Formulated Dynamic Unsupervised Machine Learning Algorithm Trained over Fused Solitary Value of Key Performance Ratios of Stock for Acquiring Optimized Portfolio

S. M. Khalid Jamal[1], A. A. Salam[2], A. R. Zaki[3], H. Abdullah[4]

[1]Department of Computer Science University of Karachi
[2]Department of Computer Science Imam Abdulrahman Bin Faisal University
[3]Department of Business Administration University of Karachi
[4]Department of Computer Science University of Karachi

Summary
One of the most important aspects of financial management is portfolio selection. Over time, its performance analysis is raising its integrity due to the fact that these affect its shareholders, thus enhancing its importance in the global financial market as well.
Traditional models like Markovitz mean variance, doesn't consider the actual financial position of the firm in its computational procedures, which is fairly injustice to the evaluation procedure. Making the presentation a great drawback as the smallest change in the model would enormously affect in decision making. For example, companies may be selected solely on basis of high return while chances of bankruptcy or financial instability being evident, making the evaluation, selection and optimization inappropriate.
Contrary to the above, a different approach to achieve efficient evaluation and selection of scripts has been introduced in the current undertaken seminal research. This is achieved by calculating the key performance rations of the firm present for the evaluation of the financial positioning. The evaluated measures will be unified (The calculated values are summed-up to a single value for each script i.e. for individual company) which in turn will be utilized with modified K-Means dynamic clustering algorithm.

Key words:
modified k-means; dynamic clustering; performance evaluation.

1. Introduction

The modern methods of portfolio selection and optimization are based on the classical approach of calculating standard deviation and variance of the return metrics whereas the financial standing and performance of the script is not being considered at all.
A novel approach has been proposed in this paper which utilizes the ratios (financial, operational and liquidity) which represents the actual performance (on financial basis) of the firm, which can be relied upon for the selection and optimization methodology of different scripts in the portfolio.

The classical k-means clustering algorithm has been intensely examined and modified for a couple of phases to be utilized with ratios representing financial standing of a firm (the algorithms is adopted due to its easiness and simplicity of implementation). In standard K-Means algorithm the total amount of clusters has been predefined by the user. The algorithm then organizes the objects in predefined number of groups whose members exhibits some kind of similarity. However, initial centroid selection in k-means algorithm is done randomly which does not bring about definiteness of cluster. Thus, existing k-means is deficient in terms of appropriate clustering methodology because clusters are being created on user's specification rather than dynamic placement based on nearest proximity of the data set, along with the issue of time invested in computation of reallocation and reassignment of cluster centroids in too many numbers of iterations.
An innovative approach has been proposed and embedded in k-means algorithm for the selection of better initial centroids through centroid estimation criteria described in section [4.1.3], which helps to improve the quality of clusters and provide less iterations with k-means. Further a newfangled approach has been formalized utilizing company's representative ratios (Key performance financial indicators) along with modification of k-means, introducing dynamic clustering methodology to place similar valued scripts in unique clusters. The maximum valued script's cluster is the desire selected and optimized portfolio.

2. Overview of Adopted and Modified Research Algorithm

Presentation of the elements of data in different clusters is considered as the basic issue in most of the fields namely knowledge discovery, data mining and pattern classification (Davidson, Ravi, 2007).

Manuscript received August 5, 2019
Manuscript revised August 20, 2019

The data mining field had attained a lot of popularity since a decade ago, however the market is facing a very tough competition currently primarily due to the time and quality of information produced. As they play an integral part for decision making for the information-based industry and attracting consideration from the society.

In the real word data is available in large amounts and is equivalently difficult to examine the database to extract meaningful data and convert it to desired information and align within the time frame. Here data mining comes to help separate the vast database into a form where it can be used for every specific errand (Wu, 2012). The use of data mining is remarkably massive and is particularly useful in applications such as market trends, customer's shopping patterns, fraud detection, science explorations, production controls and so on. On basis of random experimentation, it has been widely noted that data mining has enabled extraction of vast amount of available data. Data mining has also been able to predict the nature/pattern of various discoveries (Huang, 1998).

Clustering has eventually made relevant and delivered vital information to the fields of data compression, image processing, artificial intelligence, data mining, design acknowledgement, machine learning, market analysis, pattern recognition Clustering sorts like with like data and differentiates between the groups similar and dissimilar to each other (Celebi, 2015). By doing so it sorts a vast amount of random data into large groups of organized elements (Sreejesh, Mohapatra, Anusree, 2013).

This amount of work can be done under three categories supervised, semi-supervised or unsupervised. Some type of algorithms such as distinctive algorithms can be sorted unsupervised as the nature and description has already been made in such a manner that further distinction is not required (Xu, Tian, 2015).

In the real world its quite a task to calculate the quantity of clusters if data hasn't been entered properly for example if a small specified number of clusters could be possibly keeping different objects in the same group and similarly a large number of clusters of similar objects into different groups therefore most of the algorithms make clusters which are allocated according to the data input (Wang, Yu, Bao. Zhang, 2005).

It wouldn't be wrong to say that the algorithm k-means is one classical algorithm which has been repeatedly researched and has managed to be one of the simplest and practicably easiest algorithms (Reddy, Mishra, Jana, 2011). The proposed research proposes an improved and flexible k-means in the sense that if the end-user requires a specified amount of clusters then the algorithms automates itself for that defined set and if unspecified is required then using dynamic methods it creates an unspecified number of clusters, these are dependent on the data elements' cohesiveness and relativity. This modification enables k-means to repeat iteration and simultaneously increase the quantity of the clusters one by one till the conditions of inter and intra clusters are fulfilled. (1)

3. K-Means Clustering

K-mean clustering sets information in order, in other words it equips unordered data together. The Euclidean method is utilized to measure the distance of the centroid which is then sorted on basis of the minimum distance thus signifying the definition of K-mean clustering (MacQueen, 1967).

3.1 K-Means Clustering Algorithm

Here information collected from the database is stratified into various clusters, the term "k" is titled by the client (Tang, Khalid, 2016). Firstly, k-centroids in spaces (C1 to Ck) are settled into random locations, and then go over the data for each element with the intention to select the element closets to the centroid. For that individual xi, the distance among both the components xi and cj, for centroid (j) of each cluster is figured out (Salman, Kecman, Li, Strack, Test, 2011). Upon completion of the initial step, the cluster having the least possible distance to the closest center of the mass is selected with the point under consideration i.e. xj being assigned to the nearest centroid (Saegusa, Maruyama, 2007). At this stage, rechecking onto the groups over the K centric points have to be performed. Each centric point is recomputed, this re-computation is performed by calculating the location of each elements available in that cluster. Then, determining the vectors of each components of Xi are adjusted as to the jth cluster. This enables to formulate the average. The average is calculated by adding the components then dividing them by the aggregate of the number of elements present in the jth cluster. This results to a new centric point which would now be considered as a centroid for cluster j. Having the repetitive iteration execution done, the algorithm will reach to a point where no change or conversion to other values will be available, hence leading to the cluster members not changing their clusters (i.e. inter-cluster member switching will no longer exist). Successful completion of this stage will enable the algorithm to exit the execution phase (Mohd, Beg, Herawan, Rabbi, 2012).

3.2 K-Means Clustering Algorithm

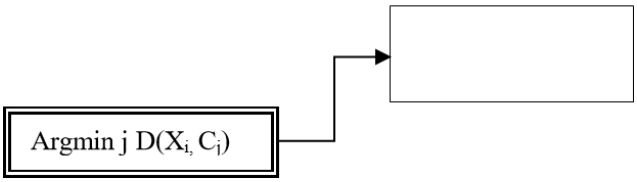
Step I: [Read Input] K, set of elements xi

Step II: A Centroid placement criterion is C1to Ck randomly at different locations in data set.

Step III: Iterate till the convergent state:

a) For each point Xi:

- Identify closest centroid Cj



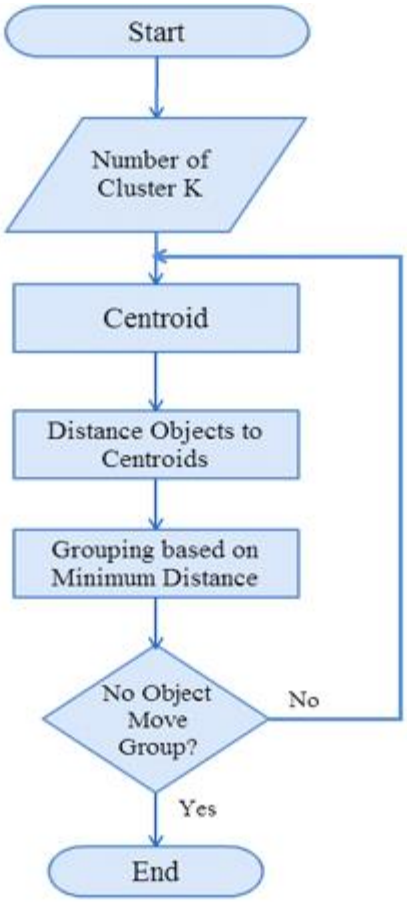
• Allocate the Xi point to cluster j

b) For each cluster j=1 …. K:

$$C_j(a) = \frac{1}{n_i} \sum \overrightarrow{x_{\iota\, to\, c_\iota}}\ x_i(a)$$

• Current centroid Cj= average of all Xi points allocate to cluster j in the above step.

Step IV: Exit → after cluster assignments remain unchanged.



3.3 Complexity of the Algorithm

The speed on the K-means algorithm is relative to its comparison. For example, if compared with any other clustering algorithm, K-means will always result to be the fastest.

In order to analyze the complexity of K-means algorithm, the number of iterations needed for consideration of the convergence would also include the centroid reallocation and reassignment of each value of the other clusters being compared in the iteration. To express in other words, the algorithm computes (for each element in data set) the distance with all the cluster centroids i.e. the total of cluster times, the total of instances times, and the total of dimensions time (Vattani, 2009). Adding more, the longer the time invested in computation the costlier it becomes to figure out that one little thing in too many numbers of iterations. The following equation for the algorithm of K-means can highlight the complexity:

$$O(d \times I \times n \times k)$$

where,

‘d’ refers to number of focuses,
‘I’ refers to number of clusters,
‘n’ refers number of iterations,
‘k’ refers to number of attributes/ dimensions.

3.4 K-Means Properties

It reduces the aggregate intra-cluster distance.

$$\sum_j \sum_{xi} \rightarrow_{cj} D(cj\ xi)^2$$

- The squared separation is added from the point to the focus of its cluster.
- The result would result the same if the distance is calculated by the Euclidian method (Manthey, Röglin, 2009).

Joins to a neighborhood least

- Initial stages differ this results to different outcomes
- Iteration repetition with arbitrary beginning stages
- The smallest aggregate distance is picked up from the group

The closest is sent to the similar group whose objects did not complete.

3.5 Assessing K-Means Clusters

Sum of Squared Error (SSE) is the most familiar method for assessment:

The distance of each point closest to the cluster is the error (Maimon, Rokach, 2006).

$$S.S.E = \sum_{j=1}^{k} \sum_{x \in cj} dist(x, mj)$$

Next the cluster with the minimum error available from the two sorts is chosen.
Another simple procedure to diminish (S.S.E) is by the expansion of the number of clusters, K.
In a good clustering, there will be a less SSE with smaller K.

3.6 Issues with Selecting Initial Points

There is a little chance for choosing a centroid from each cluster when it comes to the case of real clusters (Goyal, Kumar, 2014).

- A substantial K indicates that a moderately small chance of issue would exist.
- If the size(n) of the clusters is same, then P = $\frac{n}{\lceil \frac{Kn}{K} \rceil}$
- For example, if K=3, n=10, then probability = $\frac{10^2}{\lceil \frac{30}{3} \rceil}$ = 0.246.

At times the underlying centroids will correctly revise themselves but sometimes they don't (Kant, Ansari, 2015).

3.7 Solutions to Initial Centroids Problems

Several tries may help but at the same time may not produce a proper random centroid.
To recognize the starting centroids a sample has to be taken and then hierarchal clustering shall be applied (Luo, Fu, 2014).
Select from the K introductory centroids and take the centroids amongst it

- The centroid which is isolated foremost should be taken.

Post-processing

3.8 Limitations Of K-means

Some of the limitations of K-mean are issues of sizes, thickness and non-spherical shapes of clusters (Chang, 2009).
If the data contains outliners (i.e. divergent data) K-mean will not produce good results.
Nonetheless, k-mean tends to be the most generally practicable clustering algorithm due to its simplicity, efficiency to stream data, and reasonably scalable despite its limitations (Kant, Ansari, 2015).

4. Proposed Algorithm for Optimized Portfolio

Step 1: Select highly traded volume-based scripts.
Step 2: Extract at-least 15 sectors from highly traded scripts.
Step 3: Select 10 top most highly traded scripts by volume in each sector.
Step 4: Extract the following financial ratios, operational ratios and liquidity ratios from individual company financial report.
I. [Loop]: for (c [i] = 1; c [i] < n ; c [i] ++)
i. SET c = [company(ies)]
ii. SET r ={ratios}
iii. Let RF, RO and RL be the ratios of all companies.
• FR: = {frc1, frc2, frc3,.., frcn}
• OR: = {orc1, orc2, orc3,…, orcn}
• LR: = {lrc1, lrc2, lrc3, …., lrcn}

Where frc1 = (DYR + ROE)
orc1 = (NPR + RTP)
lrc1 = (CR + QR)

II. Compute the average of extracted financial, operational and liquidity ratios of each script and store it in an array say D[i].
D[i] = { [(frc1+orc1+lrc1) /3], [(frc2+orc2+lrc2) /3], ……, [(frci+orci+lrci) /3] }
Step 5: Apply the Modified K-Mean algorithm for dynamic clustering of optimized script values.
a) [Read Input] K, quantity of clusters (introduce k = 2 for dynamic clustering)
specific quantity of clusters = (Boolean) zero or one
D[i] = data set of average values of ratios of every script.
[initialize] Set maxValue=0, maxCluster=0
b) Compute the mean (average) of all data set values present in array D[i].
c) Estimate data_set Max_Value through mean i.e. MaxM=2$\bar{x}$
d) Division factor [DF]=k+1
e) Multiplication factor [MF]=100/[DF]
f) Percentage factor [PF] = MF / 100
g) [Centroid Estimation] = repeat for n=1 to k
i. C [i] = n*MF*MaxM
h) [End of Step centroid estimation loop].
i) Compare the mean values of each cluster (by using Euclidian distance method) with each value of the data set and values may get re-allocated to the cluster based on the similarity of values to each cluster mean.
j) If the object>maxValue
then maxValue = object's value and
maxCluster = Ki
k) Evaluate the dataset iteratively to extract the mean value unless the cluster assignments remain unchanged.
l) If quantity-of-clusters are fixed = yes, then go to step K.

m) Compute inter-cluster distance using inter=average of the centroids of all Ks.
n) Compute intra-cluster distance by computing standard deviation among the object values in each cluster and take average of all variances.
o) If the distance of old intra cluster is greater than the distance of new intra cluster AND the distance of old inter cluster is less than the distance of new inter cluster, then go to step J else go to step K
p) K=k+1, go to step b.
q) Take maxValue as the greatest value present in D.
r) Take maxCluster to find the cluster containing the greatest value from D (This cluster is the set of all scripts that are supposed to yield highest return).
s) [End of algorithm]

4.1 Description of Optimized Portfolio Algorithm

4.1.1 Data Selection Methodology

i. The process of the proposed algorithm starts by selecting the scripts based on the traded volume. The criteria shall be extended on basis of the frequency of volume. For instance, a daily highly traded volume scripts can be utilized with mathematical intersection function (set theory) in order to calculate the monthly highest traded scripts. In return theses shall be used for all 12 months to with the set intersection function to get the annual highest trade by volume.
ii. In the next step sectors of highly traded (volume-based scripts) is identified. The 15 top most sectors from the highly traded scripts shall be chosen.
iii. After examining each of 15 determined sectors, choose the top 10 highly exchanged volume-based scripts in each sector (the scripts may vary in the selected (sector wise) list when compared with the step one selected scripts due to each script in an individual sector being not present in the most highly exchanged script of the first step. This step normalizes the data (widening the base of the chosen data set).

4.1.2 Extraction of Ratios from Current Fiscal Report of The Selected Script's Companies

iv. Upon completion of the previous stage, the algorithm will extract the financial, operational and liquidity ratios from the current year financial report of each script and compute the average of each script's financial, operational and liquidity ratios. The ratios with increasing value indicating of higher return would be selected only. The average of ratios obtained will be portrayed in an array separately.

Financial Ratios

1. Dividend Yield Ratio: It shows how much "dividend" the company pays to its shareholders in a financial year.
The higher the estimation of profit, the more the extreme point is derived in the computation exhibiting the greater contribution of the specific script in algorithm.
2. Return on Equity: It computes the amount of return a company is able to generate to its shareholders. In order to attract potential investors, the company would need to yield a higher return. The proposed algorithm would calculate and the maximum point would be plotted on the outer/ higher side of the algorithm. Thus, the outer the cluster, the maximum the return.

Operational Ratios

1. Net Profit Ratio: This shows the amount of profit an entity can produce over its investment. A greater value would signify that the chances of maximum value would lie on the positive side. Making work for the selection of the specific script in the undertaken research's algorithm.
2. Receivable Turnover Period: This shows whether the organization's account receivable is effective and also whether the organization has reliable clients that payoff their obligations. Therefore, a high proportion indicates that a conventional strategy for credit is being used. Furthermore, a high value would highlight a higher chance by the algorithm to be plotted on the outer side, maximizing the probability that the script may be placed in the top/outer most cluster.

Liquidity Ratios

1. Current Ratio: This ratio represents the organization's current assets against its current liabilities. If the proportion is of high value this means that the organization has the capability and capacity to cover its short-term obligations and would present the organization more profitable for portfolio selections. Similarly, it would increase the chances of selecting a particular script within the algorithm in the selection of the portfolio.
2. Quick Ratio: As compared to current ratio, the quick ratio is more effective as it does not take current stock and other current resources into consideration which is difficult to convert to money.
A high liquid current position from the quick ratio reflects a positive attitude within the algorithm and helps in plotting the point on the outer/higher side through the modified K-means. This increases the chance of the scripts to be selected in the optimized cluster.

4.1.3 Modification Of K-means Algorithm for Plotting of Optimized Values and Dynamic Clustering Using Mathematical / Statistical Methods

K-means is used for clustering. In the standard algorithm of k-means the initial problem is the random selection of centroid, thereby it does not result in definiteness of clusters (Vij, Kumar, 2012). Furthermore, the time for computation is directly proportional to the following: number of data-points, dimensions and iterations making it very expensive to process a vast amount of data (Kumari, Maheshwa, Goyal.P, Goyal.N, 2015). Due to K-means being highly sensitive to initial centroids a proper selection of initial centroids is compulsory. The newly proposed approach is for the better selection of initial centroid which forms basis of the k-means algorithm. This makes suitable initial centroids useful for better clustering and less iterations

Selecting initial centroids needs the proposed algorithm to primarily compute the average optimized script values previously stored in array D[i]. The value obtained is multiplied by two to get an approximate of the maximum value of the rearranged scripts. The algorithm calculates the division factor in each iteration by addition one in number of clusters 'k' which increase dynamically upon satisfying through the condition of intra and inter cluster. The multiplication factor is then applied to divide the fixed value of 100 by the Division factor [DF] computed in the aforementioned step. For better initial centroid selection, the algorithm estimates the division criteria by taking the percentage of the computed value in multiplication factor.

The estimation criteria is applied "n" number of times equaling to "k" (the iteration will execute from 1 to k times). To assess the equation, multiply the n number of centroids to calculate the percentage and estimate a maximum value computed by twice of the mean.

K-means classical algorithm has been further improved by using a new approach by which the user can decide either go for definite or indefinite quantities of cluster. This forms dynamic clusters by using both inter and intra cluster distances calculated by mean and standard deviation. This has enabled research to not only figure out certain points close to each other but also creates a dynamic pattern enclosing all close neighboring points in the cluster.

For the unspecified quantity of clusters, the algorithm adds one in the quantity of clusters upon fulfilling the conditions of inter and intra cluster distances in each iteration. This helps in attain the statutory requirements of the cluster quality. In contrast with the k-mean algorithm, the modified algorithm establishes that it not only enhances the cluster's quality but also can form the clusters dynamically. Separate elements are assigned during the execution phase to the most appropriate clusters, this depends on the relativity and feasibility of clusters. This increases the population of the other clusters concurrently. When the iteration termination provision is met, the proposed algorithm furnishes the best possible clusters. These clusters have the nearest neighbor data within and have the least intra cluster distance,

The standard algorithm of k-mean finds the quantity of clusters specified by the users themselves (Le, Kim, 2015). However, in contrast, while in practical circumstances, it is very necessary to locate the quantity for uncertain informational collection on the execution time (Suryanarayana, Rao, Veeraswamy, 2015). If the number of clusters is static the Cluster's quality would be of low eminence.

The innovative approach proposed will instead find the number of clusters dynamically regarding the output quality of the cluster. There is flexibility for the user to either define the quantity of clusters or the quantity of clusters to be decided upon the algorithm (created dynamically on basis on the conception of the nearest neighbors).

The modified algorithm shall treat the predefined number of clusters the same way as the K-mean algorithm does. For example, for a predefined set of clusters the cluster counter would constantly increase by one until the conditions of the group quality limit are achieved. The proposed algorithm shall determine the combination of the above scripts that yields maximum output, in simple words it will evaluate the cluster hosting the highest value.

i. Intra-cluster Measure

The word intra is a prefix and is defined as "within". Likewise, the word Intra-Cluster measure describes the closeness or near proximity of the elements within a cluster. "Standard Deviation", a statistical tool, is required to calculate the Intra-cluster measure of each element present. A better result would mean that the clusters are near to the elements within and vice versa. Below is the equation to calculate the intra-cluster distance:

$$intra\ cluster\ distance\ = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Xi - Xm)}^{2}$$

ii. *Average Inter-cluster Measure*

The word inter is a prefix and it defines "among/between", hence the word Inter-cluster is used to calculate the separation between the centroids of the cluster. The farther the separation, the better the quality and shape of the cluster. An average of all K's is taken, this is supported by the following equation:

$$Inter\ cluster\ distance\ = \frac{m_k + m_{k+1}}{no.of\ centroids}$$

Where,
* m represents the value of centroid (mean value of a cluster)
** m_k & m_(k+1) will increase by one in each iteration and simultaneously the number of centroids would increase as well. Hence the formula for the next iteration would be as follows: m_k+m_(k+1)+ m_(k+2) and beyond (to be referred later in this section)
The computed value shall be used to compare the value of the inter-cluster distance with the next iteration (explained ahead). The above process would repeat till the condition is satisfied.
When it comes to dynamic clustering, it will start off with K=2 (K is the number of clusters). After calculating the distance of each element in the data using the Euclidian method three further steps occur before moving on to the next iteration. Firstly, the intra-cluster distance is calculated (using standard deviation) the secondly the inter-cluster method is calculated (using the average function of the two different clusters). And lastly the values attained of both intra-cluster and inter-cluster are saved for comparison with the next iteration. Once these steps are completed the next iteration shall start.
In the next (second) iteration begins with setting K=3. The steps above shall be repeated but this time they are classified to any three clusters. The centroids are re-assigned to each of the "to be created" clusters. Once the intra and inter cluster distance will be measured and those values will be compared with the previous iteration's values.
The proposed algorithm here compares the two classifications of K (2 in the first and 3 in the second). This is done by the following command:
["Compute the average of all Intra cluster distance of first iteration (in this case when K=2)" and IF it is greater than "computed average of all Intra cluster distance of second iteration (in this case when K=3)"] AND ["the computed value of inter cluster distance of first iteration (where k = 2)" IF it is less than "the computed value of inter cluster distance of second iteration (where k = 3)"]
In words, as mentioned before the algorithm will increase the value of k by one till the conditions are met. If the conditions are not met, it will exit the iteration phase first and then figure out the maximum value cluster.
Three simple rules to understand the plotted graph by k-mean algorithms are first, points of scripts near the origin mean they are performing and returns are low, second points of scripts farther from the origin mean performance and returns is high and third, points on the outer or higher side of the graph indicate the highest returns.
Dynamic clusters are created when the points are near to each other. The clusters represent selected scripts which are converted to portfolios. Applying the above rules, if clusters are near the origin this means that all the companies at those points are going to have a low return on their portfolio, in the same way clusters farther from the origin will have a higher return. Cluster even farther than these are also shown, they also contain those scripts which have a minimum distance to each other i.e. they show the nearest neighbors and have the highest return on the portfolio as they are even farther from the clusters with higher returns. The fact that closely related clusters in the proposed algorithm create dynamic clusters makes this approach innovative, for the reason that it helps identify those scripts which would give a similar return/output.
The aim primarily for this research is to identify the scripts which provide higher returns (on the outer/higher side of the graph). It is in this cluster where the max-value would be, it would be a perfect grouping of all the scripts which provide the maximum output and contribute to the most optimized portfolio.
Max-Value Cluster
Max-value clusters are found within those clusters which are at the top most and outermost cluster. The outermost cluster is found out by traversing through the data and choosing the maximum value at the start of the algorithm.
The max value cluster illustrates that at this point not only would the return be high but also the most optimized portfolio as well. Companies plotted at this point would have the highest profits.

5. Results

The current undertaken reconnaissance initially evaluates the company's financial position and its market standing through the company's current market stock traded values and book values, further it takes into account the company's financial ratio's which facilitates in appraisal of the company's current financial positioning, all the related values are amalgamated through mathematical fusion techniques (a single value is computed/premeditated for a single company) which is then provided to the modified (dynamic clustering methodology has been innovated) K-Means algorithm which places/locates the relevant value (single evaluated value for a single company) in a near neighboring set of values (multiple near-proximity values of multiples scripts each of which is represented by a single value).
The graphs plotted just below the current text (Gp-1 using Market Value of a single company's script and Gp-2 using gross-profit ratio) are the plotting of a single script's fluctuation (of dynamically evaluated values through mathematical fusion functions of different afore mentioned script related assessments) over a longer period of time. The plotted graph is hard to depict for any of the future forecasting and it also couldn't be explained for any trending as the frequency and amplitude of the wave is too

high and random in nature (i.e. having high variance, in-fact too many spikes up and down).

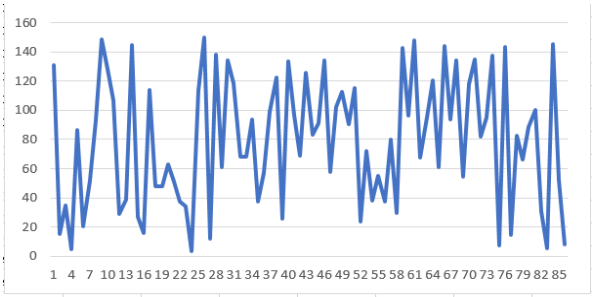


Fig. 1 Single Script Estimated Value Using "Market Value" of Script

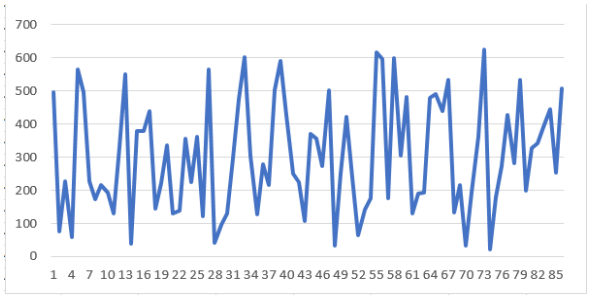


Fig. 2 Single Script Estimated Value Using "Gross Profit Ratio" of Script

After failing to meaningfully utilizing the Market Value of script initially and later onwards again missing to assess any evocative results using gross-profit ratio, a diverse mathematical fusion methodology is being insinuated to utilize all of the available ratios (namely financial / operational / liquidity ratios) to compute single value for a single script (explained in previous sections). That single value per script is plotted over a two-dimensional canvas and modified dynamic K-Means clustering is being utilized for acquiring the (random number of) clusters, the plot and clustering is presented below (Gp-3)

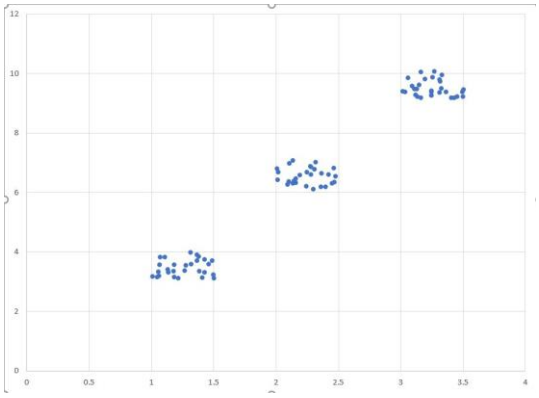


Fig. 3 Plot of different scripts utilizing modified K-Means algorithm

The plotted values are clearly lying in 3 different neighborhoods of similar domains. Straightaway it can be effortlessly observed that the upper righter most cluster items is having those scripts which are yielding maximum return, i.e. that have the highest market value and also the highest dividend yield along with other company evaluation ratios on the higher side thus represents the optimized portfolio by automated means.

6. Conclusion

In this study, the financial statement of the company is evaluated to find the financial positioning of the company through which the company's market standing has been evidently estimated. On those ratios based a single value is being computed which then is furnished to the machine learning algorithms that worked after learning through the historic data set and have successfully recommended a set of considerable high return yielding scripts which was the ultimate aspiration of the investigation i.e. optimized portfolio

References

[1] Davidson, Ian, and S. S. Ravi. "The Complexity of Non-Hierarchical Clustering with Instance and Cluster Level Constraints." Data Mining and Knowledge Discovery, vol. 14, no. 1, 2007, pp. 25–61., doi:10.1007/s10618-006-0053-7.

[2] Wu, J. (2012). Cluster Analysis and K-means Clustering: An Introduction. Advances in K-means Clustering Springer Theses, 1-16. doi:10.1007/978-3-642-29807-3_1

[3] Huang, Zhexue. "Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values." Data Mining and Knowledge Discovery, vol. 2, no. 3, Sept. 1998, pp. 283–304., doi:org/10.1023/A:1009769707641.

[4] M.Calebi. "Partitional Clustering Algorithms." Springer. 2015, doi.org/10.1007/978-3-319-09259-1.

[5] Sreejesh, S., Mohapatra, S., & Anusree, M. R. (2013). Cluster Analysis. Business Research Methods, 229-244. doi:10.1007/978-3-319-00539-3_10

[6] Xu, D., & Tian, Y. (2015). A Comprehensive Survey of Clustering Algorithms. Annals of Data Science, 2(2), 165-193. doi:10.1007/s40745-015-0040-1

[7] Wang, Daling, et al. "An Optimized K-Means Algorithm of Reducing Cluster Intra-Dissimilarity for Document Clustering." Advances in Web-Age Information Management Lecture Notes in Computer Science, 2005, pp. 785–790., doi:10.1007/11563952_81.

[8] Reddy, D., Mishra, D., & Jana, P. K. (2011). MST-Based Cluster Initialization for K-Means. Advances in Computer Science and Information Technology Communications in Computer and Information Science, 329-338. doi:10.1007/978-3-642-17857-3_33

[9] MacQueen, James B. Some Methods for Classification and Analysis of Multivariate Observations. Defense Technical Information Center, 1966.
[10] Tang, Q. Y., & Khalid, M. A. (2016). Acceleration of k-Means Algorithm Using Altera SDK for OpenCL. ACM Transactions on Reconfigurable Technology and Systems, 10(1), 1-19. doi:10.1145/2964910
[11] Salman, R., Kecman, V., Li, Q., Strack, R., & Test, E. (2011). Two-Stage Clustering with k-Means Algorithm. Communications in Computer and Information Science Recent Trends in Wireless and Mobile Networks, 110-122. doi:10.1007/978-3-642-21937-5_11
[12] Saegusa, T., & Maruyama, T. (2007). An FPGA implementation of real-time K-means clustering for color images. Journal of Real-Time Image Processing, 2(4), 309-318. doi:10.1007/s11554-007-0055-8
[13] Mohd, W. M., Beg, A. H., Herawan, T., & Rabbi, K. F. (2012). MaxD K-Means: A Clustering Algorithm for Auto-generation of Centroids and Distance of Data Points in Clusters. Communications in Computer and Information Science Computational Intelligence and Intelligent Systems, 192-199. doi:10.1007/978-3-642-34289-9_22
[14] Vattani, A. (2009). K-means requires exponentially many iterations even in the plane. Proceedings of the 25th Annual Symposium on Computational Geometry - SCG 09. doi:10.1145/1542362.1542419
[15] Manthey, Bodo, and Heiko Röglin. "Worst-Case and Smoothed Analysis of k-Means Clustering with Bregman Divergences." Algorithms and Computation Lecture Notes in Computer Science, 2009, pp. 1024–1033., doi:10.1007/978-3-642-10631-6_103.
[16] Maimon, O., & Rokach, L. (2006). The Data Mining and Knowledge Discovery Handbook (Ser. 326). New York, 233 spring street: Springer Science Business Media. doi:https://books.google.com.pk/books?id=S-XvEQWABeUC&lpg=PP1&pg=PP1#v=onepage&q&f=false
[17] Goyal, M., & Kumar, S. (2014). Improving the Initial Centroids of k-means Clustering Algorithm to Generalize its Applicability. Journal of The Institution of Engineers (India): Series B, 95(4), 345-350. doi:10.1007/s40031-014-0106-z
[18] Kant, S., & Ansari, I. A. (2015). An improved K means clustering with Atkinson index to classify liver patient dataset. International Journal of System Assurance Engineering and Management, 7(S1), 222-228. doi:10.1007/s13198-015-0365-3
[19] Luo, Ying, and Haiyan Fu. "Modified K-Means Algorithm for Clustering Analysis of Hainan Green Tangerine Peel." IFIP Advances in Information and Communication Technology Digital Services and Information Intelligence, 2014, pp. 144–150., doi:10.1007/978-3-662-45526-5_14.
[20] Chang, J. (2009). SDCC: A New Stable Double-Centroid Clustering Technique Based on K-Means for Non-spherical Patterns. Advances in Neural Networks – ISNN 2009 Lecture Notes in Computer Science, 794-801. doi:10.1007/978-3-642-01510-6_89
[21] Vij, R., & Kumar, S. (2012). Improved k- means clustering algorithm for two-dimensional data. Proceedings of the Second International Conference on Computational Science, Engineering and Information Technology - CCSEIT 12. doi:10.1145/2393216.2393327
[22] Kumari, S., Maheshwari, A., Goyal, P., & Goyal, N. (2015). Parallel Framework for Efficient k-means Clustering. Proceedings of the 8th Annual ACM India Conference on - Compute 15. doi:10.1145/2835043.2835060
[23] Le, V., & Kim, S. (2015). K-strings algorithm, a new approach based on K-Means. Proceedings of the 2015 Conference on Research in Adaptive and Convergent Systems - RACS. doi:10.1145/2811411.2811472
[24] Suryanarayana, S. V., Rao, G. V., & Veeraswamy, G. (2015). Spectral Clustering Algorithm for Navie Users. Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET 2015) - ICARCSET 15. doi:10.1145/2743065.2743076.