Project MineStock

Mining Time Series Data for Finance Applications

Abstract of Presentation

Prepared by Tong Man Kin (08502621d@polyu.edu.hk)

1. Backgrounds

   a. Portfolio Diversification

   - Derivative securities are so popular nowadays. They are perceived to be able to reduce the risk by hedging, but they actually carry an even more complicated risk structure. Individual investors should do portfolio diversification instead.

   b. Existing Systems

   - There are existing theories in mathematical finance for inducing the optimal weighting of stocks in a portfolio. But currently there is no way to determine whether a stock is suitable to be included in the portfolio.

   c. Project Objectives

   - Clustering techniques are believed to be able to solve the problem. The project is to develop such clustering techniques on stocks. Implement a software tool (MineStock Workbench) which combines the use of clustering techniques on stock and the theory in mathematical finance.

2. Methodologies

   a. Classical K-means Technique

   1. Create after users have defined the number of clusters. Randomly assign an initial stock for each of them.
   2. Asset return for each price tick is treated as an attribute for the stock.
   3. Apply the formula of Euclidean distance ($d_{euc}(S_1, S_2) = \sqrt{\sum_{t=1}^{T} \left(R_{S_1t} - R_{S_2t}\right)^2}$) to measure similarity between stocks and clusters.
   4. Stocks which are similar in price fluctuations will be grouped into one cluster. Stocks which are not similar will be assigned into different clusters.

b. Identical Sequence Extraction

1. Discretization on stocks is needed as a preprocessing step. For example, if we have a stock, $S = \{R_1, R_2, R_3, R_4, R_5, \cdots, R_{365}\}$ and we are trying to discretize them by 3 intervals:

	Result	Remarks
Discretize by price change	$S = \{U, D, N, N, D, D, \cdots, D\}$	U=Up, N=No change, D=Down
Discretize by price level	$S = \{A, A, A, A, A, M, \cdots, B\}$	A=Above mean, M=Mean, B=Below mean

2. First define a window by user desired width, and then shift this window along the discretized series of the stock to count the occurrence of the patterns (subsequence). After that, hashtable for each stock can then be formed:

$$Occurrence_S = \begin{bmatrix} \{N,N,N,N,N\} & 10 \\ \{N,N,N,N,U\} & 16 \\ \{N,N,N,N,D\} & 24 \\ \{N,N,N,U,D\} & 79 \\ \vdots & \vdots \\ \{D,D,D,D,D\} & 7 \end{bmatrix}$$

3. Each row (subsequence occurrence) in the hashtable will be treated as one attribute of the stock.
4. Apply K-means clustering on these stocks by their occurrence of each possible subsequence. If 2 stocks are similar, they should have similar occurrence of patterns.

c. Similar Motif Discovery

1. Same discretization process as sequence extraction.
2. Find the Longest Common Subsequence (LCS, or motif) between stocks

$$S_1 = \{U, U, D, U, D, D, N, D, D, D\}$$

$$S_2 = \{U, D, U, D, D, N, N, D, D, D\}$$

$$LCS(S_1, S_2) = \{U, D, U, D, D, N, D, D, D\}, Size\ of\ LCS = 9$$

3. Compute similarity between stocks = Size of LCS between them ^ a multiplier defined by users. The following similarity matrix can be acquired.

$$Similarity\ matrix = \begin{array}{cc} \begin{array}{cccc} 1.HK & 2.HK & \cdots & 8383.HK \end{array} & \\ \begin{bmatrix} 365 & 106 & \cdots & 13 \\ 106 & 365 & \cdots & 22 \\ \vdots & \vdots & \ddots & \vdots \\ 13 & 22 & \cdots & 365 \end{bmatrix} & \begin{array}{c} 1.HK \\ 2.HK \\ \vdots \\ 8383.HK \end{array} \end{array}$$

4. Apply K-means clustering on the similarity matrix, in order to form clusters of stocks. If 2 stocks are similar, they should have similar similarity to other stocks.

d. Performance Indicators

1. Mean Return ($\bar{R}$)
2. Standard deviation of return (σ)
3. Sharpe ratio ($S = \left(\bar{R} - R_f\right)/\sigma$)
4. Beta ($\beta = Cov(R, R_M)/\sigma_M^2$)
5. Alpha ($\alpha = \bar{R} - \left[\beta\left(R_M - R_f\right) + R_f\right]$)
6. Value at risk (VaR, multiple methods of calculation)

e. Portfolio Weighting Optimization

1. Find the maximum of Sharpe ratio ($S = \left(\overline{R_p} - R_f\right)/\sigma_p$)) by trial and error.
2. Portfolio return is calculated by: $\overline{R_p} = \sum_i^n w_i \bar{R}_\iota$
3. Portfolio variance is calculated by: $\sigma_p^2 = \sum_i^n w_i^2 \sigma_i^2 + \sum_i^n \sum_{j \neq i}^n w_i w_j \sigma_i \sigma_j \sigma_{ij}$

f. Six Steps for Users

1. Download data
2. Preprocessing
3. Clustering
4. Review Indicators
5. Add stocks to portfolio
6. Optimization

$$\left.\begin{array}{l}\text{1. Download data}\\ \text{2. Preprocessing}\\ \text{3. Clustering}\\ \left.\begin{array}{l}\text{4. Review Indicators}\\ \text{5. Add stocks to portfolio}\\ \text{6. Optimization}\end{array}\right\}\text{Normal system}\end{array}\right\}\text{MineStock Workbench}$$

2. Evaluation

a. Low correlation of stocks, then low total risk of portfolio ($\sigma_p^2 = \sum_i^n w_i^2 \sigma_i^2 + \sum_i^n \sum_{j \neq i}^n w_i w_j \sigma_i \sigma_j \sigma_{ij}$), and also a higher Sharpe ratio ($S = \left(\overline{R_p} - R_f\right)/\sigma_p$).

b. Therefore, higher stocks correlation within a cluster, the more effective of the algorithm.

c. In my experiments of 2 years data and 5 clusters, the following results are found:

	Average Intra-cluster Correlation	Improvements
Classical K-means Technique	~8%	N/A
Identical Sequence Extraction	~8%	<1%
Similar Motif Discovery	11~12%	~3% improvement

d. Also, the improvements become far more significant when we are trying to do further clustering on existing clusters.