Clustering based portfolio optimization using investor information in Indian stock market

A B. Tech Project Report Submitted
in Partial Fulfillment of the Requirements
for the Degree of

Bachelor of Technology

by

Patoliya Meetkumar Krushnadas
(150101045)

under the guidance of

Dr. N. Selvaraju and Dr. S. V. Rao

to the

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI
GUWAHATI - 781039, ASSAM

CERTIFICATE

*This is to certify that the work contained in this thesis entitled "**Clustering based portfolio optimization using investor information in Indian stock market**" is a bonafide work of **Patoliya Meetkumar Krushnadas (Roll No. 150101045)**, carried out in the Department of Computer Science and Engineering, Indian Institute of Technology Guwahati under my supervision and that it has not been submitted elsewhere for a degree.*

Supervisor: **Dr. N. Selvaraju**
Professor,
Department of Mathematics,
Indian Institute of Technology, Guwahati
Assam.

Co-supervisor: **Dr. S. V. Rao**
Professor,
Department of CSE,
Indian Institute of Technology, Guwahati
Assam.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In markets various types of investors are there- foreign, institutional, individual. They are having different strategies and competing against each other. Studies have shown that each of these types have their own advantage and disadvantages. Like foreign investors have better market-timing abilities than others. Institutional investors have informational advantage from other types of investors. So we can reap the benefit by emulating these investors actions.

1.1 Brief about study

We would analyze the volume of stocks traded by each of these types of investors and try to identify the stocks of interests for trading using clustering. We can try to make portfolio based on investor information and continuously optimize it over period of time to get higher return than standard asset. The proposed method is employed for top 100 stocks of National Stock Exchange of India (NSE) in terms of market capitalization. (Table 1.1).

1.2 Organization of the report

Literature survey is described in Chapter 2. In next Chapter we have described overall model for portfolio optimization. Chapter 4 presents results and conclusion of the work.

Table 1.1 Top 100 companies of NSE in terms of market capitalization.

Sr. No.	Security Name	Sr. No.	Security Name
1	Reliance Industries Limited	51	Britannia Industries Limited
2	Tata Consultancy Services Limited	52	The New India Assurance Company Limited
3	HDFC Bank Limited	53	Dabur India Limited
4	ITC Limited	54	Shree Cement Limited
5	Housing Development Finance Corporation Limited	55	Bandhan Bank Limited
6	Hindustan Unilever Limited	56	ICICI Prudential Life Insurance Company Limited
7	Maruti Suzuki India Limited	57	Zee Entertainment Enterprises Limited
8	Infosys Limited	58	Bosch Limited
9	Oil & Natural Gas Corporation Limited	59	Indiabulls Housing Finance Limited
10	State Bank of India	60	Hindustan Petroleum Corporation Limited
11	Kotak Mahindra Bank Limited	61	InterGlobe Aviation Limited
12	Larsen & Toubro Limited	62	Hindalco Industries Limited
13	ICICI Bank Limited	63	Pidilite Industries Limited
14	Coal India Limited	64	Ambuja Cements Limited
15	Indian Oil Corporation Limited	65	United Spirits Limited
16	Bharti Airtel Limited	66	Cipla Limited
17	NTPC Limited	67	Piramal Enterprises Limited
18	HCL Technologies Limited	68	Ashok Leyland Limited
19	Axis Bank Limited	69	Marico Limited
20	Wipro Limited	70	Cadila Healthcare Limited
21	Hindustan Zinc Limited	71	Siemens Limited
22	Sun Pharmaceutical Industries Limited	72	Hindustan Aeronautics Limited
23	UltraTech Cement Limited	73	NMDC Limited
24	IndusInd Bank Limited	74	UPL Limited
25	Asian Paints Limited	75	ICICI Lombard General Insurance Company Limited
26	Vedanta Limited	76	DLF Limited
27	Bajaj Finance Limited	77	Biocon Limited
28	Power Grid Corporation of India Limited	78	Bharat Electronics Limited
29	Tata Motors Limited	79	Petronet LNG Limited
30	Bharat Petroleum Corporation Limited	80	Dr. Reddy's Laboratories Limited
31	Mahindra & Mahindra Limited	81	Sun TV Network Limited
32	HDFC Standard Life Insurance Company Limited	82	Lupin Limited
33	Titan Company Limited	83	Idea Cellular Limited
34	Avenue Supermarts Limited	84	Bank of Baroda
35	Bajaj Finserv Limited	85	Aurobindo Pharma Limited
36	Bajaj Auto Limited	86	Shriram Transport Finance Company Limited
37	Eicher Motors Limited	87	Bharat Forge Limited
38	Godrej Consumer Products Limited	88	Aditya Birla Capital Limited
39	GAIL (India) Limited	89	Oracle Financial Services Software Limited
40	Adani Ports and Special Economic Zone Limited	90	Procter & Gamble Hygiene and Health Care Limited
41	Hero MotoCorp Limited	91	MRF Limited
42	Yes Bank Limited	92	Havells India Limited
43	JSW Steel Limited	93	Container Corporation of India Limited
44	Grasim Industries Limited	94	Bharat Heavy Electricals Limited
45	SBI Life Insurance Company Limited	95	Bajaj Holdings & Investment Limited
46	Motherson Sumi Systems Limited	96	TVS Motor Company Limited
47	General Insurance Corporation of India	97	Steel Authority of India Limited
48	Tata Steel Limited	98	Divi's Laboratories Limited
49	Tech Mahindra Limited	99	Colgate Palmolive (India) Limited
50	Bharti Infratel Limited	100	L&T Finance Holdings Limited

Chapter 2

Literature Survey

2.1 Management of stocks portfolio

Portfolio management involves choosing stocks to incorporate into the portfolio and choosing how much capital to distribute to each stock. It additionally includes deciding when it is required to rebalance the portfolio. All through the portfolio procedure, it is important to take investor risk tolerence level into account since certain investors are all the more eager to take higher degree of risk in exchange for possibly higher return.

There are two types of portfolio management strategies:

1. Active strategy
2. Passive strategy

Passive strategy involves replicating particular index of the market. This strategy has low risk and gives low returns. On the other hand, active strategy requires constant monitoring of the market trends to do frequent buy and sell of assets to get more return than index of the market, which comes by cost of higher risk. In this study we will use active portfolio management strategy in order to get higher returns.

2.2 Types of investors and their behaviour information

There are three types of investors.

1. Foreign Investors
2. Institutional Investors
3. Individual Investors

These investors have different level of information, hence they have different trading strategies. Thus various types of investors harvest different returns from the market. Studies [MG00] found that foreign investors, who have a data advantage and solid market-timing capacities, foresee future value returns generally well. Be that as it may, different studies have displayed contrasting results. A study [BB04] has found that institutional investors seems to have edge over other investors that who have an information disadvantage. Researchers have shown agreed upon foreign and institutional investors performing better than individual investors. So we have analyzed institutional and foreign investors trading strategies in our study.

2.3 Clustering of data

Using clustering, we can divide a group of data points into smaller subgroups such that each of the subgroup have members which are similar to each other. A data-point is included in that sub group whose distance (e.g.Euclidean, Manhattan) is minimum among all the sub groups. Clustering algorithms involve unsupervised learning. Research [LAY09] has shown that we can apply k-means algorithm to basket of stocks based on their financial data. In this study we have used Euclidean measure to find distance between two data points.

2.4 Evolutionary algorithm

Evolutionary algorithm, otherwise known as genetic algorithm is method of stochastic optimization. In this algorithm problem solution is initialized as random set of vectors, known as chromosome. These chromosome undergoes various processes like mutation, crossover and selection to generate new chromosome. Over the period of time, only fitter solution survives by replacing weaker solutions from set of chromosomes. When fitness of these chromosome converges, that is newly generated vectors are not fitter than vectors present in set, then the most fittest chromosome is declared as solution of the optimization problem. Research [YO03] has shown that genetic algorithm is proven very effective for portfolio formation. We will use vectors consisting of weights of each stock in portfolio as chromosome for genetic algorithm.

Chapter 3

Model Specification

The model has two steps to get return in a time interval. In first step, we select stocks based on investor information using k-means clustering. In second step, we assign weights to selected stocks using genetic algorithm such that we get optimized portfolio. Note that above two steps are repeated over a timeline using sliding window technique to generate new portfolio in every window.

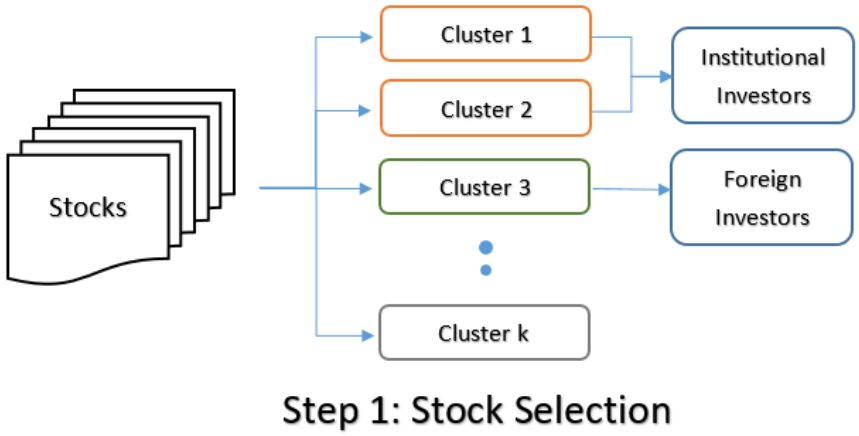

Fig. 3.1 First step of proposed model

3.1 Stock selection

In this step (Fig. 3.1), stocks highly traded stocks are selected based on investor information . These stocks are expected to give higher return. The features used in selecting stocks via k-means are listed in table 3.1.

Table 3.1 Features for k-means clustering (Time period : day t to day T).

Name	**Formula**
Proportion of volume traded by foreign investor	$= \frac{\sum_{i=t}^{T} \text{Net volume traded by foreign investor(i)}}{\sum_{i=t}^{T} \text{Volume traded (i)}}$
Proportion of volume traded by institutional investor	$= \frac{\sum_{i=t}^{T} \text{Net volume traded by institutional investor(i)}}{\sum_{i=t}^{T} \text{Volume traded (i)}}$
Proportion of volume traded by individual investor	$= \frac{\sum_{i=t}^{T} \text{Net volume traded by individual investor(i)}}{\sum_{i=t}^{T} \text{Volume traded (i)}}$
Stock returns	$= \frac{\text{Stock price at time (T) - Stock price at time (t)}}{\text{Stock price at time (t)}}$
Volume Ratio	$= \frac{\sum_{i=t}^{T} \text{Up volume (i)}}{\sum_{i=t}^{T} \text{Down volume (i)}}$

Features for clustering includes trading volume proportions for three types of investors. They indicate % volume traded by particular type of investor over given period of time. Also, return achieved by stock over given period of time is taken as another feature. Volume ratio, a ratio of volume of stock traded when stock price was increasing over volume of stock traded when stock price was decreasing is taken as a feature for clustering. These features are normalized for cluster analysis.

Let X be $n \times m$ matrix, where x_{ij} denotes the j^{th} stock's i^{th} variable ($i = 1, ..., n$ and $j = 1, ..., m$) and x_j be j^{th} column of X.

$$X = (x_1, x_2, ...x_n)$$

$$x_j = (x_{1j}, x_{2j}, ...x_{nj})^T, \text{where } j = 1, ..., m$$

Let S be the set of disjoint clusters s_c.

$$S = \{s_1, s_2, ...s_k\}, \qquad \bigcup_{c=1}^{k} s_c = X, s_c \in S$$

$$s_c = (x_1^c, x_2^2, ..., x_{l(c)}^c), \qquad c = 1, ..., k$$

where, k is total number of cluster formed by k-means clustering and $l(c)$ is the number of stocks for cluster s_c.

The objective of clustering algorithm is to choose stocks on which if trading is performed, yields higher expected returns. Mathematically clustering is made by following formula.

$$S = \arg\min_{S} \sum_{c=1}^{k} \sum_{x_j \in s_c} |x_j - \mu_c|$$

where, μ_c is center of s_c.

Once clusters are formed, they are decreasingly ordered according to mean trading volume proportion values for each investor types. The stocks of cluster with highest mean trading volume proportion are chosen for step of portfolio optimization.

3.2 Portfolio optimization

In this step (Fig. 3.2), we would be assigning weights to stocks selected using clustering in previous step to minimize the portfolio risk for given return or maximizing the portfolio return for given risk. Let w_j is proportion invested in stock j,

$$w = (w_1, w_2, ..., w_n)$$

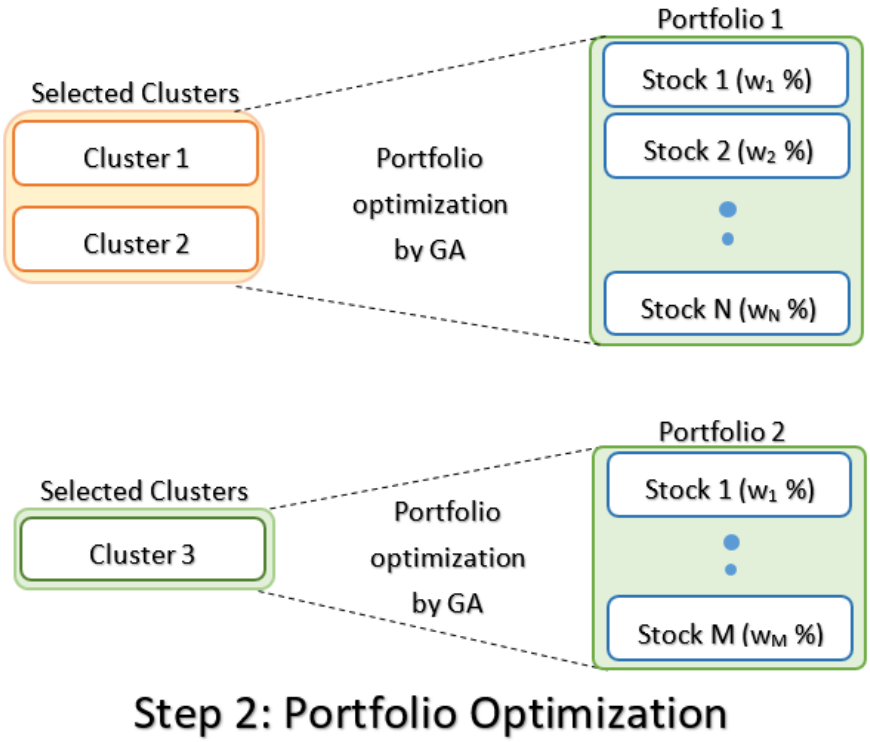

Fig. 3.2 Second step of proposed model

$$\sum_{j=1}^{n} w_j = 1, \qquad w_j \geq 0, \qquad j = 1, 2, ..., n$$

To do portfolio optimization we would be using genetic algorithm. The weight vector is taken as chromosome in our algorithm.

3.2.1 Fitness functions

We will use various fitness functions to determine fitness of chromosomes. Higher the fitness of given chromosome, better is the suitability of portfolio with that chromosome being the weight vector of selected stocks. Fitness functions used are as follows:

Maximum Sharpe ratio of portfolio

In Markowitz theory, the efficient portfolio is based on mean-variance relationship. Sharpe ratio is given by following formula.

$$\text{Sharpe Ratio } S_p = \frac{R_p - R_f}{\sigma_p}$$

where,

R_p is return of portfolio, which is weighted sum of returns on each stock of portfolio.

R_f is return on reference asset (risk free rate).

σ_p is portfolio variance, which is given by following formula.

$$\sigma_p{}^2 = w_p C_p w_p{}^T$$

where,

w_p is weight vector of portfolio

C_p is covariance matrix of stocks selected by clustering.

We will use S_p as our fitness function. We can see that for given amount of risk, portfolio yielding highest return also have highest Sharpe ratio. So optimized weight vector will be such that it will have highest return for given level of risk.

Minimum variance of portfolio

Portfolio variance is weighted sum of variances of individual stocks within portfolio. We will use $-\sigma_p{}^2$ as our fitness formula. Hence, the optimized weight vector obtained by using this fitness function will be such that it will minimize the variance of the portfolio.

3.2.2 Other weight vectors

Apart from the fitness functions to obtain weight vectors, we will also consider following weight vectors directly.

Equal weight vector

Each stock in portfolio is allocated equal amount of capital.

Market Capitalization weight vector

Each stock in portfolio is allocated weight directly proportional to its market capital at that time.

3.2.3 Various evolution processes of genetic algorithm

Population initialization

In this process, n number of chromosomes are randomly initialized. The i^{th} entry in weight vector (i.e. i^{th} gene in the chromosome) corresponds to proportion of capital allocated to i^{th} stock in portfolio.

Selection

In this processes, a chromosome or two having highest fitness among the population are selected for mutation or crossover. For selection, all the chromosomes are sorted by their fitness and chromosomes with highest fitness are chosen thereafter.

Mutation

A chromosome is mutated when it is changed at some point(s). Changed value can be anything random between 0 and 1. There can be several kinds of mutations: Single

point mutation, multi-point mutation etc. The mutation is made in chromosome chosen by Selection process with probability $mutationRate$.

The mutated chromosome is then included in population if it is fitter than least fit chromosome of population by rule of Survival of fittest. Mutation is done in genetic algorithm so that algorithm dont get stuck at local optimal.

Crossover

A pair of chromosomes chosen by Selection process can be crossed over. Process of crossover generates pair of child chromosomes as a result. In crossover each gene (w_i) is derived from one of two parent chromosomes.

The crossover is made in chromosome chosen by Selection process with probability $crossoverRate$.There are various types of crossovers: Single point crossover, k-point crossover, uniform crossover.

Note that after mutation, crossover weight vectors need to be normalized such that sum of weights in vector becomes 1. This process of evolution is repeated till the population converges, i.e. almost no better chromosome than current population can be generated.

3.3 Sliding window technique

The algorithm of the problem has various hyper parameters

- **Stocks Selection hyper parameters:** k, investor type
- **Portfolio optimization hyper parameters:** population size, mutation rate, crossover rate, number of generations, fitness function

This technique (Fig. 3.3) is to be applied for getting optimal values of hyper parameters by doing portfolio rebalancing and measurement of performance. Every window is

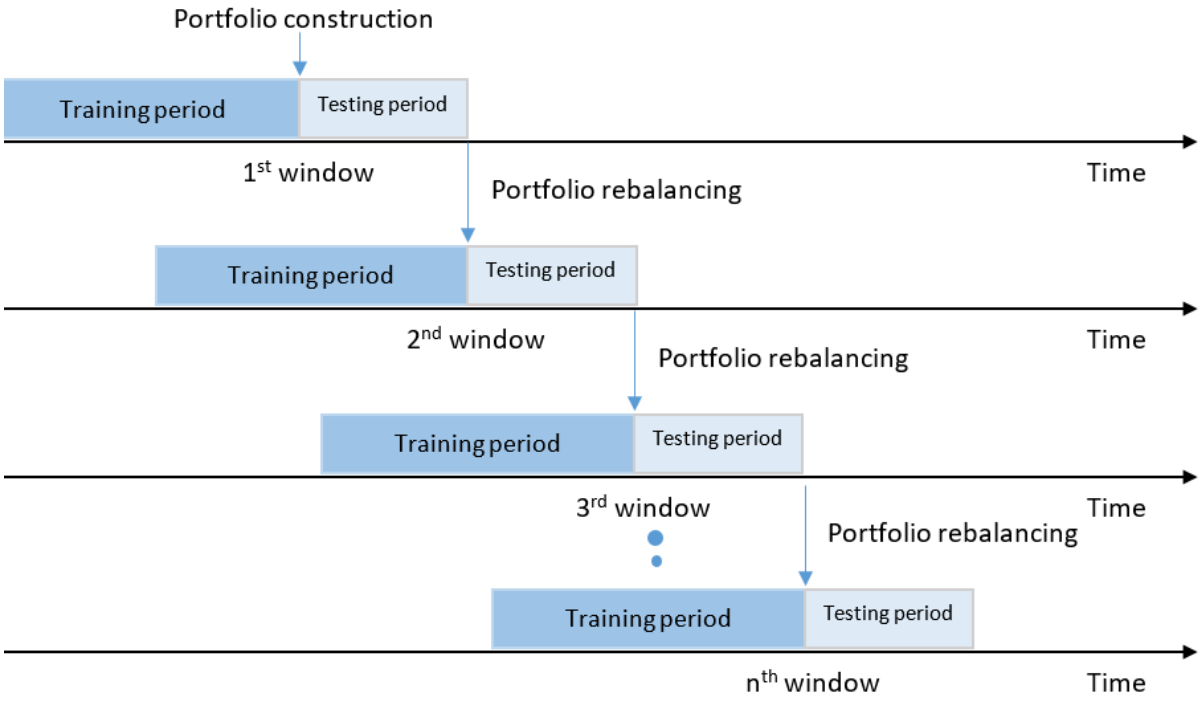

Fig. 3.3 Sliding window technique

divided in two parts : 1. Training period 2. Testing period. In training period of each window, k-means algorithm is applied for various values of ks. The k for which portfolio has maximum profit is selected. And the k selected in training period is used during testing period. Similarly, values for other hyper parameters are determined in training period and used in testing period. The weight of stocks are optimized by genetic algorithm during training and testing period. After end of every testing period, portfolio is rebalanced and window containing training and testing period is shifted towards right side by testing period. This process is repeated till whole dataset timeline is completed. Length of training period is adjusted to examine the performance of overall method and make it better. The full algorithm is given as below:

Algorithm 1 Clustering-based portfolio optimization scheme

```
1: procedure PORTFOLIO REBALANCE()
2:     while sliding window do
3:         Training period:
4:         for k varies do
5:             for Investor type varies do                    ▷ Training period
6:                 k-means clustering by changing k
7:                 Construct a portfolio with the highest mean TVP value cluster for given inv. type
8:                 GA with various fitness functions
9:                 Comparing portfolio returns in accordance with k
10:                Find k, investor type, fitness function which maximizes the return of portfolio
11:            end for
12:        end for
13:        for k, inv. type, fitness func. selected at training period do  ▷ Testing period
14:            k-means clustering with chosen k
15:            Construct a portfolio with the highest mean TVP value cluster for selected inv. type
16:            GA with selected fitness function
17:        end for
18:        return Portfolio's return
19:    end while
20: end procedure
```

Chapter 4

Empirical results and conclusion

We have considered top 100 companies of National Stocks Exchange according by their market capitalization as on 31^{st} March, 2018. We have used NIFTY50 index as standard asset and want to achieve greater return than it. The implementation code is given here.

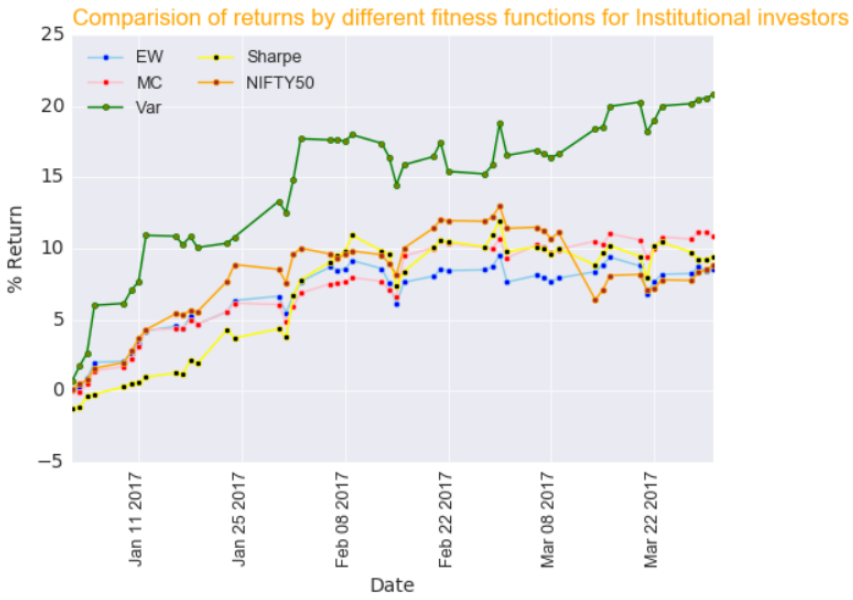

Fig. 4.1 Comparison of performance by various fitness function for institutional investors

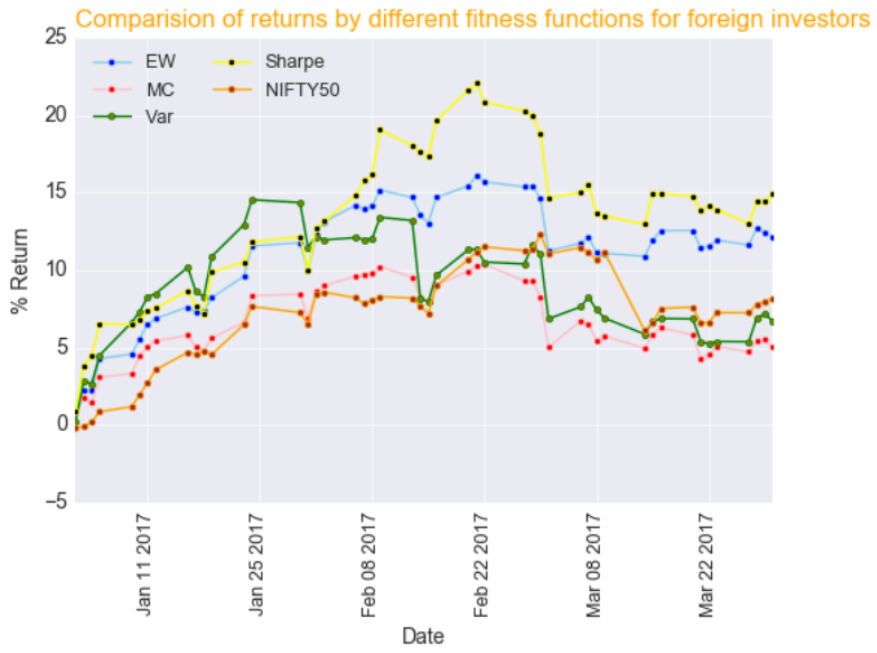

Fig. 4.2 Comparison of performance by various fitness function for foreign investors

The training period is kept six months while testing period, for evaluation of performance is fixed at one week. Inside training period, to evaluate hyper-parameters, we have again taken six months training and one week of testing performance, and the hyper-parameters who yields highest returns are selected for building portfolios in testing period.

The hyper-parameters for genetic algorithm, affects scope of search during evolutionary process. Population size was kept 50, crossover rate was kept 0.5 and mutation rate was kept 0.5/(length of chromosome) . Total number of generations was taken as 500. Risk free rate for calculation of Sharpe ratio is taken to be 6.50 %. We have neglected transaction cost here as they are broker dependent and they are constant value, so we can neglect them, if capital amount is high enough.

For institutional investors, portfolio formed with minimum variance fitness gave highest returns as shown in (Fig. 4.1) while for foreign investors, portfolio formed with

maximum Sharpe ratio fitness gave highest returns as shown in (Fig. 4.2).

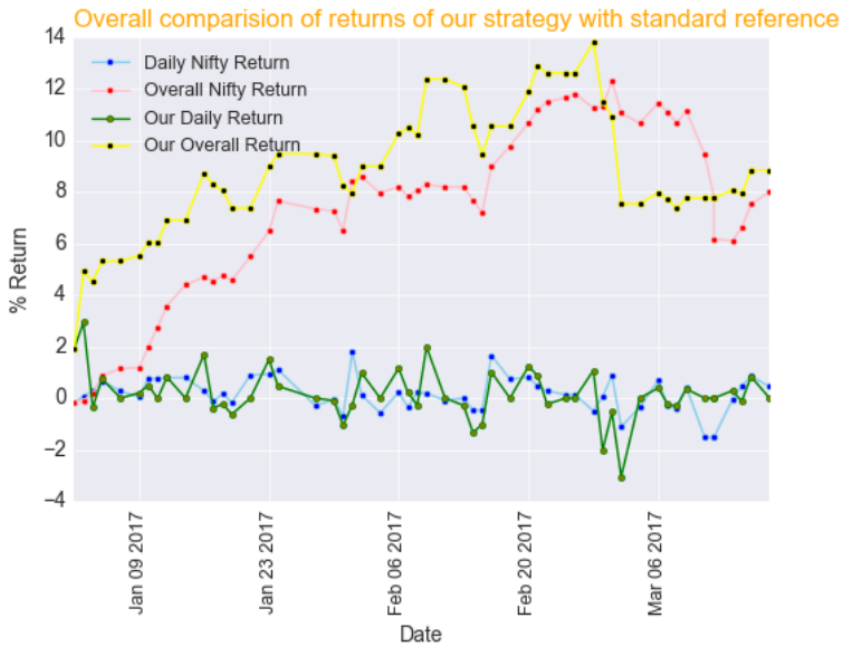

Fig. 4.3 Results of proposed algorithm

The results of algorithm simulated in first quarter of 2017 can be seen in (Fig. 4.3). It can be seen that for majority of time, the proposed strategy works better than standard index. This project shows that clustering based on portfolio optimization technique using investor trading volume information. Proposed model works well on long term (6 months) training on trading volume data of institutional and foreign investors there by yielding high return portfolios.

References

[BB04] Y.J. Liu T. Odean B.M. Barber, Y.T. Lee. Who gains from trade? evidence from taiwan, working paper. 2004.

[LAY09] S.-Y. Wang L.-A. Yu. Kernel principal component clustering methodology for stock categorization. *System Engineering-Theory Pract.*, 29:1–8, 2009.

[MG00] M. Keloharju M. Grinblatt. The investment behavior and performance of various investor types. *Journal of Financial Economics*, 55:43–47, 2000.

[YO03] G. Yamazaki Y. Orito, H. Yamamoto. Index fund selections with genetic algorithms and heuristic classifications. *Comput. Ind. Eng.*, 45:97–109, 2003.