduction algorithm. Isomap which is proposed by Tenenbaum [6] and his workmates get approximate geodesic distance from shortest path of nearest neighbor graph firstly to instead Euclidean distance which cannot represent the inner manifold structure. And then input geodesic distances into MDS to find the low dimensional coordinate embedded in high dimensional space. LLE which is proposed by Roweis and Saul [7] can map high dimensional input data to a global low dimensional coordinate, and maintain relation of neighbors, thus keep original topology structure for data after dimensionality reduction.
Some scholars use these data dimensionality reduction methods into other fields and get many interesting results. Reference [8] use shapes of mouth as input of LLE and get a low dimensional manifold. Then finding curves of each word formed on manifold, by analyzing these curves we can read the lip language. Reference [9] improve the conventional Isomap algorithm by utilizing class membership for measuring distance of data pair so as to find a low-dimensional manifold preserving the distance between classes as well as the distance between data points. Through computational experiments on artificial data sets and real facial data sets, they confirm that the proposed method gives better performance than the conventional Isomap. Reference [10] does the emulation experiment of EEG (electroencephalogram) generation source by combining Isomap and SVM (Support Vector Machine). Combining with Rényi entropy, manifold learning is used on face recognition in Reference [11].

The present paper is organized as follows. In Section 2, the source of experiment data and their pre-processing is introduced. In Section 3, Isomap is applied to stocks’ clustering. In Section 4, we compare the results of stocks’ clustering using LLE with using Isomap. And Section 5 concludes the whole paper.

2. Data Selection and Pre-Processing

CSI (China Securities Index) 300 Index reflects overall trends of A-share market and review of changes of stocks prices of Chinese stock market whose samples cover 60% of capitalization of Shanghai and Shenzhen market. It has a good representation of market. So we choose stocks of CSI 300 Index as our experimental subjects.

We take closing price of these 300 stocks for 500 days from September 22nd, 2009 to October 18th, 2011. When the stock stopped quotation on the exchange, we use the closing price of the day before to fill the gap. If the listing of a stock was after September 22nd, 2009, we use the closing price of the day before to fill the gap. If the stock stopped quotation on the exchange, we number them from 1 to 300 and divide them into 5 groups, each group corresponds to one color on the distribution graph, and data points of same color are not necessarily in a same classification.

3. Clustering Analysis of CSI 300 Stocks Using Isomap

3.1. Dimensionality Reduction of the Stocks

We run Isomap codes1 on MATLAB 7.6 (R2008a). By changing the value of parameter k, we get a series of residual variance graphs and data distribution graphs which are low dimensional embedding recovered by Isomap. Residual variance graph (Figure 1) can help to test the effectiveness of dimensionality reduction and decide the low dimension in which data embedded. Data distribution graph (Figure 2) can help to observe results of stock clustering directly.

Figure 1 shows the residual variance graph when \( k = 7 \). When changing the value of \( k \), we get nearly same results, so we use \( k = 7 \) as an example to analyze the results. X-Label represents dimensionality; Y-Label represents residual variance. Residual variance decreases as the dimensionality is increased [10]. With increasing of dimensionality, it is getting close to true dimension, error between current and true dimension is reducing, so value of residual variance is decreasing. Curves in Figure 1 reduced as dimensionality increased, so dimensionality reduction of stock data by Isomap is successful.

The “intrinsic” dimensionality of the data can be estimated by looking for the “elbow” at which this curve ceases to decrease significantly with added dimensions [6]. Before the “elbow point”, error between current and true dimensionality is huge. After the “elbow point”, errors of these dimensionalities between true dimensionality are nearly the same. Since we need to observe how data points distribute after dimensionality reduction directly, we need to choose a low dimension while keep the error as small as possible. So “elbow point” is the best choice. In Figure 1, arrow marks the true or approximate dimensionality which is two, so the data points distribute on a two-dimensional rectangular plane coordinate system.

$$S_j = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$$ (3)

\( x_{ij} \) represents original data of row \( i \) column \( j \), \( \bar{x}_j \) represents standardized data of row \( i \) column \( j \), \( \bar{x}_j \) represents mean of original data of column \( j \), \( S_j \) represents standard deviation of original data of column \( j \). In this experiment, \( i \) is from 1 to 500, \( j \) is from 1 to 300. With so many data points to observe, we number them from 1 to 300 and divide them into 5 groups, each group corresponds to one color on the distribution graph, and data points of same color are not necessarily in a same classification.

\[S_j = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2 \]

\[x_{ij} \] represents original data of row \( i \) column \( j \), \( \bar{x}_j \) represents standardized data of row \( i \) column \( j \), \( \bar{x}_j \) represents mean of original data of column \( j \), \( S_j \) represents standard deviation of original data of column \( j \). In this experiment, \( i \) is from 1 to 500, \( j \) is from 1 to 300. With so many data points to observe, we number them from 1 to 300 and divide them into 5 groups, each group corresponds to one color on the distribution graph, and data points of same color are not necessarily in a same classification.

3.1. Dimensionality Reduction of the Stocks

We run Isomap codes1 on MATLAB 7.6 (R2008a). By changing the value of parameter k, we get a series of residual variance graphs and data distribution graphs which are low dimensional embedding recovered by Isomap. Residual variance graph (Figure 1) can help to test the effectiveness of dimensionality reduction and decide the low dimension in which data embedded. Data distribution graph (Figure 2) can help to observe results of stock clustering directly.

Figure 1 shows the residual variance graph when \( k = 7 \). When changing the value of \( k \), we get nearly same results, so we use \( k = 7 \) as an example to analyze the results. X-Label represents dimensionality; Y-Label represents residual variance. Residual variance decreases as the dimensionality is increased [10]. With increasing of dimensionality, it is getting close to true dimension, error between current and true dimension is reducing, so value of residual variance is decreasing. Curves in Figure 1 reduced as dimensionality increased, so dimensionality reduction of stock data by Isomap is successful.

The “intrinsic” dimensionality of the data can be estimated by looking for the “elbow” at which this curve ceases to decrease significantly with added dimensions [6]. Before the “elbow point”, error between current and true dimensionality is huge. After the “elbow point”, errors of these dimensionalities between true dimensionality are nearly the same. Since we need to observe how data points distribute after dimensionality reduction directly, we need to choose a low dimension while keep the error as small as possible. So “elbow point” is the best choice. In Figure 1, arrow marks the true or approximate dimensionality which is two, so the data points distribute on a two-dimensional rectangular plane coordinate system.

1The Isomap code is from J. B. Tenenbaum et al. [6] with some modifications. http://isomap.stanford.edu/
3.2. Clustering Analysis of the Stocks

During the process of dimensionality reduction, Isomap produces data point distribution graph in which data points distribute on a two-dimensional rectangular plane coordinate system (Figure 2). As we can see, except a pile of points at left bottom, most points on the map distribute evenly. We choose 8 groups of them to compare their trends curves. Information of these stocks is in Table 1. Points of each group are close to each other in Figure 2, if Isomap can cluster stocks according to their trends, stocks in same group should have similar trend.

From comparisons of these stock trends in Figure 3, we can see stocks of same group from A1 to A6 are similar both in overall trends and the ups and downs at the same time. While stocks trend curves of same group in A7 and A8 are different from each other. In Figure 2, group A1 and A6 are far apart whose trends are different from each other (Figure 3). So points with different trend do not distribute together.

4. Comparison of LLE and Isomap

We also use LLE\(^2\), another manifold learning algorithm to processing and clustering the same stocks. In this experiment, we reduce the dimensionality of stock data to 2, and map these data to a two-dimensional rectangular plane coordinate system (Figure 4). We also choose 8 groups of points to compare their trend curves. The results are in Figure 5.

From the comparison in Figure 5, we can see that LLE can clustering the stocks that have similar trend curves, but the similarity is not as high as Isomap.

We also conduct following experiments: 1) the chosen stocks distribute together after being processed using Isomap, and are apart from each other after being processed using LLE. The trends of these stocks are highly similar; 2) the chosen stocks distribute together after being processed using LLE, and are apart from each other after being processed using Isomap. The trends of these stocks are not so similar with each other. We conclude that Isomap is more effective in clustering stocks.

---

Table 1. Information of 8 groups of stocks chosen in Figure 2.

<table>
<thead>
<tr>
<th>Group No.</th>
<th>Point No.</th>
<th>Stock Code</th>
<th>Distance</th>
<th>Group No.</th>
<th>Point No.</th>
<th>Stock Code</th>
<th>Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>240</td>
<td>SHA: 600998</td>
<td>0.887</td>
<td>A5</td>
<td>116</td>
<td>SHA: 600058</td>
<td>0.721</td>
</tr>
<tr>
<td></td>
<td>247</td>
<td>SHA: 601098</td>
<td></td>
<td></td>
<td>177</td>
<td>SHA: 600456</td>
<td></td>
</tr>
<tr>
<td>A2</td>
<td>53</td>
<td>SHE: 000825</td>
<td>0.649</td>
<td>A6</td>
<td>55</td>
<td>SHE: 000858</td>
<td>0.577</td>
</tr>
<tr>
<td></td>
<td>108</td>
<td>SHA: 600028</td>
<td></td>
<td></td>
<td>290</td>
<td>SHA: 601888</td>
<td></td>
</tr>
<tr>
<td>A3</td>
<td>52</td>
<td>SHE: 000807</td>
<td>2.174</td>
<td>A7</td>
<td>128</td>
<td>SHA: 600115</td>
<td>0.660</td>
</tr>
<tr>
<td></td>
<td>176</td>
<td>SHA: 600432</td>
<td></td>
<td></td>
<td>161</td>
<td>SHA: 600316</td>
<td></td>
</tr>
<tr>
<td>A4</td>
<td>256</td>
<td>SHA: 601168</td>
<td>3.092</td>
<td>A8</td>
<td>142</td>
<td>SHA: 600183</td>
<td>1.907</td>
</tr>
<tr>
<td></td>
<td>297</td>
<td>SHA: 601958</td>
<td></td>
<td></td>
<td>169</td>
<td>SHA: 600380</td>
<td></td>
</tr>
</tbody>
</table>

Figure 3. Comparison of trend curves of sample stocks chosen in Figure 2, horizontal axis represents time; vertical axis represents relative variation of stock price.
Figure 4. Data point distribution graph of stocks of CSI 300 Index after being processed using LLE. Trend curves of 8 groups (B1-B8) of stocks circled in this figure will be compared in Figure 5, each number in this graph represents a stock listed in Table 2.

Table 2. Information of 8 groups of stocks chosen in Figure 4.

<table>
<thead>
<tr>
<th>Group No.</th>
<th>Point No.</th>
<th>Stock Code</th>
<th>Distance</th>
<th>Group No.</th>
<th>Point No.</th>
<th>Stock Code</th>
<th>Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1</td>
<td>145</td>
<td>SHA: 600208</td>
<td>0.0218</td>
<td>B2</td>
<td>126</td>
<td>SHA: 600109</td>
<td>0.0074</td>
</tr>
<tr>
<td></td>
<td>152</td>
<td>SHA: 600266</td>
<td></td>
<td></td>
<td>186</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B2</td>
<td>187</td>
<td>SHA: 600528</td>
<td></td>
<td>B3</td>
<td>72</td>
<td>SHE: 002007</td>
<td>0.0356</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>98</td>
<td>SHA: 600004</td>
<td></td>
</tr>
<tr>
<td>B3</td>
<td>88</td>
<td>SHE: 002310</td>
<td></td>
<td>B4</td>
<td>210</td>
<td>SHA: 600703</td>
<td>0.0067</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 5. Comparison of trend curves of sample stocks chosen in Figure 4, horizontal axis represents time, vertical axis represents relative variation of stock price.
5. Conclusions

In this paper, we use Isomap processing closing price of stocks of CSI 300 Index. Results show that Isomap algorithm can reduce dimensionality of these high dimensional data. Reduced data points distribute in a two-dimensional coordinate system. By choosing and observing sample points, we find that for most points which distribute together have similar trend, for which distribute away from each other have different trend. There are also points with different trends distributing together. Since most stocks clustering are successful, we think Isomap can cluster stocks according to their trend.

By comparing the effects of clustering by Isomap and by LLE, we consider Isomap as a more effective method. Isomap is a global geometric manifold learning algorithm which is based on MDS (Mutidimensional Scaling). It makes every effort to keep the intrinsic geometry of data points, which is to preserve the geodesic distances between two points. For neighboring points, input-space distance provides a good approximation to geodesic distance. For faraway points, geodesic distance can be approximated by adding up a sequence of “short hops” between neighboring points [6]. While LLE recovers global nonlinear structure from locally linear fits [7]. Isomap plays a better role.

6. Acknowledgements

This study is partly supported by the Basic Research Program of China/973 Project, Basic Research on Data Storage System Theory and Technology for Complex Application Environment, 2011CB302306. The authors would like to thank reviewers for their helpful comments.

REFERENCES