A Novel Approach for Cluster Formation of Virtual Machines Using Elbow Means Technique

Avinash Kumar Sharma
Department of Computer Science and Engineering, Uttarakhand Technical University

Chanderwal, Nitin
Department of Electrical Engineering and Computer Science, University of Cincinnati

Yadav, Mitul
Department of Computer Science and Engineering, Dev Bhoomi Institute of Technology

https://doi.org/10.5109/7183443

出版情報：Evergreen. 11 (2), pp.1326-1332, 2024-06. 九州大学グリーンテクノロジー研究教育センター
バージョン：

A Novel Approach for Cluster Formation of Virtual Machines Using Elbow Means Technique

Avinash Kumar Sharma[1*], Nitin Chanderwal[2], Mitul Yadav[3]

[1] Department of Computer Science and Engineering, Uttarakhand Technical University, India

[2] Department of Electrical Engineering and Computer Science, University of Cincinnati, Cincinnati, USA

[3] Department of Computer Science and Engineering, Dev Bhoomi Institute of Technology, Dehradun, India

*Author to whom correspondence should be addressed:
E-mail: avinashsharma2006@gmail.com

(Received September 14, 2022: Revised April 8, 2024: Accepted June 14, 2024).

Abstract: Cloud computing has been adopted by the various industries for applications that are demanding for the huge amount of compute power or resources. The main advantage of using the cloud computing is that various resources like servers, disks and bandwidth are allocated as per the demand of the application, which may be elastic in nature. This removes the overhead of the end user to search for the pool of resources, instead are made available as the virtual machines. The virtual network embedding is a technique to embed networks on virtual machines that are compatible of performing tasks and run optimally with low latency. In the paper k means based approach is used to create the cluster having virtual machines and virtual networks that will first allocated with respect to geolocations then the compatible cluster is to be calculated if the conditions not satisfied. This will serve as a huge time decrement of finding optimal clusters and with lower latency task will performance the same has been justified from the results.

Keywords: Cloud computing; Energy-efficient model; Virtual machine management; Virtual cluster; Elbow Means

1. Introduction

Cloud computing has enabled the business and users with the virtual resources that can be used to execute the jobs on these virtual devices with the help of internet. In Cloud computing paradigm the various resources are kept in a pools by the service providers like Amazon, Google and Microsoft and the client may request the resources and have to pay for the resources that have been used by them. Pay as per model is so flexible that client nowadays can even for minutes not even hours. Pay per as usage has made the complete system very economical and client doesn't even have to think about the maintenance. Another major advantage of using the cloud infrastructure is availability of the data as these machine keep the replication of data and are geographically apart which make the more reliable in comparison the standalone machines[1)]. The service providers would be able to provide the services if they can generate the profit from these infrastructure or services. As there is heterogeneity in the demand by different client hence the quality of service is to dictated by the client and has registered in the Service level agreements. Once the cloud service provide gets the requirements from the different clients the main objective of the service provider should be allocation the resources such that the profit should be maximized. This maximization is also constrained by the quality of service for the different resources that were recorded in service level agreements. The allocation can be optimized to gain the maximum profit and maintaining the response time, ensuring the availability, reliability [2)]. Determining the right amount of cloud resources for executing the user requests depends on the incoming request at the particular time.

The proposed solution is based on virtual network embedding in an optimal means using machine learning algorithms. This paper briefs cluster formations of virtual machines having communication and performing tasks of word count etc. With each other there are three parameters used that is processing, memory and latency. These three are based on the actual distances between the virtual machines and the sub clusters used for communicating between two virtual machines of different clusters having separated with relocation [7)]. Proposed methodology consists of geo-locations having latitudes and longitudes. The dataset is also provided to have a check on latency. A model named K-virtual-means [1)] is fitted on location parameters of virtual machines of different clusters. Once the cluster is allotted we need to check if the clusters are compatible enough to perform tasks if not as mentioned the generation of clusters at dynamic will serve the

purpose for allocation of optimal clusters. Now if this is the case we will increment the number of data points cluster and try to generate optimal clusters for the task out paper has both algorithms that are being implemented and is serving high quality, low latency and minimum time.

2. Related Work

The comparison of the various related works by the different authors in the recent years have been summarized in the table 1.

3. Tools and Technologies Used

3.1 Oracle virtual box

The setup of virtual machines are done here n1 number virtual machines belong to cluster-1 and n2 number of virtual machines belongs to cluster-2. The clusters are intercommunicating [10] and they are sending files to each and performing specific tasks.

3.2 Hadoop

It provides a GUI and a running environment for all virtual machines and respected clusters. Here we can see the queues and data of CPU cores and memory usage[5-9]. It is a solution towards big data as we are using it for a computational environment. In this scenario we have created clusters having name nodes and data nodes. The name node is responsible for maintaining the metadata of the data stored on the data nodes.

3.3 Machine Learning

We have used machine learning model to compute virtual machines having minimum distances between clusters and form a centroid by our proposed solution, The advantage of using machine learning is to find a minimum spanning tree with a good accuracy and in dynamic time very quickly as model is pre-trained and will take value of latitude and longitude as input of one virtual machine to other and will form a minimum and optimal path. The proposed algorithm is an extension to a k-means [2] which we are using.

4. Algorithm

In the current scenario, we have used the variant of k-Means clustering algorithm, which used the Euclidian distance to identify the dissimilarity between the two clusters. The formula for the Euclidean distance is described in equation (1):

$$D_i = \sum_{k=1}^{n_i} \quad \sum_{j=1}^{n_i} \quad \left|\left|d_k - d_j\right|\right| \qquad (1)$$

The main challenge of k means algorithm is to find the optimal values of k. These values of k can be used to make the different clusters with the different data nodes.
The solution for placing the data nodes into different cluster is done with the help of Elbow method which is used to decide the optimal number of clusters required.

- Elbowformation(X)

$$W_k = \sum_{i=1}^{k} \quad \frac{1}{n_i} D_i \qquad (2)$$

- where, i , iteration in range of values for optimal cluster
- W_k , average weight for a cluster 1 to k that needs to be minimized
- n_i , number of samples in X
- D_i , Euclidean distances in all pair of clusters

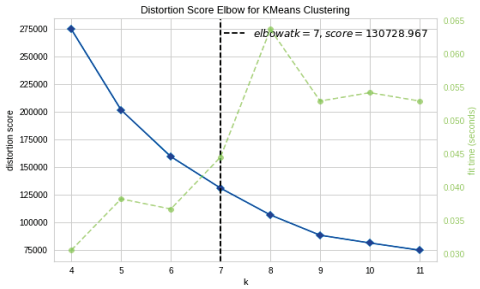

Fig. 1: K Means Clustering

We need to minimize W_k then return k where W_k becomes flat, as shown in Fig. below elbow method flattens at the value of 7 it is computed with the above algorithm and can be visualized.

4.1 Algorithm: K-Virtual-Means (X)

Step 1: Call Elbowformation(X).
Step 2: Value of k will get computed from step 1.
Step 3: Initialize centroids and select k data points for centroid (μ_k).
Step 4: While the data points assignment becomes flat, keep iterating.
Step 4.1: Call Objective function (X_i, μ_k) Where, X_i = data points μ_k= centroid
Step 4.2: Assign each data point (X_i) to the closest cluster (μ_i) depending on value return from step 4.1
Step 4.3: Compute centroids for the cluster by taking the average of all data points that belong to each cluster.
Objective function (X_i, μ_k)

$$J = \sum_{i=1}^{n} \quad \sum_{j=1}^{k} \quad w_{ij} | \quad |X_i - \mu_j| \quad |^2 \qquad (3)$$

4.2 Centroid

The centroids can be chosen and can be plotted the same as shown in Fig. 2 [3] . These are optimized clusters that have become a label of data set clusters as optimized labels and it is the end result of the mapping of clusters [6].

5. Data Description

Table 1. Attributes of dataset

Attributes	Type
Country code	String
Latitude	Geolocation vector
Longitude	Geolocation vector
Cluster	Nearest label

Table 2: Describes the various proposals by different authors to address the problem

S. No.	Article	Methodology	Prons
1	Chavan et'al [12]	Used K-means for clustering	Higher availability of resources Easy Scheduling
2	Rajabzadeh et'al [13]	absorption mode in Simulated annealing	Energy saving
3	liu et'al [14, 22]	greedy mechanism for winner determination and payment algorithms	Best overall use of resources
4	Aktan et' al[15]	hybrid algorithm based with SA along with GA and DE along with greedy algorithm	fast completion time and effective load balancing
5	Asghari et'al [16, 28]	coral reefs optimization algorithm and multi-agent deep Q-network	Reduced energy consumption
6	Raju et' al [17, 25]	K-medoid particle swarm approach for makespan	effective resource sharing
7	Muthusamy et'al[18, 23]	K means clustering based on average task length	Minimum make span time Low execution time less deviation from workloads
8	kiani et'al [19, 26]	Used Chemical reaction optimization based on network awareness and power efficient	fast execution and power efficient
9	Asenio et'al [20]	Used the concurrent scheduling algorithm based on greedy randomized adaptive search procedures	Close to optimal solution Very small time required for scheduling
10	Yin et'al [21, 27]	Proposed the inter and intra cluster algorithms for allocation	Energy efficient Workload balance

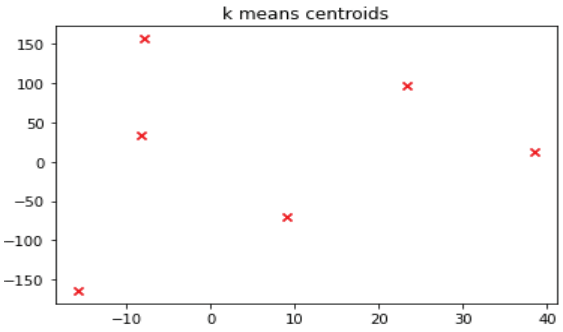

Fig. 2: K Means Centroids

6. Data Description

6.1 Dataset Description

Figure 3 describes the snapshot of the dataset used.

6.2 Inference Statistics

Applying statistics on data set we will specifically look in three concepts that are normal distributions, skewness, and kurtosis. Here normal distribution is how data is spread across as values we will also have some insights on the dataset which is normally distributed. The skewness describes about the symmetry of data set values or does normal distribution. The kurtosis describes the peak of normal distribution.

6.3 Normal Distribution of longitudes

The spread of latitude values are equally distributed having skewness=0, the graph in Fig. 4 shows symmetric and the model can be fit Optimized.

6.4 Normal Distribution of latitudes

The spread of latitude values are equally distributed having skewness=0, the graph Fig. 5 symmetric and the model can be fit Optimize.

6.5 Dataset link

The data set is publicly available and best suited to serve our purpose and fitted well on our algorithm: https://github.com/TusharRajVerma/GEOLOCATION/blob/master/world_country_and_usa_states_latitude_and_longitude_values.csv.

```
df.describe()
```

	latitude	longitude	usa_state_latitude	usa_state_longitude
count	244.000000	244.000000	52.000000	52.000000
mean	16.253109	13.294814	39.153437	-92.824719
std	27.031206	73.976477	6.967697	19.431647
min	-75.250973	-177.156097	18.220833	-155.665857
25%	-0.301710	-38.092008	35.438381	-102.009600
50%	16.869235	18.182149	39.435516	-89.093198
75%	38.965238	49.046734	43.220246	-78.291302
max	77.553604	179.414413	63.588753	-66.590149

Fig. 3: Screenshot of dataset

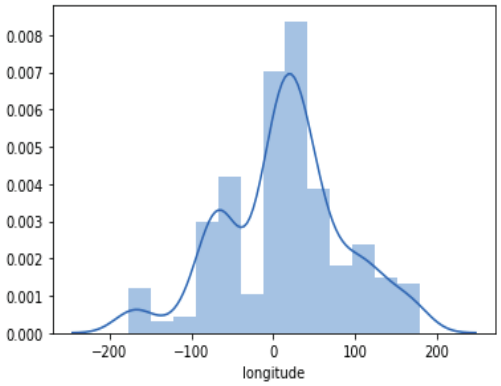

Fig. 4: Normal distribution of longitude

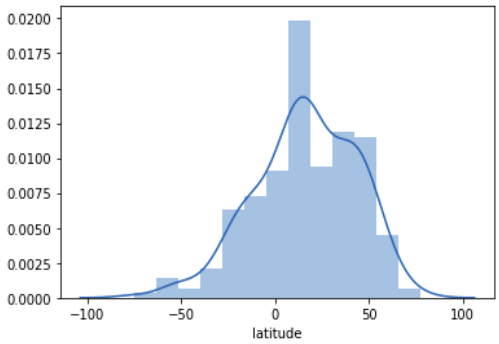

Fig. 5: Normal distribution of latitude

7. Allotment of Virtual Machines

Once the algorithm is able to find the optimal path after iterations become flat, we will allot the task [11, 24] to the respective clusters and the path will be formed. The clusters will form a path from source to destination as accuracy of our model is high, the path will overcome Minimum spanning tree, Bellmanford, Dijkstra algorithm as our algorithm works on cluster formation on geographic points on clusters of virtual machines. As geolocations are dynamically available it will become a test set and cases that will be satisfied.

7.1 Random cluster distribution for condition having cluster of virtual machines lack of resources

Once we have allocated a cluster to perform tasks and allocation is done using the main algo now we have developed an algorithm for satisfying this case also the conceptual study is to have a well refined cluster or virtual machine which can satisfy the resource allocation of the task allocated to it with lower latency.[4]The algorithm k+1 means solve the problem as:

7.2 Algorithms: K+1 means(x)

Step-1: For a iteration from i=k+1
Step-2: Initialize centroids and select i data points for centroid (μ_k).
Step-3: While the data points assignment becomes flat, keep iterating
Step 3.1: Call Objective function (X_i, μ_k)

Where, X_i = data points μ_k = centroid
Step 3.2: Assign each data point (X_i) to the closest cluster (μ_i) depending on value return from step 4.1.
Step 3.3: Compute centroids for the cluster by taking the average of all data points that belong to each cluster.
Step-4 Allot the tasks to the clusters so formed of geo locations
Step-5 If the cluster is not satisfying the resources needed then go to step-7
Step-6 Else we have found a compatible cluster we can end it
Step-7 Set i=i+1 and go to step-2

7.3 Machine Learning Application Flow

Here is step by step explanation of each step for inference.

As from data we will use the two columns latitude and longitude but there values needs to be preprocessed before feeding to clustering model

- First we have to apply a forward fill method to remove NaN values. The reason behind using the forward fill method is the data frame is sorted according to regions hence a forward fill can work for the nearest region.
- Now after removing NaN values let us now start considering about scaling and normalization of data the reason of doing this step is mainly due to high variation of values in latitude and longitude that needs to be scaled and then normalized.
- Applying principal component analysis is one of the best steps while working high variance data for dimensionality reduction on samples. This step really helps the values of latitude and longitude

This step of PCA is applied on normalized data coming from step-b

Figure 6 shows how meaningful our data of latitudes and longitudes looks:

	P1	P2
0	-0.740853	-0.567411
1	-0.744049	0.625422
2	-0.875043	0.385111
3	0.493621	-0.870045
4	0.456448	-0.886996

Fig 6.: Describes latitude and longitudes

As these values make sense now let us get back to the clustering part as we see in elbow function, we can take number of clusters=2 or 3. This experiment let's club the data with training and prediction with two clusters as described in the Fig. 7.

Now here is result of the grouping/ clustering of dataset with respect to optimal clusters found by our algorithm

This figure will have clear understanding of how we have grouped our data of longitudes and latitudes each row in two clusters such that the resources by virtual network embeddings can be Optimizely found with less latency

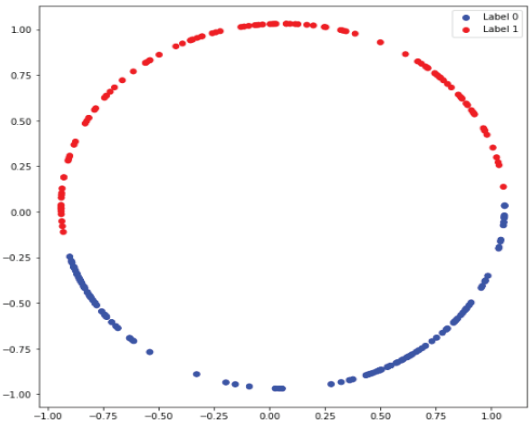

Fig. 7: Described the details of two clusters using K means

7.4 Inference Time

After all pre-processing steps on data the model should be given an optimal number of clusters for the virtual machine to get allocation on different regions. Advantages of using real time application

- During load balancing the virtual machine in a cluster label can request for a virtual machine within same cluster and it can be allocated if that virtual machine also cannot be allocated for some reason than it can try in some other cluster having our Euclidean distance algorithm in the use case.
  - During Auto scaling the same use case applies with virtual machine allocation
  - Due to our algorithm for nearest cluster or nearest virtual machine allocation the latency is also less

Here is a use case explanation flow diagram as described in Fig. 8.

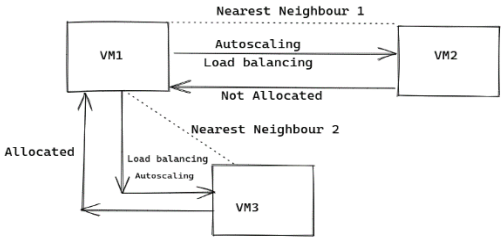

Fig. 8: Describes the flow with auto scaling and load balancing

Here a virtual machine-1 request virtual machine-2 for auto scaling or load balancing but it is not allocated for the nearest neighbour-1 but when virtual machine-1 request to virtual machine-3 which is the nearest neighbour-2 it got allocated. By this flow the idea of approach as well our implementation is demonstrated.

8. Experiment Results

The cluster metrics are so formed using algorithms and task allocations and running with low latencies. The tasks are being allocated for example word count of a document. The apps are submitted, running, completed and memory used. The complete cluster of virtual machines is as shown in Fig. 9.

9. Conclusion

The main components of the cloud environment are virtual machines. The key challenge is how these virtual machines are linked and in the case of failure. How will the migration of the loads should be done? As we prefer that the nearby location based on the distance as the ideal

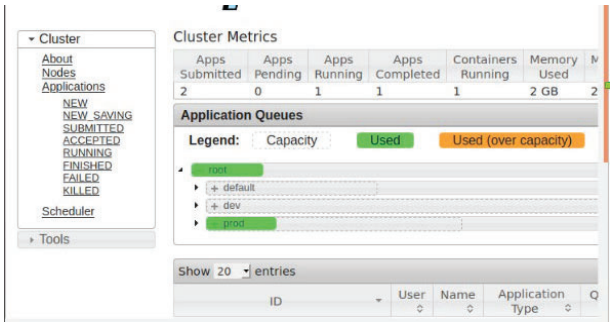

Fig. 9: Describes the details of various applications and VM

source of transferring the loads. However, there is the possibility that the resources available on the particular machine might not be sufficient for the computation. Therefore, there is the need of migrating the task to other VMs on which resources are available. In this approach machine learning based algorithm using K means clustering is proposed and has successfully took the decision for designing the cluster. Also helps in the migration based on the distances for the jobs in order to achieve optimal load balance. The cluster formed have placed the balance on the criteria that is the geographical location and also prevented the overloading of the certain system that are nearby, However during the study the power consumption was not considered as one of the constrains.

References

1) Likas A, Vlassis N, Verbeek JJ. The global k-means clustering algorithm." *Pattern recognition*. (2003) Feb 1;36(2):451-61.
2) Bradley PS, Fayyad UM. Refining initial points for k-means clustering. InICML 1998 Jul 24 (Vol. 98, pp. 91-99).
3) Kanungo T, Mount DM, Netanyahu NS, Piatko CD, Silverman R, Wu AY. An efficient k-means

clustering algorithm: Analysis and implementation. IEEE transactions on pattern analysis and machine intelligence. 2002 Aug 7; 24(7):881-92.

4) Zha H, He X, Ding C, Gu M, Simon HD. Spectral relaxation for k-means clustering. InAdvances in neural information processing systems 2002 (pp. 1057-1064).
5) Zhang W, Rajasekaran S, Wood T, Zhu M. Mimp: Deadline and interference aware scheduling of hadoop virtual machines. In2014 14th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing 2014 May 26 (pp. 394-403). IEEE.
6) Chavan V, Kaveri PR. Clustered virtual machines for higher availability of resources with improved scalability in cloud computing. In2014 First International Conference on Networks & Soft Computing (ICNSC2014) 2014 Aug 19 (pp. 221-225). IEEE.
7) Akula GS, Potluri A. Heuristics for migration with consolidation of ensembles of virtual machines. In2014 Sixth International Conference on Communication Systems and Networks (COMSNETS) 2014 Jan 6 (pp. 1-4). IEEE.
8) Beloglazov A, Buyya R. Energy efficient allocation of virtual machines in cloud data centers. In2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing 2010 May 17 (pp. 577-578). IEEE.
9) Xu G, Xu F, Ma H. Deploying and researching Hadoop in virtual machines. In2012 IEEE International Conference on Automation and Logistics 2012 Aug 15 (pp. 395-399). IEEE.
10) Dash P. Getting started with oracle vm virtualbox. Packt Publishing Ltd; 2013 Dec 12.
11) Mann ZÁ. Allocation of virtual machines in cloud data centers—a survey of problem models and optimization algorithms. Acm Computing Surveys (CSUR). 2015 Aug 10;48(1):1-34.
12) Chavan, V., & Kaveri, P. R., Clustered virtual machines for higher availability of resources with improved scalability in cloud computing. In 2014 First International Conference on Networks & Soft Computing (ICNSC2014), IEEE , pp. 221-225.
13) Rajabzadeh, M., Toroghi Haghighat, A., & Rahmani, A. M., New comprehensive model based on virtual clusters and absorbing Markov chains for energy-efficient virtual machine management in cloud computing. The Journal of Supercomputing, 2020, 76(9), 7438-7457.
14) Liu, X., & Liu, J. A truthful online mechanism for virtual machine provisioning and allocation in clouds. Cluster Computing, 2022, 1-15.
15 Aktan, M. N., & Bulut, H., Metaheuristic task scheduling algorithms for cloud computing environments. Concurrency and Computation: Practice and Experience, 2022, 34(9), e6513.
16) Asghari, A., Sohrabi, M.K. Combined use of coral reefs optimization and multi-agent deep Q-network for energy-aware resource provisioning in cloud data centers using DVFS technique. Cluster Comput 25, 119–140 (2022). https://doi.org/10.1007/s10586-021-03368-3
17) Raju, Y., & Devarakonda, N. A cluster medoid approach for cloud task scheduling. International Journal of Knowledge-based and Intelligent Engineering Systems, 2021, 25(1), 65-73.
18) Muthusamy, G., & Chandran, S. R. Cluster-based Task Scheduling Using K-Means Clustering for Load Balancing in Cloud Datacenters. Journal of Internet Technology, 2021., 22(1), 121-130.
19) Kiani, M., & Khayyambashi, M. R., A network-aware and power-efficient virtual machine placement scheme in cloud datacenters based on chemical reaction optimization. Computer Networks, 2021, 196, 108270.
20) Asensio, A., Masip-Bruin, X., Garcia, J., & Sánchez, S., On the optimality of Concurrent Container Clusters Scheduling over heterogeneous smart environments. Future Generation Computer Systems, 2021, 118, 157-169.
21) Yin, X., Zhang, K., Li, B., Sangaiah, A. K., & Wang, J., A task allocation strategy for complex applications in heterogeneous cluster–based wireless sensor networks. International Journal of Distributed Sensor Networks, 2018, 14(8), 1550147718795355.
22) Nisrina, Nora, et al. "The Effect of Genetic Algorithm Parameters Tuning for Route Optimization in Travelling Salesman Problem through General Full Factorial Design Analysis." *EVERGREEN Joint Journal of Novel Carbon Resource Sciences & Green Asia Strategy*. (2022): 163-203. doi.org/10.5109/4774233
23) Dwivedy, Bhoopendra, Anoop Kumar Bhola, and C. K. Jha. "Clustering Adaptive Elephant Herd Optimization Based Data Dissemination Protocol for VANETs." *EVERGREEN Joint Journal of Novel Carbon Resource Sciences & Green Asia Strategy*. (2021): 812-820. doi.org/10.5109/4742126
24) Khan, Shahroz Akhtar, et al. "A Perspective on Advances in Cloud-based Additive Manufacturing." *EVERGREEN Joint Journal of Novel Carbon Resource Sciences & Green Asia Strategy*. (2022): 861-869. doi.org/10.5109/4843119
25) Mishra, Arpana, et al. "Qualitative Analysis of Intra-Class and Inter-Class Clustering Routing and Clusterization in Wireless Sensor Network." *EVERGREEN Joint Journal of Novel Carbon Resource Sciences & Green Asia Strategy*. (2021): 358-373. doi.org/10.5109/4480718
26) Prasetyo, Hoedi. "On-grid photovoltaic system power monitoring based on open source and low-cost internet of things platform." *EVERGREEN Joint Journal of Novel Carbon Resource Sciences & Green Asia Strategy*. (2021): 98-106.

doi.org/10.5109/4372265
27) Huzaifi, Hanzalah, Arif Budiyanto, and Juanda Sirait. "Study on the carbon emission evaluation in a container port based on energy consumption data." *EVERGREEN Joint Journal of Novel Carbon Resource Sciences & Green Asia Strategy*. (2020): 97-103. doi.org/10.5109/2740964
28) Srivastava, Ashish Kumar, et al. "Statistical optimization by response surface methodology of process parameters during the CNC turning operation of hybrid metal matrix composite." *EVERGREEN Joint Journal of Novel Carbon Resource Sciences & Green Asia Strategy*. (2021): 51-62. doi.org/10.5109/4372260