CHAPTER 4
CHAPTER 4

NOVEL FRAMEWORK FOR THE PORTFOLIO DETERMINATION USING PSO ADOPTED CLUSTERING TECHNIQUE

4.1 Introduction

Decision making for choosing and Investment in the stocks are very critical and challenging task [49] because the stock market is a complex, non stationary, chaotic and non-linear dynamic system. Portfolio management is a key issue to be considered when one wants to increase their assets by using the stock market. The pain of managing the portfolio by the investor paved the way of research in this area. The dynamic nature of the stock market makes this problem an interesting and challenging one. Portfolio is the basket consisting of the stocks, securities and the bonds related to the asset creation held by the particular investor for the game to play and win money from the market.

This chapter deals with the methodology for the portfolio determination using the PSO adopted K-Means algorithm for the creation of the clusters from the stock data considered. This is the important and novel approach used in the paper. The seeding of K-Means is an important problem in the research. In this paper attempts have been made to solve this issue by using the PSO. Then the NARX based algorithm is
used for the calculation of the expected returns of the stocks. Then the modern portfolio theory by the Markowitz model has been used by the minimization of the covarience of the stocks to build the portfolio. Portfolio created by K-Means and PSO adopted K-Means has been demonstrated and compared with the Nifty returns.

4.2 Related works

This section deals with the related works of the proposed methodology. The clustering algorithm is used as the back bone of the method. It is adapted with the particle swarm optimization to improve its effectiveness. The PSO theory and the basic working model of the PSO can be explained as:

4.2.1 Particle Swarm Optimization

Particle Swarm Optimization is the population based algorithm inspired by the social behaviour of birds. It is the algorithm given by Eberhart and Kennedy [50-52]. It is based on the social behaviours of birds flocking or fish schooling. The major advantage of this PSO is it doesn’t use the gradient of the problem to be optimized; hence it could be used for the variety of the optimization problem [53]. Particle swarm optimization is a global optimization algorithm for dealing with problems in which a best solution can be represented as a point or surface in an N–dimensional space. Hypothesis are plotted in this space and seeded with an initial velocity, as well as communication channel between the
particles. Particles then move through the solution space and are evaluated according to some fitness function after each timestamp. Over time particles are accelerated towards those particles within their grouping which have better fitness values [54].

PSO performs a population-based search, using particles to represent potential solutions within the search space. Each particle is characterized by its position, velocity, and a record of its past performance. Particles are influenced by their leaders, which are the best performers either from the entire swarm or their neighborhood. At each flight cycle, the objective function is evaluated for each particle, with respect to its current position, and that information is used to measure the quality of the particle and to determine the leader in the sub-swarms and the entire population [55].

4.2.1.1 Pseudo code for PSO

The pseudo code of the PSO algorithm as stated in [56] as given below

FOR each particle

 initialize particle

END

Do

For each particle

 Calculate fitness value

 If the fitness value is better than its personal best
set current value as the new pBest

End

Choose the particle with the best fitness value of all as gBest

For each particle

Calculate particle velocity

Update particle position

WHILE maximum iterations or minimum error criteria is not attained

The particle velocity is calculated based on the following formula

\[ v[] = v[] + c1 \times \text{rand()} \times (pbest[] - \text{present[]}) + c2 \times \text{rand()} \times (gbest[] - \text{present[]}) \]

The particle position is given by

\[ \text{present[]} = \text{present[]} + v[] \]

where, \( v[] \) is the particle velocity, \( \text{present[]} \) is the current particle (solution). \( \text{pbest[]} \) and \( \text{gbest[]} \) are defined as stated before. \( \text{rand()} \) is a random number between (0,1). \( c1, c2 \) are learning factors (usually \( c1 = c2 = 2 \)).

PSO is initialized with a group of random particles (solutions) and then searches for optima by updating generations. At every iteration, each particle is updated by following two "best" values. The first one is the best solution (fitness) it has achieved so far. (The fitness value is also stored.) This value is called pbest. Another "best" value that is tracked by the particle swarm optimizer is the best value, obtained so far by any
particle in the population. This best value is a global best and called as gbest. When a particle takes part of the population as its topological neighbours, the best value is a local best and is called lbest.

The features that attract towards PSO is its ease of implementation, fewer parameters to adjust, each particle remembers its own previous best as well as neighborhood best and it is efficient in maintaining diversity.

4.2.2 Cluster analysis of a dataset

Due to the stock market volatility, needs to be captured from hidden information in the database, forecasting stock market movement for intelligent decision making is a quite difficult task [57]. Data mining is being actively applied to stock market since 1980s. The various aspects of stock market to which data mining has been applied include predicting stock indices, predicting stock prices, portfolio management, portfolio risk management, trend detection, designing recommender systems etc [58].

Clustering is an adaptive procedure in which objects are clustered or grouped together, based on the principle of maximizing the intra-class similarity and minimizing the inter-class similarity [59]. Clustering is an unsupervised process i.e., it analyses the data set without the knowledge of the labels under which it must be grouped. Clustering is used to divide a data set into classes (by generating labels for them) using the principle of maximizing the intra class similarity. Within the data set clusters are
formed so that objects which are similar are grouped together and objects that are very different fall into other clusters.

Cluster analysis is the automatic identification of groups of similar objects or patterns. For example, if a set of data denoted by \( x \), is very similar to a few other sets of data, we may intuitively tend to group \( x \) and these sets of data into a natural cluster. By maximizing inter group similarity and minimizing intra group similarity; a number of clusters would form on the measurement/observation space. It becomes easy to recognize and assign to the clusters suitable label or feature description. There are generally two types of learning approaches relevant to cluster analysis. The parametric partitional approach attempts to cluster the set directly, in a manner that depends on a set of parameters. These parameters are then adjusted to optimally satisfy a chosen criterion of separation and compactness of clusters. Whereas, the nonparametric approach hierarchical approach proceeds from a provisional initial clustering and iteratively merges/or split clusters until a required degree of similarity holds for the elements of the clusters [58].

The basic Clustering is usually considered to be the problem of partitioning a single set of unlabeled points. Clustering algorithms may be categorized by how they form groups of clusters. Hierarchical algorithms work on either successive splitting (divisive) or merging (agglomerative) of groups to form a hierarchy of clusters based on a specified measure of
distance or similarity between objects. Alternatively, partitioning algorithms search for a partition of the data that optimizes a global measure of quality for the groups, usually based on distance between objects.

One of the most common iterative algorithms is the K-means algorithm [60], broadly used for its simplicity of implementation and convergence speed. K-means also produces relatively high quality clusters considering the low level of computation required. K-means is one of the simplest algorithms known to perform well with many data sets, but its good performance is limited mainly to compact groups. When the points are drawn from a mixture of Gaussian distributions, the K-means algorithm is a gradient descent algorithm that minimizes the quantization error [61]. As with many gradient descent algorithms, one drawback of K-means is that it can reach a local minimum of the objective function instead of the desired global minimum, meaning that convergence is reached but the solution is not optimal.

### 4.2.3 Clustering Using PSO

In the context of pattern recognition theory, each object is represented by a vector of features, called a pattern. Clustering can be defined as the process of partitioning a collection of vectors into subgroups whose members are similar relative to some distance measure. A clustering algorithm receives a set of vectors, and groups them based
on a cost criterion or some other optimization rule. The related field of pattern classification, which involves simply assigning individual vectors to classes, has developed a theory based on defining error criteria, designing optimal classifiers, and learning. In comparison, clustering has historically been approached heuristically; there has been almost no consideration of learning or optimization, and error estimation has been handled indirectly via validation indices [62].

The main drawback of the K-means algorithm is that the result is sensitive to the selection of the initial cluster centroids and may converge to the local optima [63]. Therefore, the initial selection of the cluster centroids affects the main processing of the K-means and the partition result of the dataset as well [64]. The processing of K-means is to search the local optimal solution in the vicinity of the initial solution and to refine the partition result. The same initial cluster centroids in a dataset will always generate the same cluster results. However, if good initial clustering centroids can be obtained using any other techniques, the K-means would work well in refining the clustering centroids to find the optimal clustering centers [65].

4.3 Problem formulation

Decision making in investment is a challenging job. The systematic way to make an investment and to keep track of the investment to yield a positive output is through the proper asset allocation and appropriate
portfolio management. A formal model for an efficient portfolio was given by Markowitz [66]. This model is based on return of an asset is its mean return and the risk of an asset is the standard deviation of the asset returns. It aims to find an optimal allocation of capital among a set of assets by simultaneously minimizing the risk and maximizing the return of the investment. Risk was quantified in such a way that investors could analyze risk return choices. Moreover, risk quantification enabled investors to measure risk reduction generated by diversification of investment. So diversification of investment is essential to create an efficient portfolio [67].

To create the efficient portfolio and manage it, we need to have

i) Diversification - helps to identify the stock from various or less correlated stocks.

ii) More stable portfolio - the insensitivity towards the daily changes of the market. This is an ideal situation but it more or less could be achieved by the deployment of new theories.

iii) less risk – to condense the possibility that actual future returns will be different from expected return, to reduce the volatility and to reduce the standard deviation from the asset returns.

iv) Rely on correlations between the stock to overcome the dependency on expected returns as the expected returns are tough to predict in the dynamic market situations.
Thus the problem is formulated in such a way that the methodology must be able to find the diversified and more stable portfolio with less risk and results with more returns.

4.4 Proposed Methodology

The above problem could be encountered in two stages. The first stage concentrates on the diversification of the portfolio thus reducing the risk. The second stage is creation of the efficient portfolio. The proposed methodology could be depicted with the following frame work.

**Stage 1:** In this stage the stock data to be considered is taken. The data is applied with the PSO algorithm to determine the centroid for the forthcoming clustering algorithm. This is the scrutinizing step where the following clustering algorithm will be seeded with the initial centroid. The clustering algorithm here adopted is the K-means algorithm, which is a partitioning based clustering algorithm. This PSO adopted clustering algorithm is used to find the Clusters in the initial stock data provided.

The clustering process aims at least diversity within a group and find most difference among groups is to be reached. The K-means algorithm is used for the clustering purpose since the K-means clustering algorithm offers a good compactness compared to other clustering techniques such as Self organizing maps and Fuzzy K-means [1]. But the K-means algorithm suffer from the problem of fixing the initial centroid.
In order to rectify this limitation, this research uses the PSO to fix the centroids.

The ability of globalized searching of the PSO algorithm and the fast convergence of the K-means algorithm are combined. The PSO algorithm is used at the initial stage to help discovering the vicinity of the optimal solution by a global search. The result from PSO is used as the initial seed of the K-means algorithm.

**Figure 4.1: Framework of the proposed work**
Stage 2: In this stage the Markowitz based selection for the efficient portfolio is built. The Markowitz model measures the risk reduction generated by the diversification of investment. The return of the portfolio is the weighted return of the stocks present. According to Markowitz, if an investor holds a portfolio of two assets he or she can reduce portfolio risk below the average risk attached to the individual assets. This can be achieved by investing in assets that have low positive correlation, or better still, a negative correlation.

4.5 Proposed Algorithm

The proposed algorithm for the efficient portfolio determination is presented as in the pseudo code format.

Stage 1(a) /* PSO to fix the initial seed of the K-means algorithm */

Input: Stock data

BEGIN

Each particle randomly chooses k numbers of vectors from the stock data as the cluster centroid vectors.

FOR EACH

Compute pBest and the gBest

Calculate particle velocity

Update particle position

END FOR
Repeat until maximum iterations or a minimum error criterion is not attained.

**END**

**OUTPUT** : K Centroids

**Stage 1(b) /* PSO Adopted K-means algorithm */**

**INPUT**: Stock data, Initial centroids by PSO

**BEGIN**

Make initial partition of objects into K clusters by assigning objects to closest K centroids given by PSO

Calculate the centroid(mean) of each of the K clusters.

i) For object i, Calculate its distance to each of the centroids.

ii) Allocate object i to cluster with closest centroid.

iii) If object was reallocated, recalculate centroids based on new clusters.

Repeat Until for object i= 1 to N

Repeat until no reallocations occur

**END**

**OUTPUT**: Clusters from Stock data

**Stage 2: /* determination of efficient portfolio through Markowitz model */**

**INPUT**: Clusters from Stock data
/* stage 2(a) – determination of expected return of each stock */

FOR each cluster

/* Calculation of return of the \( i^{\text{th}} \) stock, \( r_{it} \) is the anticipated return at time \( t \) per unit invested in security \( i \). 

\( r_{it} \) is decided based on the Nonlinear Autoregressive eXogenous (NARX) Network architecture in association with Levenberg Marquardt algorithm. */

Select the appropriate number of neurons, input delays and feedback delays in NARX.

Select the Levenberg Marquardt algorithm.

Select the appropriate performance function.

Perform the Training and Testing for all pairs of inputs.

Generate the Network Output \( Y \) until it is equivalent to Target \( T \).

If the network produced results for all pairs of input is good

Save it as \( r_{it} \)

Else

Retrain the network

END

// stage 2(b) Formulation of the portfolio

FOR every pair of stocks in a cluster

// Assignment of the weight to the stock
\[ w_i = \frac{\text{return value of the stock } i}{\text{total return value of portfolio}} \times 100 \]

// Calculation of expected return

\[ E(R_p) = \sum_{i} w_i E(r_{it}) \]

/* Where, \( E(R_p) \) is the expected return of the portfolio

\( w_i \) is the weight of the stock \( i \)

\( E(r_{it}) \) is the expected return of the stock \( I \) */

// Calculation of the covariance

\[ \sigma_p^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} COV_{ij} x_i x_j \]

/* where, \( \sigma_p^2 \) is the variance in portfolio return

\( COV_{ij} \) is the covariance matrix between stock \( i \) and \( j \)

\( x_i \) is the fraction of the portfolio devoted to stock \( i \)

\( n \) is the number of stocks */

END

FOR Each cluster

Find the minimum \( \sigma_p^2 \)

END

Construct the portfolio with minimum risk by the minimization of \( \sigma_p^2 \)

OUTPUT: Portfolio with the stocks and the weights associated.
4.6 Data description

The data employed for the proposed approach is the historical data that has been collected between the period of March 2010 to October 2010. The stocks from various sectors are collected to create a diversified portfolio. The various sector indices like financial, Healthcare, Basic materials, Automobiles were collected for the experiment purpose from the National Stock exchange.

4.7 Experimental details

The data were processed for the clustering process first using the K-means clustering, then the PSO adopted K-means clustering is applied for the data.

4.7.1 Metrics

The various metrics used for the validity of the cluster is adopted from [1] is given as follows:

**Silhouette index**: Better quality of a clustering is indicated by a larger Silhouette value.

**Davies–Bouldin index**: The lower the value the better the cluster structures.

**Calinski–Harabasz index**: It evaluates the clustering solution by looking at how similar the objects are within each cluster and how dissimilar the different clusters are. It is also called a pseudo F statistic.
Krzanowski–Lai index: Optimal clustering is indicated by maximum value.

Dunn’s index (DI): This index is proposed to use for the identification of "compact and well-separated clusters". Large values indicate the presence of compact and well-separated clusters.

Alternative Dunn index (ADI): The aim of modifying the original Dunn’s index was that the calculation becomes simpler, when the dissimilarity functions between two clusters.

4.7.2 Building of the portfolio

The efficient portfolio has been built based on the clusters formed in the stage 1 and in the stage 2, Markowitz model is used for building the portfolio. The stage 2 is using the NARX model for the estimation of the returns, the returns are then used to find the return of the portfolio. The variance is then calculated based on the weights associated and the covariance.

4.8 Results and discussions

The results obtained based on the various cluster metrics are displayed as shown in the table.
Table 4.1: The validity metric values based on the K-Means Clustering

<table>
<thead>
<tr>
<th>Cluster Metric</th>
<th>Number of clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Silhouette</td>
<td>0.375</td>
</tr>
<tr>
<td>Davies–Bouldin</td>
<td>1.042</td>
</tr>
<tr>
<td>Calinski–Harabasz</td>
<td>18.3</td>
</tr>
<tr>
<td>Krzanowski–Lai</td>
<td>1.119</td>
</tr>
<tr>
<td>Dunn’s Index</td>
<td>0.895</td>
</tr>
<tr>
<td>Alternative Dunn's</td>
<td>0.422</td>
</tr>
</tbody>
</table>
Table 4.2: The validity metric values based on the PSO adopted K-Means Clustering

<table>
<thead>
<tr>
<th>Cluster Metric</th>
<th>Number of clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Silhouette</td>
<td>0.462</td>
</tr>
<tr>
<td>Davies–Bouldin</td>
<td>0.987</td>
</tr>
<tr>
<td>Krzanowski–Lai</td>
<td>1.7138</td>
</tr>
<tr>
<td>Dunn’s Index</td>
<td>0.0359</td>
</tr>
<tr>
<td>Alternative Dunn's</td>
<td>0.7561</td>
</tr>
</tbody>
</table>
The Table 4.1, Table 4.2 clearly shows the performance of the proposed method over the traditional K-means algorithm. The PSO is suitably used to find the cluster centroid and the K-means algorithm uses the initial seeding from the PSO. The K-means are used to find the clusters among the stock, which is further used for the portfolio creation. The cluster produced must be a qualitative for backing such kind of applications, which is highly dynamic and involves the money of the investor.
Figure 4.2: Comparison based on Silhouette Measure for cluster validity

![Comparison based on Silhouette Measure](image1)

Figure 4.3: Comparison based on Davies Bouldin Measure for cluster validity

![Comparison based on Davies Bouldin Measure](image2)
Figure 4.4: Comparison based on Calinski-Harabasz Measure for cluster validity

Figure 4.5: Comparison based on Krzanowski-Lai Measure for cluster validity
Figure 4.6: Comparison based on Dunn’s Index Measure for cluster validity

Figure 4.7: Comparison based on Alternative Dunn’s index Measure for cluster validity
The Figure from 4.2 to 4.7 clearly depicts the graphical representation for the differentiation between the K-means enable clustering and the PSO adopted K-means. The proposed method outperforms the existing one.

The portfolio has been identified by both the methods. The monthly returns from the nifty have been taken and the returns of the proposed method have been computed on the basis of K-means and the PSO adopted K-means.

**Table 4.3: Weights of stock taken for the portfolio by both the methods**

<table>
<thead>
<tr>
<th>Portfolio by K-means</th>
<th>Portfolio by PSO K-means</th>
</tr>
</thead>
<tbody>
<tr>
<td>Companies</td>
<td>Weights</td>
</tr>
<tr>
<td>Reliance</td>
<td>0.15</td>
</tr>
<tr>
<td>Herohonda</td>
<td>0.29</td>
</tr>
<tr>
<td>Bharti</td>
<td>0.06</td>
</tr>
<tr>
<td>Tata steel</td>
<td>0.2</td>
</tr>
<tr>
<td>Ranbaxy</td>
<td>0.3</td>
</tr>
</tbody>
</table>
Table 4.4: Returns from the portfolio created

<table>
<thead>
<tr>
<th></th>
<th>Mar-10</th>
<th>Apr-10</th>
<th>May-10</th>
<th>Jun-10</th>
<th>Jul-10</th>
<th>Aug-10</th>
<th>Sep-10</th>
<th>Oct-10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monthly returns from Nifty</td>
<td>6.6</td>
<td>0.6</td>
<td>-3.6</td>
<td>4.4</td>
<td>1</td>
<td>0.6</td>
<td>11.6</td>
<td>-0.2</td>
</tr>
<tr>
<td>Returns from K-Means</td>
<td>6.261</td>
<td>2.4579</td>
<td>1.0132</td>
<td>3.593</td>
<td>5.003</td>
<td>3.072</td>
<td>7.899</td>
<td>2.72</td>
</tr>
</tbody>
</table>

Figure 4.8: Comparison based on returns of the portfolios with respect to the Nifty

Returns of the Portfolios with respect to Nifty

- Monthly returns from Nifty
- Returns from K-Means
- Returns from PSO K-Means
The Figure 4.8 clearly depicts the returns got along the mentioned time period for the portfolio depicted by K-means and the PSO adopted K-means. It has been compared with the returns from the Nifty at the time period under consideration.

4.9 Conclusion

A novel model for the portfolio creation through the PSO adopted K-means algorithm along with the Markowitz model has been demonstrated. The need for the PSO for seeding the K-means is clearly stated and the results also ensure the effectiveness of the novel approach adopted. The returns are slightly promoted by the proposed approach. This shows the need of improvement in the stage 2 of the algorithm discussed. The future work concentrates on the improvement of the portfolio creation method, which could be an alternative to the Markowitz model, but we can’t completely deny the modern portfolio theory. The problem could be adopted with the methods like multi objective optimizations through one of the evolutionary algorithms.