A Hybrid Machine Learning and Dynamic Nonlinear Framework for Determination of Optimum Portfolio Structure

Sayan Gupta, Gautam Bandyopadhyay, Sanjib Biswas and Arun Upadhyay

Abstract Capital market investment is a growing stream of the economic literature. It has been a prime concern of a large number of investors belonging to different clusters or income groups for two reasons mainly. First, the construction of a portfolio, which deals with the selection of the stocks. Second, the formulation of an appropriate investment strategy, which calls for minimizing the risk while maximization of the return, i.e., optimization of the constructed portfolio. Following the broad framework as suggested in the seminal work of Markowitz [1], this research attempts to address the issue of portfolio optimization based on risk and return parameters while dynamically allocating the weights to the constituent stocks. In the first part of this study, k-means clustering is applied to a heterogeneous sample of 53 number of stocks enlisted with the NSE during the year 2012–2017. The purpose is to classify the stocks in three categories (such as low stock price, medium stock price, and high stock price) based on their monthly closing return. In the second phase, this study focuses on finding out the distribution of weights among the stocks belonging to the portfolio by using the generalized *reduced gradient* (GRG) method under the dynamic environment. Finally, this study attempts to validate the results by applying perception mapping. We have found eight stocks in the cluster of low stock price which is the sample studied in this research. We have observed that dynamic allocation of weights led to minimization of risk and the finding is validated through a perceptual map.

Keywords Portfolio · Nonlinear optimization · K-means clustering · Generalized reduced gradient method

S. Gupta (✉) · G. Bandyopadhyay
Department of Management Studies, NIT Durgapur, Durgapur, India
e-mail: sg.17ms1101@phd.nitdgp.ac.in

G. Bandyopadhyay
e-mail: gautam.bandyopadhyay@dms.nitdgp.ac.in

S. Biswas
Calcutta Business School, Calcutta, India

A. Upadhyay
NSHM Business School, Durgapur, India

1   Introduction

The construction of the optimal portfolio plays a critical role behind any investment decision as it is premised on the concept of maximization of the return on investment. It is imperative to contemplate on the selection of appropriate stocks in the right proportion for constructing a portfolio with an objective to increase return while minimizing risk. The essence of effective portfolio management lies in optimum utilization of the capital in terms of allocation of the stocks constituting the portfolio and balancing the same taking risk-return trade-off into account, since risk tolerance level and expectation of return vary from investor to investor. Stated in this field, Markowitz's mean–variance approach [1] is considered to be the first systematic attempt to formulate a diversified investment decision with an objective of ensuring high return at low-risk level [2, 3]. In tune with this work, Tobin [4] postulated the concept of *efficient frontier* and *capital market line*, wherein the author argued for the persistence of the portfolio structure irrespective of risk tolerance level at a consistent expectation level; only relative proportions of stocks change otherwise. Extending the work, Markowitz [5] introduced expected return–semi-variance based analysis which considers both the extremes while determining the efficiency of the portfolio. Sharpe [6, 7] further worked on Markowitz's analysis and noted that optimal portfolio (among all efficient portfolios) depends on expected return and risk preference of the investors. These classical approaches paved the way to the *modern portfolio theory* (MPT). MPT aims to maximize the expected return of the portfolio at a given risk level or minimizing portfolio risk at a given return by optimally allocating the total available fund to different assets [8]. However, while forming the portfolio, it is essential to consider the influence of the assets on each other, i.e., only on the basis of individual stock performance one cannot form a portfolio [9]. In effect, the distribution of the returns decides effective optimization of the portfolio. In the context of portfolio management, a risk is perceived as the total risk of a portfolio which has two components; *systematic risk* or *market risk* and *unsystematic risk* or *diversifiable risk*. By prudent stock selection and distribution of appropriate weightage for them within a portfolio, unsystematic risk can be reduced to a considerable extent [10].

In line with the seminal work of Markowitz [1], over the years, several studies have been made on portfolio selection and optimization. There has been a growing stream of alternative methods suggested by several researchers and practitioners in the stated field. Examples include artificial neural network [11], genetic algorithm [12], particle swarm optimization [13], simulated annealing [14], ensemble [15], decision tree [16], clustering [17], multi-criteria decision-making approach [18] to name a few. It is evident from these researches that in order to optimize the portfolio and correctly predict its return, selection of the right number of right stocks in the right proportion is of paramount importance. There has been a plethora of research conducted on portfolio selection using unsupervised learning methods like clustering. Clustering finds its importance in pattern identification, classification, and detection of an anomaly while selecting stocks to form a portfolio. Further, it is useful in finding interrelationship or co-movements of the stocks [19]. Although hierarchical

clustering is a dominant method in this regard, k-means clustering and C-means clustering have also drawn significant attraction of the researchers [20].

In this study, the authors have used k-means clustering for segregating 53 heterogeneous companies with different sectors enlisted with the NSE during the year 2012–2017 on the basis of monthly stock return (60 months) into three distinct clusters; *low stock price* (LSP), *mid stock price* (MSP), and *high stock price* (HSP) in order to construct a portfolio. Further, GRG method has been used to optimize the constructed portfolio of the stocks belonging to the MSP cluster through dynamic allocation of the weights to different stocks with an objective to minimize risk while maximizing return. The rest of this paper is organized as follows. In Sect. 2, data and methodology are discussed. Section 3 presents the findings. Finally, Sect. 4 concludes the paper while highlighting some of the implications and future scope.

2 Data and Methodology

The broad objective of this study is to find out the distribution of the stocks under the portfolio ensuring maximum return at minimum risk. It requires appropriate weight assignment to individual stocks forming the portfolio.

2.1 Sample

The sample for this study consists of 53 numbers of heterogeneous companies with different sectors enlisted with the National Stock Exchange (NSE), India through convenience sampling (refer Table 1). The study period is the year 2012–2017. The data were collected from the published secondary database. Monthly returns of those stocks for 60 months have been considered for analysis. In order to calculate monthly returns in case of the Index or Stock (since the data collected is month wise), we have used the formula $Ln(P_1/P_0)$, $Ln(P_2/P_1)$ and so on since the behavior of data is continuous. After clustering, we have worked on MSP cluster.

2.2 Methods

In this study, a three-stage approach has been followed. First, k-means clustering, an unsupervised learning technique has been applied in order to classify the stocks based on the monthly returns for constructing the portfolio; second, GRG, a nonlinear optimization technique has been selected for deciding weights for the stocks with an objective to minimize risk while maximizing return. Finally, we have applied perception mapping using mean and standard deviation of the monthly returns of the stocks belong to MSP.

Table 1 Stocks under the study (primary level)

Sector(s)	Companies	Total
Automobile	AMARA RAJA BATT; MRF; TVS MOTOR; TATA MOTORS; HERO MOTOCORP; BAJAJ AUTO; MARUTI SUZUKI; BHARAT FORGE; M&M; ASHOK LEYLAND	10
Banking	CANARA BANK; SBI; BANK OF BARODA; HDFC BANK; YES BANK; INDUSIND BANK; PNB; FEDERAL BANK	8
Oil and natural gas	IOCL; GAIL; RIL; NTPC; ONGC; BPCL	6
FMCG sector	ITC; HUL; GODREJ CP; MARICO; COLPAL; DABUR; P&G; GSK CP; EMAMI LTD; GODREJ IND	10
Food and beverages	MCDOWELL; TATA GLOBAL; UBL; UB (H) L; BRITANNIA; JUBILANT FOODWORKS	6
IT/ITes	KPIT TECH; HCL TECH; INFOSYS; WIPRO; TCS; TECH MAHINDRA; MINDTREE; OFSS	8
Others	TATA ELXSI; RIL INFRA; HIND PETRO (HPCL); POWER GRID CORP; TATA POWER	5
Grand total		53

The expected return on a portfolio is computed as follows:

$$E(R_p) = \sum_{i=1}^{N} w_i E(R_i) \tag{1}$$

where

$E(R_p)$ the expected return on the portfolio
N   the number of Index or stocks in the portfolio
w_i the proportion of the portfolio invested in Index or ith Stock
$E(R_i)$ the expected return on ith Stock.

The risk is calculated as follows:

$$\sigma_p^2 = (w_A)^2 \sigma_A^2 + (w_B)^2 \sigma_B^2 + 2\, w_A w_B\, \sigma_{A,B} \tag{2}$$

where σ_p: Standard deviation of the portfolio; σ_A and σ_B: Standard deviation of the stocks A and B; w_A and w_B: Weights assigned to the stocks A and B.

K-means Clustering. It is a type of unsupervised learning, which distribute the unlabeled data into a specific number of groups or clusters (represented by the variable K) on the basis of feature similarity. Each such cluster is represented by its centroid which is a collection of feature values pertaining to that cluster. This algorithm takes a set of "m" number of data points in "n" dimensions into "K" number of clusters through an iterative process. The objective is to minimize within-cluster sum of squares [21]. It starts with an initial estimate of clusters wherein a

particular data point x_j(where, $j = 1, 2, \ldots m$) is allocated to a particular cluster c_i(where, $i = 1, 2, \ldots k$) depending on the squared Euclidean distance between the data point and the centroid of the respective cluster. The points which find their distances with respect to the centroid of a particular cluster minimum are assigned to that particular cluster. Then, the process recomputes all the centroids and again finds the distances of all the data points with respect to each newly constructed centroids for assigning the data points to newly formed clusters. This process repeats itself until limiting condition is reached, i.e., either there is no change of clusters by the data points (i.e., within a particular cluster all data points are homogeneous) or within group sum of squares are minimized or the prefixed maximum number of iterations is conducted.

Generalized Reduced Gradient Method (GRG). This is a general version of the reduced gradient method, which solves optimization problems with nonlinear constraints and arbitrary bounds. The method is described as follows as explained by Lasdon et al. [22].

The structure of general nonlinear problem is given by

$$\text{Minimize } f_0(X) \tag{3}$$

$$\text{subject to the constraints } g_i(X) = 0; \text{ where, } i = 1, 2, \ldots m \tag{4}$$

$$\text{Where, and } l_j \leq X_j \leq u_j; \text{ where, } j = 1, 2, \ldots n \tag{5}$$

Here, u_j and l_j indicate the upper boundary and lower boundary, respectively, where, $u_j > l_j$ assuming m < n in order to avoid infeasibility of the solution or unique solution. The above forms are general since inequality constraints get transformed to equalities by adding slack variables. Following this, for solving the problem, basic variables (m) get expressed in terms of remaining nonbasic (n-m) variables. If $\bar{X}$ denotes a feasible solution point and Y be the vector representing basic variables and Z be the vector of nonbasic variables on X, then after partitioning, Eqs. (4) and (5) can be written as

$$X = (Y, Z), \bar{X} = (\bar{Y}, \bar{Z}) \tag{6}$$

$$g_i(Y, Z) = 0 \tag{7}$$

Here, it is assumed that both the objective and constraint functions are differentiable. The transformed objective function is given as

$$F(Z) = f(Y(Z), Z) \tag{8}$$

Accordingly, the nonlinear problem is transformed at least for Z close to $\bar{Z}$, to a reduced problem given by

$$\text{Minimize F(Z)} \tag{9}$$

$$\text{Subject to } l' \leq Z \leq u' \tag{10}$$

The GRG method actually solves the original problem as stated above by solving Eqs. (9) and (10). This study addresses a nonlinear problem since in our study, it has been observed that the monthly rate of return is having quadratic and cubic nature. Hence, the data suits the applicability of GRG method in our case.

3 Results and Discussions

Table 2 describes the distribution of the stocks (Table 1) into three defined clusters such as LSP, MSP, and HSP.

In our study, we have selected MSP for further analysis since in LSP cluster we have 44 companies which make the portfolio too stretched for a common investor. Also, HSP cluster shows opposite nature and therefore, it has not been considered. Also, LSP signifies comparatively prematured or poor performance at the market and HSP indicates a bit saturated performance and comparatively less growth. We have considered to form a portfolio based on the stocks belonging to the MSP cluster. Table 3 lists out the stocks under study along with their monthly expected rate of return and standard deviations.

Table 2 Number of companies in each cluster

Cluster	No. of companies
LSP	44
MSP	8
HSP	1
Total	53

Table 3 Stocks under MSP cluster

	Hero Moto-corp	Bajaj Auto	Maruti Suzuki	Britannia	P&G	GSK CP	TCS	OFSS
Monthly expected rate of return (AVROR)	0.01148	0.00746	0.03127	0.03742	0.02045	0.00902	0.01276	0.00364
Standard devia-tion (SD)	0.06651	0.06602	0.08861	0.07129	0.05079	0.07688	0.06413	0.05535

Table 4 Normality test

	Hero Moto-corp	Bajaj Auto	Maruti Suzuki	Britannia	P&G	GSK CP	TCS	OFSS
Kolmogorov-Smirnov Z	0.760	0.427	0.762	0.461	0.510	0.986	0.740	0.430
Asymp. Sig. (two-tailed)	0.610	0.993	0.607	0.984	0.957	0.285	0.644	0.993

Table 5 Correlation matrix

	Hero Moto-corp	Bajaj Auto	Maruti Suzuki	Britannia	P&G	GSK CP	TCS	OFSS
Hero Moto-corp	1							
Bajaj Auto	0.541032	1						
Maruti Suzuki	0.371683	0.420821	1					
Britannia	0.251049	0.325695	0.203601	1				
P&G	0.016686	0.066044	0.287807	0.271632	1			
GSK CP	0.076709	0.056452	0.244738	0.50983	0.374539	1		
TCS	0.207421	−0.00987	−0.16676	0.082849	0.04115	0.185365	1	
OFSS	0.210708	0.342858	0.312976	0.184763	0.171558	0.129388	0.257402	1

Further, we have performed a normality test (refer Table 4) in order to comply with the conditions of Markowitz [1]. It is seen from the result that the rate of returns of the stocks satisfies normality condition.

In order to understand whether diversification is possible among the stocks, we have conducted a correlation analysis (refer Table 5). The determinant value (0.16548) obtained from the correlation matrix suggests that the stocks under the portfolio can be diversified. Hence, it is a problem of dynamic allocation of weights among the stocks.

Table 6 shows the variance (diagonal values) and covariance (off-diagonal values) analysis for determining risk.

Table 7 shows the weights of the stocks calculated dynamically using GRG method. The optimum risk of the portfolio under study is 0.11% as calculated through the GRG method.
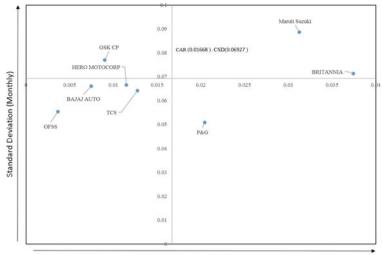
For validation purpose, we have performed perception mapping (refer Fig. 1). Table 8 shows the combined values of AVROR and SD. These two values are calculated as follows:

Table 6 Variance and covariance table

	Hero Motocorp	Bajaj Auto	Maruti Suzuki	Britannia	P&G	GSK CP	TCS	OFSS
Hero Motocorp	0.004350	0.002336	0.002154	0.001171	0.000055	0.000386	0.000870	0.000763
Bajaj Auto	0.002336	0.004286	0.002421	0.001507	0.000218	0.000282	−0.000041	0.001232
Maruti Suzuki	0.002154	0.002421	0.007721	0.001265	0.001274	0.001640	−0.000932	0.001509
Britannia	0.001171	0.001507	0.001265	0.004997	0.000967	0.002748	0.000372	0.000717
P&G	0.000055	0.000218	0.001274	0.000967	0.002537	0.001438	0.000132	0.000474
GSK CP	0.000386	0.000282	0.001640	0.002748	0.001438	0.005813	0.000899	0.000541
TCS	0.000870	−0.000041	−0.000932	0.000372	0.000132	0.000899	0.004044	0.000898
OFSS	0.000763	0.001232	0.001509	0.000717	0.000474	0.000541	0.000898	0.003013

Table 7 Distribution of the weights to the stocks under the portfolio

Companies	Hero Motocorp	Bajaj Auto	Maruti Suzuki	Britannia	P&G	GSK CP	TCS	OFSS	Total
Weights (%)	10.3	12.3	0.0	3.9	35.2	2.6	19.5	16.2	100.0

Fig. 1 Perceptual map

Table 8 Combined AVROR and SD

CAR	0.0166888
CSD	0.0692789

$$\text{Combined AVROR(CAR)} = \text{Average(AVROR)} \qquad (11)$$

$$\text{Combined SD(CSD)} = \frac{\sqrt{\sum\left(d_i^2 + SD_i^2\right)}}{n}; \qquad (12)$$

where

d_i Average return for ith Stock-CAR; i = 1, 2, …n.

Further, we have used the values of CAR and CSD to construct a perceptual map, wherein the axes are shifted from the origin to the point having CAR and CSD values. This results into generation of four new quadrants representing (High Return, High SD), (High SD, Low Return), (Low SD, High Return), and (Low SD, Low Return), respectively.

It is seen from the above figure that only P&G falls in the third quadrant, which characterizes high growth and low risk. It suggests that out of the stocks considered for the study, P&G stands alone, matched with the stated requirements. Therefore, it is logical to state that the earlier findings are being validated by the perceptual map.

4 Conclusion

In this study, we have attempted to assign optimal weights to the individual stocks constituting a portfolio in a dynamic environment. In order to select a portfolio, we have applied k-means clustering on a heterogeneous set of stocks listed in the NSE which are selected based on convenient sampling. Next, we have used a nonlinear optimization technique such as GRG for allocating weights with an objective to minimize the risk while maximizing the return. Allocation of the weights to the individual stocks forming the portfolio is having a significant impact on return on investment from the same. Further, in order to validate the results, a perception mapping of stocks under study has been performed which commensurate the earlier findings. This study is unique in the sense that initially, it started with unsupervised learning, but in the process, it came up with the bridging effect of unsupervised learning to a supervised learning validating all the way. This study may further be extended to analyze the clusters in the same way other than MSP for investigating any relation with the present one. Further, sector-wise portfolio performance may also be assessed for a comparative analysis in a dynamic and uncertain environment.

References

1. Markowitz H (1952) Portfolio selection. J Fin 7(1):77–91
2. Steinbach MC (2001) Markowitz revisited: mean-variance models in financial portfolio analysis. SIAM Rev 43(1):31–85
3. Rubinstein M (2002) Markowitz's "portfolio selection": a fifty-year retrospective. J Fin 57(3):1041–1045
4. Tobin J (1958) Liquidity preference as behavior towards risk. Rev Econ Stud 25(2):65–86
5. Markowitz H (1959) Portfolio selection. Yale University Press, New Haven, Connecticut
6. Sharpe WF (1963) A simplified model for portfolio analysis. Manage Sci 9(2):277–293
7. Sharpe WF (1966) Mutual fund performance. J Bus 39(1):119–138
8. Fabozzi FJ, Gupta F, Markowitz HM (2002) The legacy of modern portfolio theory. J Invest 11(3):7–22
9. Elton EJ, Gruber MJ, Brown SJ, Goetzmann WN (2009) Modern portfolio theory and investment analysis, 6th edn. Wiley
10. McClure B (2017) Modern portfolio theory: why it's still hip. Investopedia. http://www.investopedia.com/articles/06/MPT.asp#axzz1g3JQY7nY. Accessed 21 June 2018
11. Nazemi A, Abbasi B, Omidi F (2015) Solving portfolio selection models with uncertain returns using an artificial neural network scheme. Appl Intel 42:609–621. https://doi.org/10.1007/s10489-014-0616-z
12. Chang TJ, Yang SC, Chang KJ (2009) Portfolio optimization problems in different risk measures using genetic algorithm. Expert Syst Appl 36(7):10529–10537
13. Zaheer KB, Aziz MIBA, Kashif AN, Raza SMM (2018) Two stage portfolio selection and optimization model with the hybrid particle swarm optimization. Matematika 34(1):125–141
14. Crama Y, Schyns M (2003) Simulated annealing for complex portfolio selection problems. Eur J Oper Res 150(3):546–571
15. Nagy G, Barta G, Henk T (2015) Portfolio optimization using local linear regression ensembles in RapidMiner. arXiv:1506.08690
16. Tucker CS, Kim HM (2009) Data-driven decision tree classification for product portfolio design optimization. J Comput Inf Sci Eng 9(4):041004

17. Nanda SR, Mahanty B, Tiwari MK (2010) Clustering Indian stock market data for portfolio management. Expert Syst Appl 37(12):8793–8798
18. Ehrgott M, Klamroth K, Schwehm C (2004) An MCDM approach to portfolio optimization. Eur J Oper Res 155(3):752–770
19. Liao SH, Chou SY (2013) Data mining investigation of co-movements on the Taiwan and China stock markets for future investment portfolio. Expert Syst Appl 40:1542–1554. https://doi.org/10.1016/j.eswa.2012.08.075
20. Pai GV, Michel T (2009) Evolutionary optimization of constrained k-means clustered assets for diversification in small portfolios. IEEE Trans Evol Comput 13(5):1030–1053
21. Hartigan JA, Wong MA (1979) Algorithm AS 136: a k-means clustering algorithm. J R Stat Soc Series C (Applied Statistics) 28(1):100–108
22. Lasdon LS, Fox RL, Ratner MW (1974) Nonlinear optimization using the generalized reduced gradient method. Revue française d'automatique, informatique, recherche opérationnelle. Recherche opérationnelle 8(V3):73–103