Analyzing public companies in the life sciences and designing an investment portfolio

Rosa Ma, Vandon Duong
{rosaxma, vandon} @ stanford.edu

Abstract

Stocks of public companies in the biotech sector tend to be extremely volatile and largely dissociated from traditional fundamentals (e.g. EPS, P/E ratio). There are numerous obscure factors that drive the valuation of companies in the life sciences. We propose a deep learning-based investment strategy, which includes diversification through clustering and portfolio construction by selecting stocks in each cluster based onSharpe ratio. Specifically, our clustering model has two phases: (1) parameter initialization with a deep convolutional autoencoder and (2) parameter optimization (i.e., clustering), where we iterate between computing an auxiliary target distribution and minimizing the Kullback-Leibler (KL) divergence to it. Experimental results show the diversification accomplished by our clustering method suggested portfolios that outperform multiple indexes during a downturn in the biotech sector.

Data & Features

- Stock data downloaded from EDE
- 267 stocks (initialization: K = 5)

Processing of daily stock data:
- A = [closing + Volume]
- Low price
- High price
- Open price
- Close price

Training and testing paradigm:
- Evaluation of dataset:
  - General increase in biotech IPOs over time
  - Significant price and volume volatility

Baseline

- K-means clustering (initialization: K-means++)
- Hierarchical clustering (linkage: Ward’s method)

Model

Σ_i (soft assignment) 
\[ q_{ij} = \frac{1 + \| x_i - \mu_j \|^2}{\sigma^2} \]

Student’s t-distribution:
\[ f_\theta(x) = \frac{1}{\sigma \sqrt{2 \pi}} \left(1 + \frac{x^2}{\sigma^2}\right)^{-\frac{1}{2}} \]

Auxiliary target distribution:
\[ p_{ij} = \frac{q_{ij}}{\sum_j q_{ij}} \]

Elbow curves

Results

Evaluation of clustering:
- Clustering based on closing & volume is similar with clustering based on all features (possibly redundancy)
- Parameter optimization using Deep embedded clustering (DEC) makes the clusters well-separated

Risk-adjusted performances of clusters during training

Discussion & Future Directions

- Compared to the general market (S&P 500), the biotech sector (XBI, $XBI, XBI, $XBI, $XBI, $CNCR) had a strong performance during the training period, but significant retracement in the testing period.
- Baseline clustering (hierarchical and K-means) tend to overfit to the training period.
- Using the autoencoder with DEC has a robust portfolio performance across the different clustering models. Clustering based on closing price only is the most effective in mitigating losses.
- The model based on both closing price and volume, and the model based on all features tend to overfit to the training period compared with the model based only on closing price.
- The current clustering method in the custom layer requires pre-specification of the number of clusters and initial node selection. Thus, other methods that do not require pre-specification such as modularity optimization could be incorporated in the future.
- We will interpret the model to understand the nuances of the markets, i.e., what do the features extracted by the model represent. We will also apply the model to other sectors such as automobile, energy, and real estates or even the general market. We expect the model to have similar performance in these markets.

Reference


G. Hu et al., "DEEP STOCK REPRESENTATION LEARNING FROM CANDLESTICK CHARTS TO INVESTMENT DECISIONS."