Scholar Publishing Group
Distributed Processing System
https://doi.org/10.38007/DPS.2021.020301
ISSN 2790-0916 Vol. 2, Issue 3: 1-11

Distributed System Optimization Based on K-means Clustering

Bisen Mayanking*

University of New South Wales Sydney, Australia

**corresponding author*

Keywords: K-means Algorithm, Clustering Partition, Distributed System, System Optimization

Abstract: The regional energy internet is a distributed complex system with deep integration of energy and information. If different energy systems are planned and operated independently and lack of coordination with each other, problems such as low energy utilization rate, weak self-healing ability, and low system security and reliability will be caused. Therefore, scientific and reasonable planning methods and operation strategies are crucial to the overall efficiency and economy of distributed multi-energy systems. The main purpose of this paper is to study the optimization of distributed systems based on K-Means clustering. In this paper, a multi-energy unified clustering model based on the K-means algorithm is established to analyze the data characteristics in the cluster, and at the same time, the energy-side data is clustered to evaluate the available value of various types of energy. Experiments show that according to the 2/5/10 principle, when the system response time is within 2s, the user experience is very good. When the response time is between 2s and 5s, the user experience is better. Because the number of concurrent users of the system exceeds 500, the response time of the system is within the normal response time range acceptable to users and meets the needs of enterprises.

1. Introduction

The economy and society are developing rapidly, and people's living standards are improving steadily. my country pays more and more attention to the green sustainability of development, and the regional integrated energy system is an emerging energy supply method, which can realize the balance of energy supply and demand in the region. Its diversification and flexibility can successfully play the role of energy saving and emission reduction. However, the regional integrated energy system contains various types of energy-consuming loads, and the coupling between loads is complex, so the problem of system design difficulties needs to be solved urgently

[1-2].

In related research, Aski et al. believe that distributed applications can use Web services to transfer data. A neuro-fuzzy system including clustering is used to evaluate the trustworthiness of individual web services [3]. Nine criteria were considered, and eight neuro-fuzzy membership functions were considered using k-means clustering in order to obtain a neuro-fuzzy system with high prediction accuracy. The main goal is to evaluate the trustworthiness of individual web services using nine criteria. Top et al. proposed a parallel and distributed k-means clustering algorithm with naive patch centroid initialization for image segmentation [4]. Adopting the Message Passing Interface (MPI) standard to utilize the computing power of distributed computing nodes in a high-performance computing cluster, approximately 104 times the clustering time is achieved.

Aiming at the regional comprehensive energy system composed of complex loads, this paper proposes a partition optimization design method based on clustering algorithm and genetic algorithm. First, the paper establishes a multi-energy unified clustering model based on the K-means algorithm, analyzes the data characteristics in the cluster, such as thermoelectric ratio, time-varying characteristics, and load synchronization and complementarity, etc., and summarizes the energy consumption habits in different partitions. At the same time, the data on the energy side is clustered to evaluate the available value of various types of energy. Unified analysis of the available value of energy and load characteristics, through the development of unified evaluation criteria, select the appropriate energy supply equipment. On the premise of meeting the energy supply, the equipment is formed into an alternative structure set.

2. Design Research

2.1. Distributed Energy Clustering and Partitioning

After standardizing the original data of different indicators that constitute the clustering and partitioning optimization model, with the goal of minimizing transmission loss and balancing supply and demand [5-6], large-scale distributed energy resources and loads are clustered and partitioned to form a regional energy network. The main principles followed are as follows:

(1) Reduce the transmission loss between distributed energy and loads

Load moment is introduced as an important indicator to reflect line loss. Distributed energy resources and loads with smaller load moment should be divided into the same regional energy network as much as possible [7-8].

(2) Try to meet the supply and demand balance of distributed energy and load

The fact that the regional energy network can achieve a balance of supply and demand is a manifestation of the effective utilization of distributed energy. When clustering and partitioning, the total capacity and load demand of distributed energy divided into the same regional energy network should be as close to balance as possible [9-10] .

2.2. Distributed Model Predictive Control

As the scale of the system gradually increases and the structure becomes more complex, the disadvantages of the centralized control method become more obvious. In the development process of predictive control, in view of the shortcomings of centralized control methods, decentralized MPC (Model Predictive Control) was proposed [11-12]. Its core idea is to convert a single complex optimization problem into the solution of multiple subsystems, which has a simple structure and is easy to implement. However, the distributed MPC does not consider the coupling between the

subsystems and cannot reflect the interaction between the subsystems, so the control effect of the strongly coupled system is not good. On the basis of distributed MPC, distributed model predictive control (DMPC) considers the system coupling effect by strengthening the requirements of the communication network between subsystems, which not only reduces the difficulty of solving centralized control, but also fully considers the interconnection of the system. It has been widely used in many fields [13-14].

Figure 1 shows the specific structure of the DMPC [15-16].

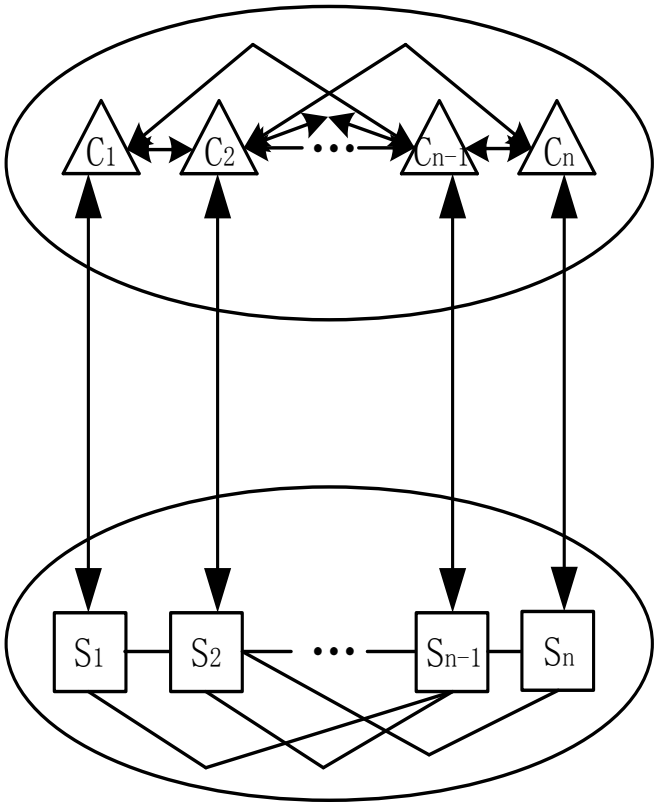

Figure 1. Schematic diagram of DMPC structure

2.3. K-means Algorithm

K-means-based multi-energy unified clustering method The core of the method of regional positive S partition optimization design lies in the cluster analysis of user load and energy conditions in the early stage. This step greatly reduces the workload of subsequent capacity allocation. Especially in the case of a large amount of data, the K-means algorithm is more efficient than other algorithms [17-18].

Specific steps are as follows:

Step 1: Select A: points as the initial cluster center;

Step 2 forms A clusters by assigning each point to its nearest cluster center, and recalculates the center of each cluster;

Step 3 Repeat step 2 until the cluster center does not change.

The clustering objects considered in this paper are various load data of various buildings in a certain area throughout the year, including various energy output information on the energy side and user energy consumption data on the load side. After data preprocessing, the data types on the user

side have become a large number of high-dimensional vectors. Therefore, the K-means algorithm is selected as the clustering method, which can accurately solve the problems caused by large numbers and high dimensions, and can control the running time within a reasonable range. Inside, the design efficiency is guaranteed. The calculation method has strong universality and can meet the distance measurement between general data points. The corresponding calculation formula is as follows.

$$d(x,y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2} \tag{1}$$

In the formula, X represents the cluster center point, and y represents each data point participating in the clustering. For determining the number of clusters, the root mean square standard deviation (RMSSTD) is an effective method to determine the optimal number of clusters, which can quantitatively evaluate the performance of the clustering model. The formula is as follows:

$$RMSSTD=\sqrt{\frac{\sum_{j=1\mathrm{K}\ d}\sum_{i=1\mathrm{K}\ k}^{n_{ij}}(x_q-\bar{x}_j)^2}{\sum_{j=1\mathrm{K}\ d}\sum_{i=1\mathrm{K}\ k}(n_{ij}-1)}} \tag{2}$$

In the formula, k represents the number of clusters, d represents the number of variables or the dimension of the data, nij represents the value of the jth dimension data of the ith cluster, and xj represents the average value of the jth dimension data.

3. Experimental Study

3.1. System Non-Functional Requirements

In order to allow users to use the system more safely and reliably, various indicators will be constrained when the system is running, because the system needs to process data under a large amount of data at the same time, and needs to ensure efficiency. Therefore, the system mainly has the following performance requirements:

(1) The amount of data processed concurrently

This system mainly performs automatic data classification for the data under the large amount of data, which should meet the parallel processing requirements under the large amount of data. Since the maximum upper limit that needs to be classified in an enterprise is generally less than 30,000, in order to ensure that the system can classify the enterprise, the system is required to process no less than 50,000 data concurrently.

(2) Response time

The system realizes automatic classification through the interaction between the front-end page and the back-end cluster. In order to ensure the service quality of the system and improve user satisfaction, the maximum response time of the system should not exceed 5 seconds, and the average response time of the system should be less than 2 seconds.

(3) Scalability

In order to ensure that the system can achieve load balancing, when the amount of data increases, the distributed system can realize distributed operation under a cluster composed of multiple machines by adding servers. And within a reasonable range, the operating efficiency of the system

continues to improve as the number of cluster nodes increases.

3.2. System Application Architecture

This paper designs a set of distributed systems, users can run distributed algorithms by setting the corresponding parameters. The user uploads the target set to be processed and sends a corresponding processing request, the system will process the data, and then return the final processing result to the user. Among them, the user only needs to participate in the data upload and result download process, and the specific implementation process does not require the user's participation. The distributed system adopts a browser/server mode (B/S) architecture, which is based on a wide area network. The network structure of this system is shown in Figure 2.

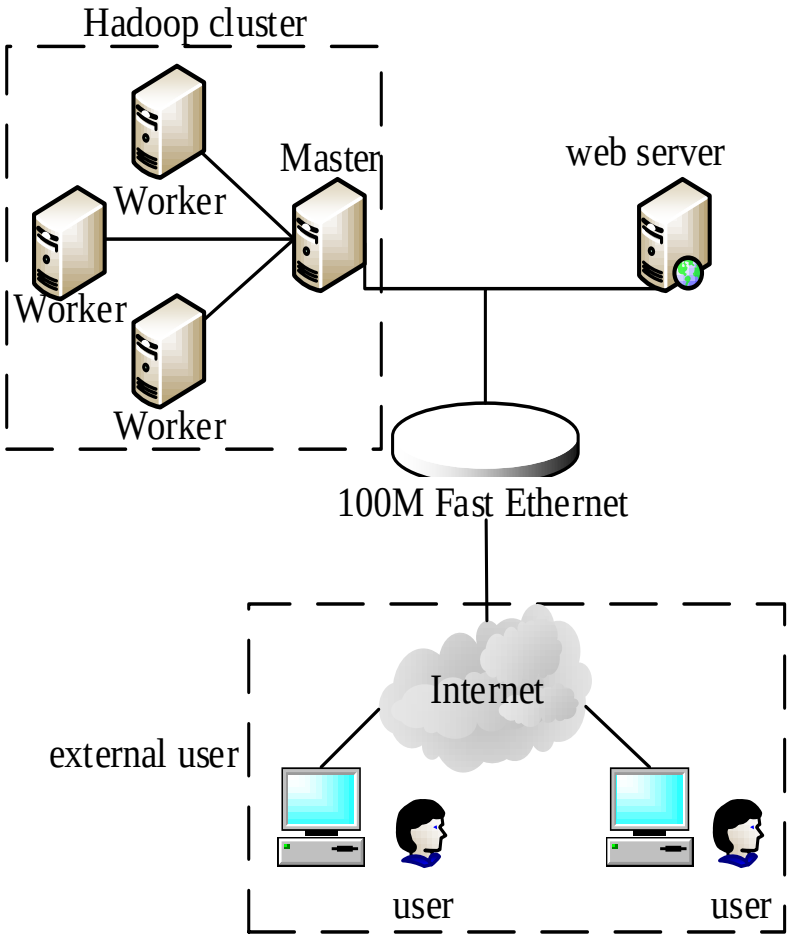

Figure 2. System network structure diagram

External users send service requests to the Internet through the client, and the web server receives the requests sent by the client through the switch and uploads them to the Hadoop cluster through the local network. Hadoop clusters process specific operations after receiving requests. A Hadoop cluster has a master node and other nodes are worker nodes. The master node is used to organize idle worker nodes for transaction processing, and the worker nodes are used to process Map Reduce jobs. After the Hadoop cluster finishes running the Map Reduce job, it outputs the results to the web server, and the web server sends the results back to the client.

The entire system is based on the Hadoop platform, with K-Means as the core algorithm of the cluster, and the Hadoop HDFS distributed file system as the storage space of the entire system. The system combines the Spring MVC framework, and separates the front-end and business-end thinking through layered design, thereby reducing the coupling of each module between the systems. The system is mainly divided into four layers: presentation layer, control layer, business logic layer and storage layer.

(1) Presentation layer

The presentation layer of the clustering system is mainly realized by the view module in the Spring MVC framework combined with the Java server page (JSP). The presentation layer is completely separate from the control layer, business logic layer and storage layer. It does not need to know the specific implementation of the business, but only needs to pass the user's page request to the control layer, and then receives and displays the views passed back. This system mainly includes data transmission page, preprocessing page, cluster analysis page and result download page.

The main function of the data transfer page is that the user selects the target set and uploads it to HDFS from the front page. The data uploaded to HDFS will be used for subsequent preprocessing and clustering process. The user operation preprocessing page performs word segmentation, stop word filtering, feature extraction, and vector building operations. The main function of the clustering analysis page is to select the operation mode and configure the corresponding parameters of the selected K-Means clustering algorithm or Canopy-K-Means clustering algorithm. After the user configures the parameters and clicks to run, the background will call the parallelized clustering analysis process based on the Hadoop platform. The main function of the result processing page is to check whether the clustering processing result is successful, and download the final clustering result.

(2) Control layer

The control layer is responsible for the interaction between the interface and the background. When the presentation layer sends a user request, the control layer directs the user request to the corresponding business logic. After the business logic layer processes the user request, it sends the processing result to the control layer, and the control layer forwards it to the presentation layer, thus completing the process of displaying the corresponding result to the user. In the Spring MVC framework, the Controller module is used to create the control layer. The DispatcherServlet view controller sends the view request to the corresponding controller. The controller invokes data source delivery, preprocessing, summary analysis, and result compilation at the business logic level. After the process is completed, the corresponding Model And View object is returned to the DispatcherServlet. Finally, the DispatcherServlet returns the output generated by the View on the first page.

(3) Business logic layer

The business of this system is implemented in Service, and the Service layer is mainly composed of data transmission, preprocessing, K-Means clustering, Canopy-K-Means clustering and result processing. The data transmission service implements the process of reading data information from the data source and uploading the data source to the HDFS space. The preprocessing business realizes data preprocessing. On the Hadoop platform, the Map Reduce programming model is used to realize the process of word segmentation, stop word filtering, feature dimension reduction, TF-IDF weight calculation and vector space generation. Cluster analysis business includes K-Means cluster analysis and Canopy-K-Means cluster analysis. The result processing business is mainly to check whether the clustering is running normally, package the clustering result into Zip, and then

pass it to the front-end page.

(4) Storage layer

This system mainly completes the data storage process through HDFS. The HDFS distributed file system provides an effective way for the distributed computing of this system. It can be deployed on inexpensive hardware resources and has high throughput. In the data source transmission phase, data is first transferred to HDFS through the front-end page. Subsequently, the data in the preprocessing and cluster analysis stages are stored on HDFS for distributed computing.

4. Experiment Analysis

4.1. Data Source Transmission Module Test

The data source transmission module mainly includes the selection of data sources and the upload of data sources, as well as the operations of adding, editing and deleting texts in the uploaded data sources. The test case description of the data source transmission module is shown in Table 1.

Table 1. Data source transport module test cases

	Test items	Test execution steps	Expected outcome	Actual results
1	Target text set selection	Click the "Choose File" button after the data source upload page; select the target text set	The file manager appears; the page displays the selected file path	Same as expected result
2	Target text set upload	Click the "Save" button	Upload the file and jump to the data source modification page	Same as expected result
3	Text addition	Click the "Add" button on the data source modification page; add text and save	The file manager appears; the file is added successfully, and the data source modification page is returned.	Same as expected result
4	Text Editor	Click the "Edit" button on the data source modification page; edit the text and save	The text modification page appears; if the text is modified successfully, return to the data source modification page	Same as expected result
5	Text deletion	Click the "Delete" button on the data source modification page; delete the text	The system pops up a prompt whether to delete, click "Yes" to delete, click "No" to not operate	Same as expected result

After testing, it is found that the data source transmission module test case is the same as the expected result

4.2. Clustering Result Processing Module

The clustering result processing module mainly includes test cases for processing result query and result download. The test case description of the clustering result processing module is shown

in Table 2.

Table 2. Clustering result processing module test cases

	Test item	Test execution steps	Expected outcome	Actual results
1	Process the result query	Check whether the final clustering is successful on the clustering result processing page	Display processing successfully	Same as expected result
2	Result download	"Click Result Download" button to download	Downloaded successfully	Same as expected result

After testing, it is found that the test case of the clustering result processing module is the same as the expected result

4.3. System Performance Test

In order to ensure the reliability and stability of the system, this paper will test the performance of the system through the system response time and the number of concurrent users. As an open source project, JMeter has a graphical user interface, is easy to learn and operate, and has good scalability, so we will use JMeter to test the system's concurrent access capability.

The method of stress testing of the system is: using JMeter to simulate different numbers of concurrent users sending requests for adding text and saving to the system. Because the system is mainly used for enterprise staff to extract and classify text data topics, in the actual application process, the number of concurrent users is rarely more than 500. Therefore, we define the maximum number of virtual users as 1000, which fully meets the needs of enterprises. For daily requirements, the results of JMeter's performance test on the system are shown in Table 3.

Table 3. System concurrent access capability test

Numbering	Number of virtual users	90% response time (ms)	Maximum response time (ms)	Error rate
1	50	97	156	0.00%
2	100	164	231	0.00%
3	200	255	462	0.00%
4	400	592	722	0.00%
5	800	1632	2067	0.00%
6	1000	2164	2910	0.00%

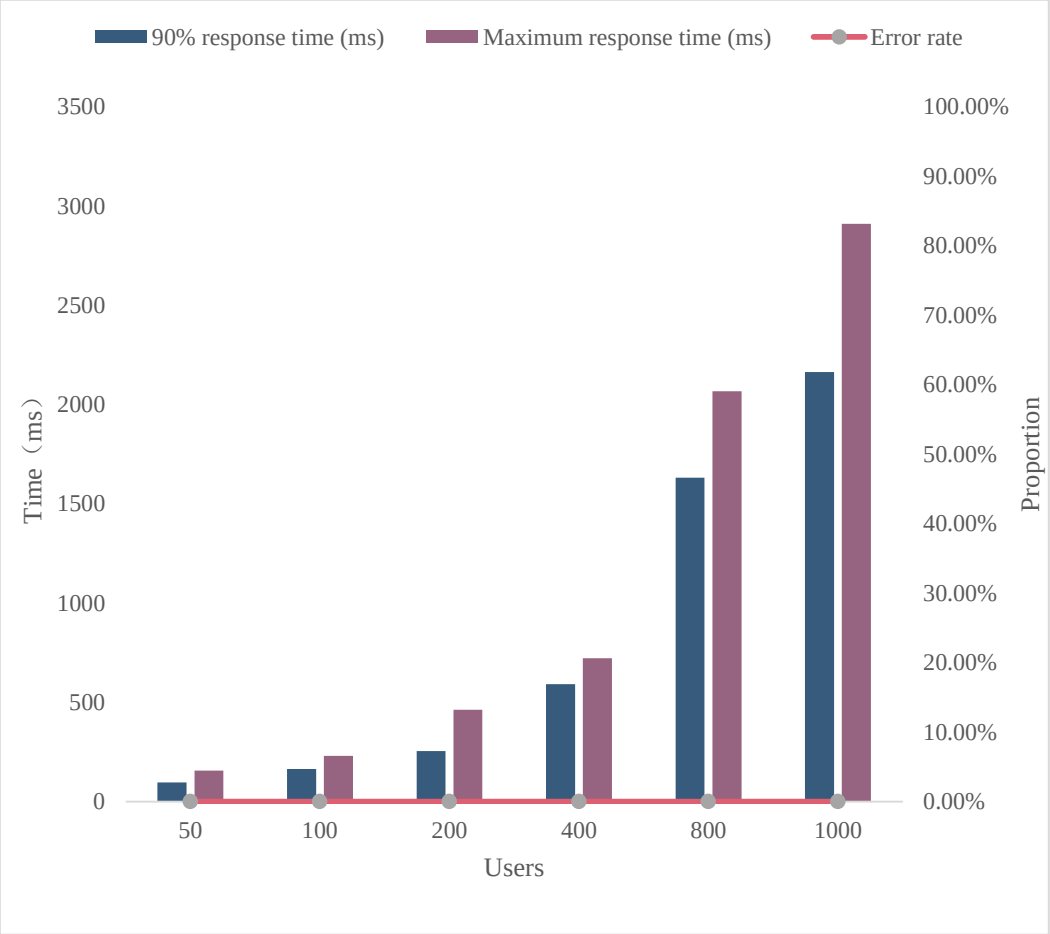

Figure 3. Analysis of system concurrent access capability test results

According to the 2/5/10 principle, when the system response time is within 2s, the user experience is very good. When the response time is between 2s and 5s, the user experience is better. Because the number of concurrent users of the system exceeds 500, the response time of the system is within the normal response time range acceptable to users and meets the needs of enterprises.

5. Conclusion

The regional energy internet is a distributed complex system with deep integration of energy and information. If different energy systems are planned and operated independently and lack of coordination with each other, problems such as low energy utilization rate, weak self-healing ability, and low system security and reliability will be caused. Therefore, scientific and reasonable planning methods and operation strategies are crucial to the overall efficiency and economy of distributed multi-energy systems. Data mining involves many aspects of knowledge, whether it is the preprocessing stage or the cluster analysis stage, each stage is worthy of in-depth research. With the further development of clustering technology, people can realize data mining more quickly and efficiently. In this paper, a distributed clustering system based on K-Means is designed, which can be used to quickly and efficiently cluster data under large amounts of data, thereby realizing automatic classification. Although the function and performance of this system have met expectations, there are still many deficiencies. For example, the clustering system is more suitable

for a large number of short clusters. If the single data is too large, the vector dimension will be too high, and the efficiency will drop significantly.

Funding

This article is not supported by any foundation.

Data Availability

Data sharing is not applicable to this article as no new data were created or analysed in this study.

Conflict of Interest

The author states that this article has no conflict of interest.

References

[1] Babu K R , Singal A , Sahiti K , et al. Performance Analysis Of Brain Tumor Detection Using Optimization Based Fcm Technique On Mri Images. International Journal of Scientific & Technology Research, 2020, 8(11):1717-1722.

[2] Kazama T , Umeki T , Shimizu S , et al. Over-30-dB gain and 1-dB noise figure phase-sensitive amplification using a pump-combiner-integrated fiber I/O PPLN module.. Optics express, 2021, 29(18):28824-28834. https://doi.org/10.1364/OE.434601

[3] Aski B S , Haghighat A T , Mohsenzadeh M . Evaluating single web service trust employing a three-level neuro-fuzzy system considering k-means clustering. Journal of Intelligent and Fuzzy Systems, 2021, 40(1):1-15. https://doi.org/10.3233/JIFS-201560

[4] Top A E , FÜ Torun, Kaya H . Parallel K-Means Clustering With Nave Sharding For Unsupervised Image Segmentation Via Mpi. Mühendislik Bilimleri ve Tasarım Dergisi, 2020, 8(3):791-798. https://doi.org/10.21923/jesd.748209

[5] Subbulaxmi M , Dr G A . Hybrid sampling Algorithm Based on Ant Colony Optimization and k-means Clustering. Indian Journal of Computer Science and Engineering, 2021, 12(2):445-455.

[6] Yadav S , Mohan R , Yadav P K . Task Allocation Model for Optimal System Cost Using Fuzzy C-Means Clustering Technique in Distributed System. Ingénierie des Systèmes D Information, 2020, 25(1):59-68. https://doi.org/10.18280/isi.250108

[7] Kuppan P , Shanmugam A , Ponnusamy S P , et al. K Means Algorithms For Cluster Analysis Using Machine Learning. Gedrag en Organisatie, 2020, 33(3):1594-15600. https://doi.org/10.37896/GOR33.04/074

[8] Ibrahim C , Mougharbel I , Kanaan H Y , et al. Industrial loads used as virtual resources for a cost-effective optimized power distribution. IEEE Access, 2020, 8(99):14901-14916.

[9] Choy K , Sin S , Tong Y , et al. Upper airway effective compliance during wakefulness and sleep in obese adolescents studied via two-dimensional dynamic MRI and semiautomated image segmentation.. Journal of applied physiology (Bethesda, Md. : 1985), 2021, 131(2):532-543.

[10] Ginting B , Riandari F . Implementasi Metode K-Means Clustering Dalam Pengelompokan Bibit Tanaman Kopi Arabika. Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI), 2020, 3(2):151-157. https://doi.org/10.32672/jnkti.v3i2.2381

[11] Goh Y L , Goh Y H , Yip C C , et al. Prediction of Students' Academic Performance by K-Means Clustering. Science Proceedings Series, 2020, 2(1):1-6. https://doi.org/10.31580/sps.v2i1.1205

[12] Carvalho L O , Ribeiro D B . Application of kernel k-means and kernel x-means clustering to obtain soil classes from cone penetration test data. Soils and Rocks, 2020, 43(4):607-618. https://doi.org/10.28927/SR.434607

[13] Afifi W , Nastiti D R , Aini Q . Clustering K-Means Pada Data Ekspor (Studi Kasus: Pt. Gaikindo). Simetris Jurnal Teknik Mesin Elektro dan Ilmu Komputer, 2020, 11(1):45-50.

[14] Yoseph F , Malim N , Heikkil M , et al. The impact of big data market segmentation using data mining and clustering techniques. Journal of Intelligent and Fuzzy Systems, 2020, 38(1):1-15. https://doi.org/10.3233/JIFS-179698

[15] Mitrentsis G , Lens H . Unsupervised learning method for clustering dynamic behavior in the context of power systems. IFAC-PapersOnLine, 2020, 53(2):13024-13029.

[16] Sabri I , Rashid A . Multi-Robot Localization System using an Array of LEDs and LDR Sensors. International Journal of Computer Applications, 2020, 176(10):9-12. https://doi.org/10.5120/ijca2020920001

[17] Alenazi M , Almutari A , Almowuena S , et al. NFV Provisioning in Large-Scale Distributed Networks with Minimum Delay. IEEE Access, 2020, PP(99):1-1.

[18] Sinambela Y , Herman S , Takwim A , et al. A Study Of Comparing Conceptual And Performance Of K-Means And Fuzzy C Means Algorithms (Clustering Method Of Data Mining) Of Consumer Segmentation. Jurnal Riset Informatika, 2020, 2(2):49-54. https://doi.org/10.34288/jri.v2i2.116