SCIENTIFIC OASIS

Decision Making: Applications in Management and Engineering

Journal homepage: www.dmame-journal.org
ISSN: 2560-6018, eISSN: 2620-0104

DECISION MAKING: APPLICATIONS IN MANAGEMENT AND ENGINEERING

Application of Big Data Technology Combined with Clustering Algorithm in Manufacturing Production Analysis System

Yu Liu[1], Zhengchao Zhang[2*], Shicao Jiang[1], Yunfei Ding[1]

[1] School of Economics & Management, Liaoning University of Technology, Jinzhou, 121000, China
[2] College of Economic, Bohai University, Jinzhou, 121000, China

ARTICLE INFO

Article history:
Received 15 August 2023
Received in revised form 13 November 2023
Accepted 20 November 2023
Available online 4 December 2023

Keywords: Clustering algorithm; Big data; Apriori; Production analysis; Genetic algorithm.

ABSTRACT

Production data analysis is crucial for production planning in the manufacturing industry, and accurate and a comprehensive and accurate analysis can improve production planning. To analyze the production of manufacturing industry, the research proposes the big data technology research method of the set clustering algorithm. In the process of this research method, the K-means clustering algorithm is first used to build the production measurement data system, and the Apache Hadoop Big data framework is applied. Then it introduces the Apriori algorithm for data association mining, and finally uses the genetic algorithm for production scheduling in view of big data analysis. The experiment showcases that the research method achieved a support level of 0.002 with 729 association rules when testing the number of association rules. When conducting data throughput testing, the research method achieved a data throughput of 7789 threads/s when the number of threads reached 8 in the Sleep scenario. In the analysis error testing, the error rate of the research method in the retained data fluctuates around 3.1%. When testing the number of processing operations in the process, the maximum error in the analysis results of the processing operations in the research method is 2. The results indicate that the research method possesses exceptional computational performance, carrying out manufacturing production analysis effectively and efficiently. The manufacturing production analysis system designed by the research institute offers valuable reference solutions for the informationization and intelligent development of the manufacturing industry.

1. Introduction

The manufacturing industry is one of the important pillars of the national economy, and its development is of great significance for economic growth and employment creation [1]. With the widespread application of information technology such as sensors and IoT devices in the manufacturing industry, manufacturing enterprises have generated a large amount of data, such as production process monitoring data, equipment operation data, quality inspection data, etc [2]. The data holds significant value, but its analysis encounters significant obstacles, encompassing

[*] *Corresponding author.*
E-mail address: bdglzzc@163.com

https://doi.org/10.31181/dmame712024897

diversity, scale, and real-time requirements [3]. The manufacturing data has a wide range of structured and unstructured data types, such as sensor and image data. This presents a challenging task to effectively merge and scrutinize these distinct data categories. The K-means algorithm can be used for data clustering, dividing manufacturing data into different categories, which helps to discover patterns and patterns in the data [4]. There are many variables and indicators involved in the manufacturing industry's production process, including complex correlation relationships and influencing factors. Traditional analysis methods often find it difficult to capture the inherent laws and key factors among them. The Apriori algorithm can mine association rules in data, help understand the relationships between different factors, and provide a reference for the optimization of manufacturing processes [5]. The manufacturing industry needs to ensure both output and product quality, and it is difficult to propose effective optimization strategies based on data analysis. Genetic algorithms (GA) offer a solution by optimizing problems, boosting production efficiency, and improving resource utilization through optimal solution search. In view of this, research attempts to combine K-means clustering algorithm (CA), Apriori algorithm, and GA in the context of big data technology. It designs a new manufacturing production analysis system to provide feasible technical references for the intelligent development of the manufacturing industry.

The research mainly focuses on four. The first discusses the current research on production analysis techniques and CA in the manufacturing industry. The second is the design of the manufacturing production big data analysis system combined with CA. The third is to test and empirically analyze the performance of the research method. The last section provides a comprehensive discussion and summary of the entire paper.

2. Related Works

In the context of the development of intelligent manufacturing concepts, data analysis in the manufacturing industry has become an important research topic in the industrial field, and production analysis is an important component of it. To provide a more accurate analysis of manufacturing production, some scholars have engaged in relevant research on manufacturing production analysis. Da Silva et al. proposed a method for analyzing manufacturing production in the context of green manufacturing. The goal was to improve manufacturing efficiency in lean manufacturing strategies. This method establishes new indicators in view of multi standard Data envelopment analysis in the process, and introduces the Levene test to statistically verify the program. The experiment showcases that the proposed method can effectively improve manufacturing efficiency [6]. Wang et al. proposed an intelligent prediction based analysis method for production performance analysis in the manufacturing industry. This method establishes a cloud-edge cooperative environment for resource distributed control during the process, and constructs an indicator analysis framework in view of colored Petri nets. The experiment showcases that the proposed method can effectively ensure the manufacturing production schedule [7]. Paul and Chowdhury proposed a production analysis based approach to the scheduling problem of manufacturing in the context of a global epidemic. This method uses mathematical modeling to analyze production requirements during the process, and makes plan adjustments in view of profits and production projects. The experiment showcases that the proposed method is feasible in ensuring manufacturer profits [8]. Scholars such as Zhou et al. proposed a method for analyzing manufacturing production to address the issue of modern intelligent development in the manufacturing industry. This method utilizes digital dual manufacturing units for knowledge driven processes, integrating intelligent perception, simulation, and prediction functions. The experiment showcases that the proposed method can effectively provide reference for manufacturing decision-making [9]. Scholars such as Oliveira have proposed a manufacturing production analysis method for problem search in the manufacturing process. This method uses a two-stage solution in the process and introduces the factor Sorting algorithm for data identification and diagnosis. The experiment showcases that the proposed method has good detection accuracy [10]. Their work contributes to the development of effective production analysis methods in the manufacturing industry.

CA can be used to handle complex data types in the big data of manufacturing production. Scholars have conducted relevant research on CA. Wang and other scholars proposed a CA based method for data mining problems. This method introduces a fuzzy membership matrix in the process and defines it in view of the geometric structure of the dataset and the degree of inter class separation, representing the optimal clustering partitioning results in the form of minimum values. The experiment showcases that the proposed method has good exponential adaptability [11]. Tang and other scholars proposed an improved method in view of CA for density based data clustering problems. This method sorts the data points and identifies the cluster structure during the process, searches for the boundary points of the augmented cluster order from an optical perspective, and determines the corresponding cluster's neighborhood radius. The experiment showcases that the proposed method has good Time complexity [12]. Dalmaijer et al. [13] proposed a recognition method in view of CA for the problem of identifying discrete subgroups of data in residual medicine. This method simulates the analysis results during the process, reconstructs the subgroup volume and covariance structure, and performs dimensionality reduction operations on the resulting dataset. The experiment showcases that the proposed method has good recognition accuracy and speed [13]. Zhang and other scholars proposed a classification method in view of CA for large-scale text information classification. In the process of this method, multiple texts to be classified are preprocessed, Vector space model is introduced to represent the text, and the classifier is constructed based on secondary Fuzzy clustering. The experiment showcases that the proposed method has a strong classification performance [14]. Yang and other scholars proposed a segmentation technique in view of CA for image segmentation of human brain magnetic resonance images. This method introduces local spatial information for data determination during the process, incorporates intuitive fuzzy thinking for noise processing, and combines local grayscale and spatial information. The experiment showcases that the proposed method has good segmentation effectiveness [15]. These scholars have conducted research on CA and their applications across various fields, influencing their respective domains. The summary of evaluated literature is shown in Table 1.

Table 1
Summary of evaluated literature

Serial Number	Document Name
6	Improving manufacturing cycle efficiency through new multiple criteria data envelopment analysis models: an application in green and lean manufacturing processes
7	A proactive manufacturing resources assignment method based on production performance prediction for the smart factory
8	A production recovery plan in manufacturing supply chains for a high-demand item during COVID-19
9	Knowledge-driven digital twin manufacturing cell towards intelligent manufacturing
10	On the influence of overlap in automatic root cause analysis in manufacturing
11	A new validity function of FCM clustering algorithm based on intra-class compactness and inter-class separation
12	An improved OPTICS clustering algorithm for discovering clusters with uneven densities

Serial Number	Document Name
13	Statistical power for cluster analysis
14	Text information classification method based on secondly fuzzy clustering algorithm
15	Noise robust intuitionistic fuzzy c-means clustering algorithm incorporating local information

In summary, although CA has been studied in multiple fields, there is still relatively little research on big data in the industrial field. Manufacturing production big data analysis systems also necessitate robust data processing technology for support. However, the research on CA in these other fields has demonstrated their good data processing performance, indicating the feasibility of applying CA to the analysis of big data in manufacturing production. In view of this, this research proposes a manufacturing production big data analysis system combined with CA, to offer additional technical references for the development of manufacturing industry.

3. Design of Big Data Analysis System for Manufacturing Production Combined with CA

The manufacturing production analysis system can provides enterprises with a reference to improve production quality. This section designs the technical means used in the research and design of the production big data analysis system. Firstly, the K-means CA was used to analyze production measurement data. Subsequently, the Apache Hadoop big data framework was designed as the operational framework of the analysis system. Then, the improved Apriori algorithm was used to mine association relationships from various data sets. Finally, the GA was applied to the production scheduling process based on big data. The manufacturing industry leverages multiple technical methods to analyze production data.

3.1 Production Measurement Data System in view of K-means Algorithm

With the development of intelligent manufacturing concepts, industrial manufacturing data, in addition to basic production data, has also begun to integrate task allocation information, fault information, warehousing information, testing information, etc. of manufacturing equipment into manufacturing data, resulting in a gradually large data volume [16-17]. When conducting data analysis, using manual analysis of large amounts of data can lead to problems such as long analysis time and insufficient analysis depth [18-19]. The existing manufacturing production big data analysis systems include Excel data processing macro technology based on Java language, resource visualization scheduling technology based on automatic guidance devices, and production data monitoring technology based on service-oriented architecture and a visual integration cloud platform. These methods can analyze and process some data in the operation of manufacturing enterprises, but most of them are limited by technical frameworks, leading to a limited amount of homogenous data being processed. Some can only provide assistance for the allocation of human and equipment resources in enterprises, while others can only analyze the data collected on the production line, posing a challenge in selecting effective data amidst multiple data types. In order to design a comprehensive analysis method for manufacturing production related data, the study chose to use a combination of multiple technical means to design a manufacturing production big data analysis system. The K-means algorithm belonged to an iterative clustering analysis algorithm with high computational efficiency when processing data. The research has introduced the K-means algorithm for the analysis of production measurement data. Many data in the manufacturing industry existed in discrete form, and the Apriori algorithm can mine the correlations in discrete data, helping to discover hidden patterns in various production data. When analyzing big data in the manufacturing industry, it is necessary to search for optimization methods for related problems. GA had strong flexibility and can maintain good operational performance when dealing with different types of problems. The research chose to combine K-means algorithm, Apriori algorithm, and GA to design a manufacturing production big data analysis system. When clustering, the cluster center calculation is shown in Equation (1).

$$u_i = \frac{1}{|C_i|}\sum_{x \in C_i} x \tag{1}$$

In Equation (1), u_i represents the cluster center, also known as the centroid. C represents the cluster. x represents the data sample. It calculates the distance between the points in the cluster and the cluster center, as shown in Equation (2).

$$E(C_k) = \sum_{x_i \in C_k} \|x_i - \mu_k\|^2 \tag{2}$$

In Equation (2), E represents the distance between the point and the cluster center. It calculates the distance sum in the dataset, as shown in Equation (3).

$$E(C) = \sum_{K=1}^{K}\sum_{x_i \in C_k} \|x_i - \mu_k\|^2 \tag{3}$$

In Equation (3), $E(C)$ represents the sum of distances. K represents the number of categories divided. During algorithm iteration, the values of distance and are continuously compressed until distance and cannot be further compressed. The K-means CA process is shown in Figure 1.

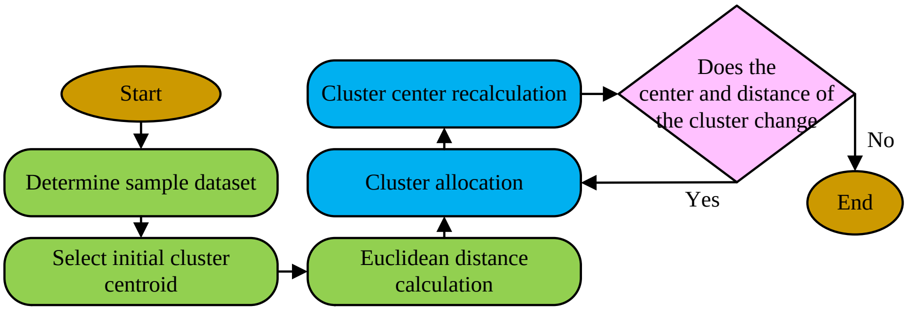

Start
Determine sample dataset
Select initial cluster centroid
Euclidean distance calculation
Cluster allocation
Cluster center recalculation
Does the center and distance of the cluster change
No
Yes
End

Fig. 1. K-means algorithm process

Figure 1 shows that before conducting calculations, it is necessary to first determine the sample dataset, and then select the initial clustering centroid in the sample dataset. It calculates the Euclidean distance between the initial clustering center and other data, assigns clusters to the data in view of the distance, and then calculates the center of the cluster. If the cluster's center and distance are altered, the data is reallocated within the cluster. It guides the end of the loop when the center and distance of the cluster remain unchanged, and outputs the result. The research is carried out in the Apache Hadoop Big data framework, which has distributed file systems and parallel clusters, providing robust load management capabilities. At the data transmission layer, it is implemented by Sqoop and Flume modules. Sqoop can transfer batch data from relational databases to distributed file systems, while Flume can collect logs generated by distributed platform tasks. The Kafka module operates simultaneously in the data storage layer and data transmission layer, allowing for the transmission and collection of extensive log data and supporting parallel data loading. In the data storage layer, the Hbase module also exists as a database, providing real-time read/write and random access functions for big data sets. In the data computing layer, there are modules such as Storm, Spark, Mahout, Hive, etc. The Storm module has high scalability and fault tolerance, which can reduce Hadoop's computing latency. The Spark module is a cluster computing engine. The Mahout module provides multiple machine learning algorithms to facilitate application development. The Hive module is used for storing large-scale data. Within the task scheduling layer, the Oozie module is used for task combination, followed by input into the logical work unit to achieve the processing of large tasks. The Zookeeper module belongs to a subsystem that lacks a data source layer. It provides corresponding service management and API provisioning when servers perform distributed tasks, simplifying application tasks [20-21]. To store large-scale data, the Hadoop distributed file system is introduced in the data storage layer, and the system architecture is shown in Figure 2.

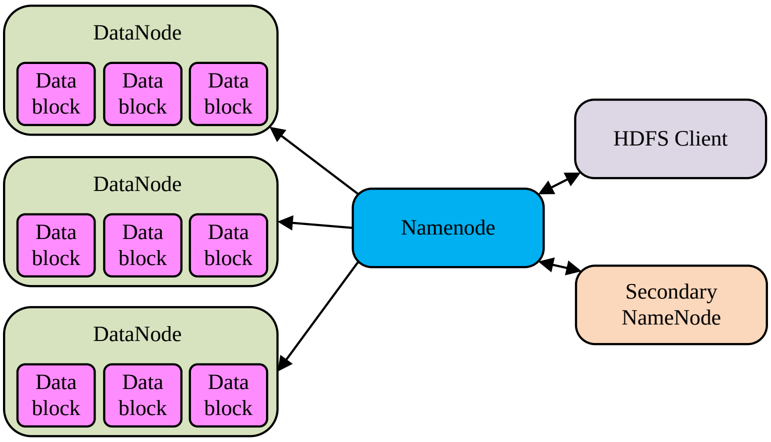


Fig. 2. Hadoop distributed file system architecture

Figure 2 shows that the Hadoop distributed file system consists of metadata management, auxiliary metadata management, data nodes, and clients. Metadata management is the only primary node in the system, serving as the management center and capable of managing the mapping information for data blocks. The client is responsible for generating relevant commands when accessing and managing the file system, which are executed by the data node. During system operation, the auxiliary metadata management is coordinated and monitored.

3.2 Production Big Data Analysis and Scheduling Technology Integrating Apriori Algorithm and GA

In the production process of the manufacturing industry, multiple data have interrelated relationships. To mine association relationships in data, the study introduces the Apriori algorithm [22]. The Apriori algorithm takes searching for frequent itemsets as its goal, and calculates support and confidence after randomly combining the itemsets. If the support and confidence are greater than or equal to the minimum value, the itemset will be judged as a frequent itemset [23-24]. The higher the support and confidence of the association rules of the two item sets, the more Strongly correlated material rule conditions are available for the two item sets, and the possibility of association rules is higher. The support calculation is shown in Equation (4).

$$\sup port\left(A => B\right) = P\left(A \cup B\right) \tag{4}$$

In Equation (4), A and B represent the itemset. P represents the probability of the occurrence of the union of itemsets. The confidence level calculation is shown in Equation (5).

$$confidence(A \Rightarrow B) = P(B|A) \tag{5}$$

In Equation (5), $B|A$ represents the probability ratio of the union of two itemsets to the A-itemset. When performing calculations, the selection criteria for the itemset are shown in Equation (6).

$$A \neq \phi, B \neq \phi, A \cap B = \phi \tag{6}$$

To improve the rigor of the algorithm, this study introduces an improvement rate to judge the relationship between frequent itemsets. The improvement rate formula is shown in Equation (7).

$$Up(A,B) = \left[\frac{P(A \cup B) P(\bar{A})}{P(\bar{A} \cup B) P(A)} - 1 \right] \times 100\% \tag{7}$$

In Equation (7), Up represents the improvement rate. The improved Apriori algorithm process obtained is shown in Figure 3.

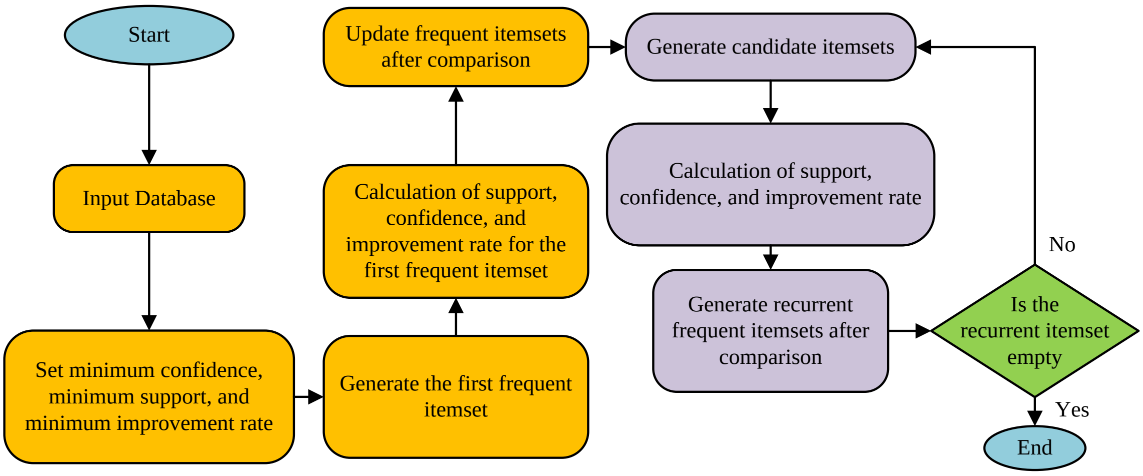

Start
Input Database
Set minimum confidence, minimum support, and minimum improvement rate
Generate the first frequent itemset
Calculation of support, confidence, and improvement rate for the first frequent itemset
Update frequent itemsets after comparison
Generate candidate itemsets
Calculation of support, confidence, and improvement rate
Generate recurrent frequent itemsets after comparison
Is the recurrent itemset empty
No
Yes
End

Fig. 3. Improving the apriori algorithm process

Figure 3 shows that after the algorithm starts running, it is first necessary to input the database, and then set the minimum confidence, minimum support, and minimum improvement rate. After generating the first frequent itemset, it calculates the support, confidence, and improvement rate of the first frequent itemset, and updates the frequent itemset through comparison. After generating candidate itemsets, their support, confidence, and improvement rates are calculated and compared to generate recurrent frequent itemsets. Then it continues to loop until the frequent items in the loop are integrated into an empty set, at which point the algorithm ends. In the production scheduling process of the manufacturing industry, discrete data and complex calculation are common problems. GA is used for production scheduling in view of big data. Using the roulette wheel method for chromosome selection, the sum of population fitness is first calculated, as shown in Equation (8).

$$SUM = \sum_{i=1}^{pop-size} f(x_i) \tag{8}$$

In Equation (8), SUM represents the sum of fitness. f represents the individual fitness of chromosomes. It calculates the selection probability of a single chromosome, as shown in Equation (9).

$$p_i = f\left(x_i\right) / SUM \tag{9}$$

In Equation (9), p_i represents the selection probability of a single chromosome. It calculates the intermediate cumulative probability of each chromosome, as shown in Equation (10).

$$q_i = \sum_{j=1}^{i} p_j \tag{10}$$

In Equation (10), q_i represents the intermediate cumulative probability of chromosomes. Afterwards, it places the chromosomes in a buffer until the buffer is filled. The operation of recombining chromosomes is called a crossover operation, which primarily considers the position and content of the crossover. There are three methods for performing crossover operations: single point crossover, multi-point crossover, and uniform crossover, as shown in Figure 4.

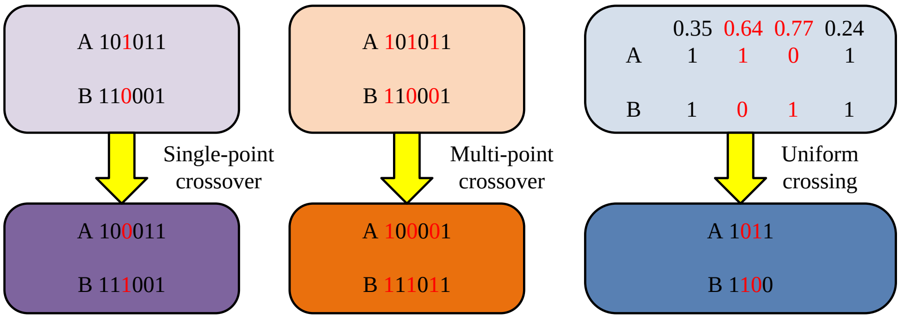


Fig. 4. Cross operation

Figure 4 shows that during single point crossing, only two gene points of two chromosomes are exchanged, which results in new chromosomes. When conducting multipoint crossing, multiple gene points in two chromosomes are exchanged to produce new chromosomes. When conducting uniform crossover, it is necessary to set a minimum crossover probability and generate the crossover probability in gene order. If the crossover probability is smaller than the set minimum crossover probability, the corresponding gene points will be exchanged. The mutation mode affects the local search capability of GA, and in binary encoding, there are two mutation modes: basic bit mutation and binary mutation [25]. The overall framework of GA is shown in Figure 5.

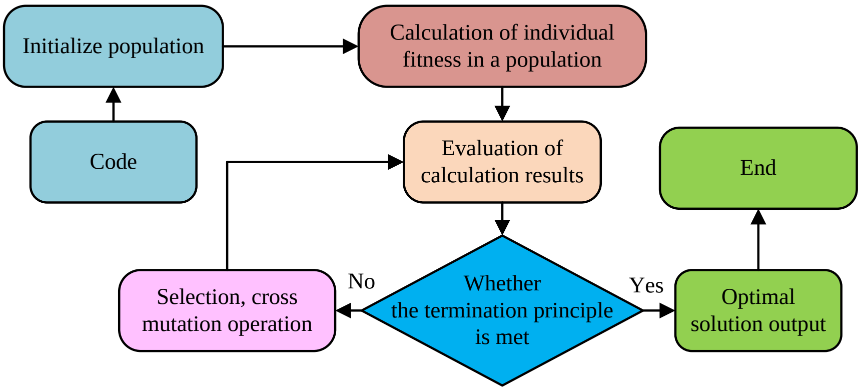

Initialize population
Calculation of individual fitness in a population
Code
Evaluation of calculation results
End
Selection, cross mutation operation
No
Whether the termination principle is met
Yes
Optimal solution output

Fig. 5. General framework of GA

Figure 5 shows that before running the algorithm, it is necessary to first select an appropriate encoding method for the problem and then initialize the population. After completing the population initialization, use the fitness function to calculate the fitness of individuals in the population and evaluate the calculation results. If the result does not meet the termination principle, then select, cross, and mutate before conducting the fitness calculations and evaluations again. It continuously loops until the calculation results meet the termination principle, outputting the optimal solution and ending the algorithm. This study proposes a matrix encoding method to enhance the production scheduling process that uses manufacturing processes and equipment as the matrix elements. The construction of the encoding matrix is shown in Equation (11).

$$T_{n\times m}=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix} \tag{11}$$

In Equation (11), T represents the matrix. n represents the processing task to be completed. m represents the process required for processing. The number and calculation of parallel machines in processing are shown in Equation (12).

$$R=\sum_{j=1}^{r}R_j,R_0=0 \tag{12}$$

In Equation (12), R represents the number of parallel machines. The range of values for the elements of the encoding matrix is shown in Equation (13).

$$a_{ij}\in\left[1+\sum_{e=0}^{j-1}R_e,1+\sum_{e=0}^{j}R_e\right] \tag{13}$$

In Equation (13), the upper limits of i and j are n and m, respectively. Treating the coding matrix as a chromosome reduces the difficulty and workload of coding. When performing selection operations, it merges the new and old populations, compares individuals, and selects more excellent individuals to retain to obtain a standard size population. The objective function of an individual is shown in Equation (14).

$$F_f = \max_{1 \le i \le n}\left(C_{im}\right) \tag{14}$$

In Equation (14), F_f represents the objective function. C_{im} represents the completion time. The fitness function is shown in Equation (15).

$$fit = C_{\max} = F_f \tag{15}$$

In Equation (15), fit represents the fitness function. The fitness function and function are interrelated, and the resulting scheduling calculation method is shown in Figure 6.

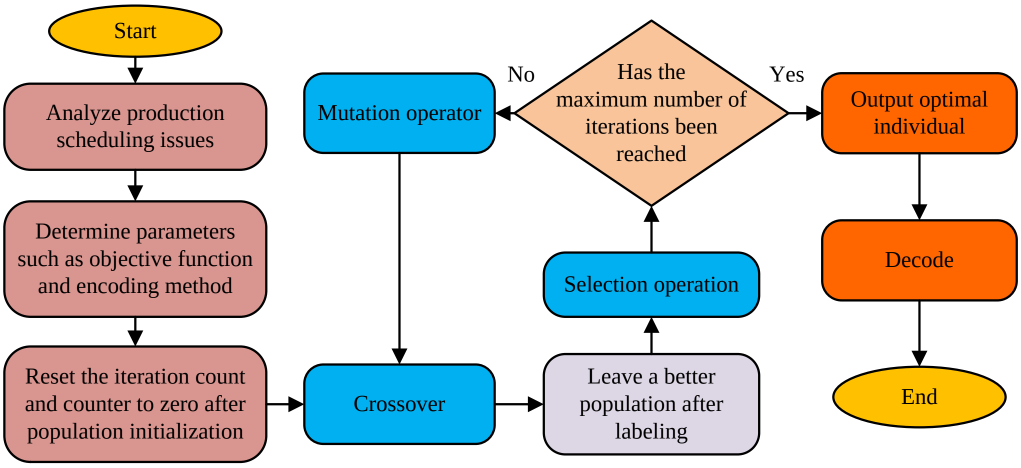

Start
Analyze production scheduling issues
Determine parameters such as objective function and encoding method
Reset the iteration count and counter to zero after population initialization
Mutation operator
Crossover
No
Has the maximum number of iterations been reached
Yes
Selection operation
Leave a better population after labeling
Output optimal individual
Decode
End

Fig. 6. Production scheduling calculation method process

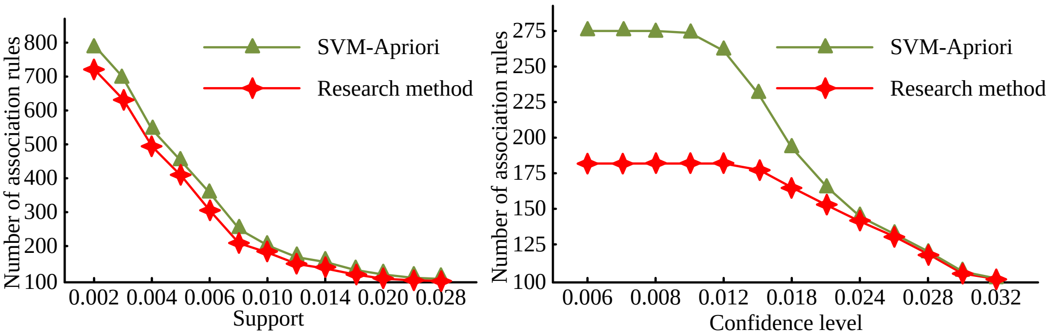
Figure 6 shows that when scheduling production in view of big data of manufacturing production, first analyze the scheduling problem, and then determine the objective function, coding method and other parameters. After population initialization, the iteration count and counter are reset to zero, and the cross operation is initiated to produce a new population. The process marks the optimal individual and retains a higher quality population. It continuously iterates until the maximum number of iterations is reached, outputting and decoding the optimal individual. When the manufacturing production big data analysis system constructed is running, the Apache Hadoop big data framework implements distributed storage of data, facilitating data preprocessing for the system. After completing data preprocessing, the K-means algorithm extracts features and converts the data into a form that can be processed by the algorithm. After mining association rules and anomalies in data using the Apriori algorithm, GA are used to calculate relevant optimization techniques and processing methods that can maximize the utility of big data analysis systems.

4. Performance Test and Application Analysis of Manufacturing Production Big Data Analysis System Combined with CA

Effective analysis of big data in production can provide an objective overview of the current situation. This section will test the performance of the research method, and analyze the application of the research method with actual data to determine the effectiveness of the research method.

4.1 Performance Test of Manufacturing Production Big Data Analysis System Combined with CA

To analyze the effectiveness of the big data analysis system in manufacturing production, the performance test and application analysis of the research method are carried out. When conducting performance testing on research methods, it uses datasets containing manufacturing production losses, capacity, and electricity for testing. The testing process examines the number of association rules generated by the research method at various support and confidence levels, with comparisons made to the Support Vector Machine Apriori Algorithm (SVMAA), as shown in Figure 7.

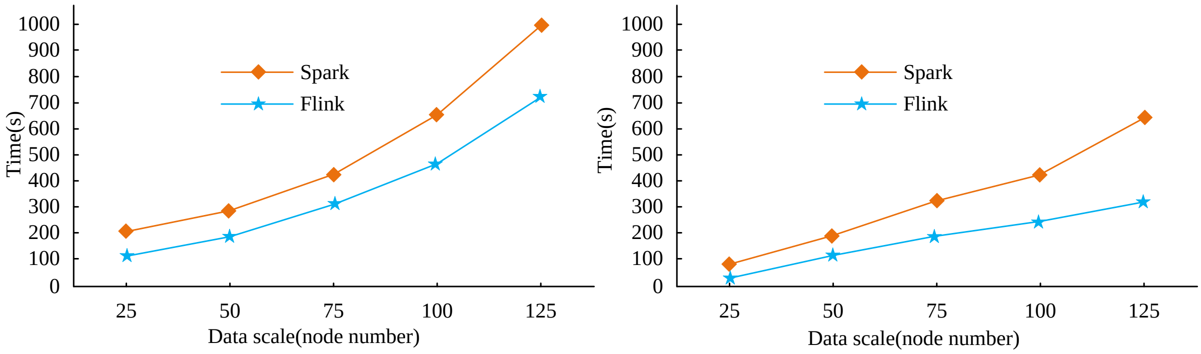


(a)Different levels of support

(b)Different levels of confidence

Fig. 7. Testing the number of association rules

Figure 7 (a) shows that the number of association rules for the SVMAA is 788 with a support level of 0.002. The number of association rules with a support level of 0.028 is 123. The number of association rules in the research method with a support level of 0.002 is 729. The number of association rules with a support level of 0.028 is 119. Figure 7 (b) shows that the number of association rules in the SVMAA is 275 at a confidence level of 0.006, and it begins to decrease at a confidence level of 0.008. The number of association rules at a confidence level of 0.032 is 106. The number of association rules in the research method is 182 at a confidence level of 0.006, and it begins to decrease at a confidence level of 0.012. The number of association rules at a confidence level of 0.032 is 104. This research method exhibits improved association rule conciseness in low confidence and low support situations. It tests the calculation time of the research method under changes in data size, as shown in Figure 8.



(a)SVM-Apriori

(b)Research method

Fig. 8. Scalability testing

Figure 8 shows that the processing time of the SVMAA and research methods increases continuously with the increase of data size under different frameworks. The processing time of the SVMAA using the Spark framework is 203 seconds when the number of data scale nodes is 25. When the data scale reaches 125 nodes, it increases to 992s. The processing time for the Spark framework used in the research method is 82 seconds when the number of data scale nodes is 25. When the data scale reaches 125 nodes, it increases to 634 seconds. The processing time of the Vector Machine Apriori algorithm using the Flink framework increases to 721 seconds when the data scale reaches 125 nodes. The processing time of the Flink framework used in the research method is 23 seconds when the number of data scale nodes is 25. When the data scale reaches 125 nodes, it increases to 318s. With an increase in data size, the processing time increases more slowly and the processing time is shorter, indicating that the framework has better scalability. It tests the data throughput of the research method, as shown in Figure 9.

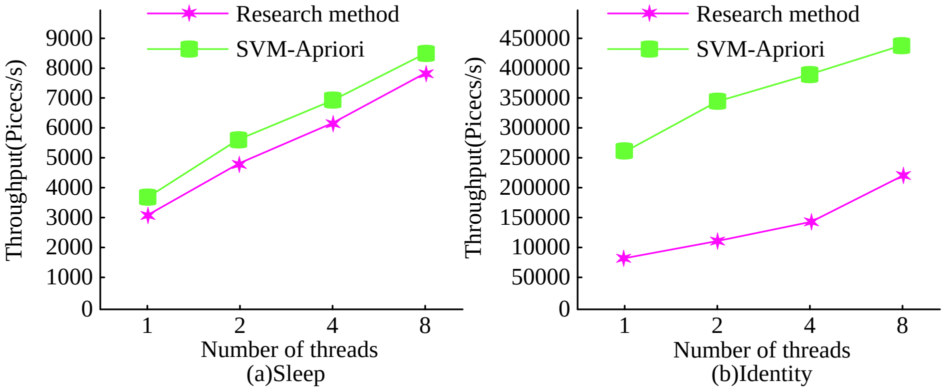


Fig. 9. Throughput

Figure 9 shows that in both the simpler Sleep scenario and the more complex Identity scenario, the data throughput of both methods increases with the number of core threads. In the Sleep scenario, the SVMAA has a data throughput of 3734 threads/s when the number of threads is 1. When the number of threads reaches 8, the data throughput increases to 8475 threads/s. The research method has a data throughput of 3067 threads/s when the number of threads is 1. When the number of threads reaches 8, the data throughput increases to 7789 threads/s. In the Identity scenario, the SVMAA has a data throughput of 261243 threads/s when the number of threads is 1. When the number of threads reaches 8, the data throughput increases to 439742 threads/s. The research method has a data throughput of 77843 pieces/s when the number of online processes is 1. When the number of threads reaches 8, the data throughput increases to 224360 threads/s. This indicates that the research method has better computational simplicity and can reduce the hardware burden. The analysis error of the research method is tested, and the grey correlation analysis method designed by scholars such as Chen et al. is introduced for comparison [26]. As shown in Figure 10.

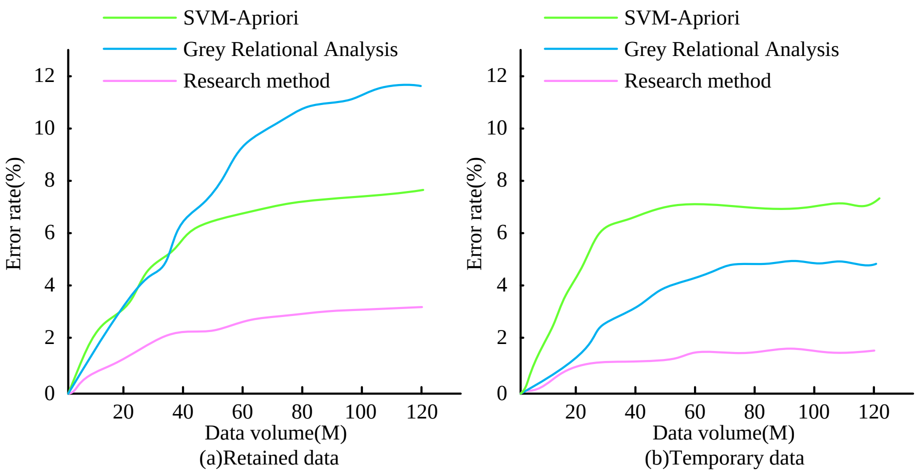


Fig. 10. Data error rate

Figure 10 shows that the data error rates of different methods tend to stabilize after rising to a certain value. When preserving data, the data error rate of the SVMAA tends to stabilize when the data volume reaches about 176M, fluctuating around 7.6%. The data error rate for the grey correlation analysis method tends to stabilize when the data volume reaches about 104M, fluctuating around 11.7%. The data error rate for the research method tends to stabilize when the data volume reaches about 57M, fluctuating around 3.1%. In temporary data, the data error rate for the SVMAA tends to stabilize when the data volume reaches about 46M, fluctuating around 7.1%. The data error rate for the grey correlation analysis method tends to stabilize when the data volume reaches about 72M, fluctuating around 4.8%. The data error rate for the research method tends to stabilize when the data volume reaches about 24M, fluctuating around 1.6%. This indicates that the research method has good analytical accuracy.

4.2 Application Analysis of Manufacturing Production Big Data Analysis System Combined with CA

The actual data of a manufacturing plant is used to analyze the application of the research method. It is tested in view of the number of processing operations, as shown in Figure 11.

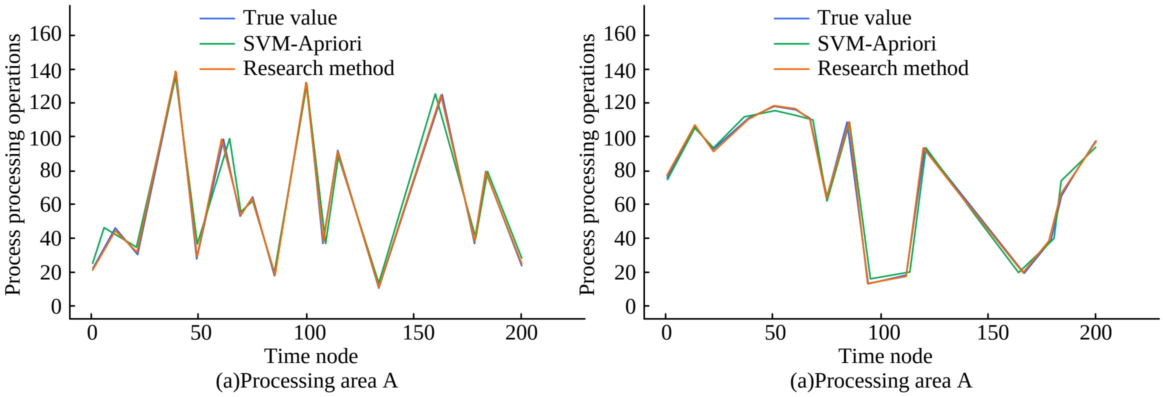


Fig. 11. Analysis of process processing operations

Figure 11 shows that both the SVMAA and research methods have effectively analyzed the number of processing operations and the time nodes involved. In the processing area A, the SVMAA allows a maximum analysis error of 4 for the number of processing operations in the process, and a maximum analysis error of 7 nodes at the corresponding time node. The maximum error in the analysis results of the number of processing operations in the research method is 2, and the maximum error in the analysis at the time node is 1 node. In the processing area B, the SVMAA allows for a maximum analysis error of 8 for the number of processing operations in the process, and a maximum analysis error of 5 nodes at the corresponding time node. The maximum error in the analysis results for the number of processing operations in the research method is 2, and the maximum error in the analysis at the time node is 1 node. This indicates that the research method is more accurate in analyzing the actual production situation. The processing time for three products was optimized and simulated, and the results are shown in Table 2.

Table 2
Optimization of product processing time

Product	Method	1	2	3	4	5	6	7	8	9	10
1#	SA(s)	270	268	272	274	270	269	273	271	267	270
	SVM-Apriori(s)	272	271	278	269	275	271	269	274	277	271
	Research method(s)	267	264	266	264	264	266	262	267	261	263
2#	SA(s)	98	96	101	97	99	101	102	97	98	99
	SVM-Apriori(s)	98	97	99	96	95	98	96	97	95	96
	Research method(s)	91	89	93	92	91	90	94	93	91	92
3#	SA(s)	564	567	569	567	565	566	571	573	569	568
	SVM-Apriori(s)	572	574	575	574	576	572	576	573	574	575
	Research method(s)	553	554	551	556	552	553	551	552	554	555

Table 2 shows that all three methods have generated time optimization results for the product. In product 1 #, the result range of the simulated degradation algorithm is between 268s and 274s. The result interval of the SVMAA ranges from 269s to 278s. The result range of the research method ranges from 261s to 267s. In product 2 #, the result range of the simulated degradation algorithm is between 96s and 102s. The result interval of the SVMAA ranges from 95s to 99s. The result range of the research method ranges from 89s to 94s. In product 3 #, the result range of the simulated degradation algorithm is between 564s and 573s. The result interval of the SVMAA ranges from 572s to 576s. The result range of the research method ranges from 551s to 556s. This indicates that the research method can effectively optimize product processing time to a greater extent.

4.3 Discussion on Performance Testing and Application Analysis Results

The results obtained during the performance testing and application analysis of the research method have been screened and abnormal data and non representative data content have been removed. When conducting system performance testing, multiple experiments have been conducted to verify that the results obtained are consistent, indicating that the performance test results obtained are accurate and effective. When conducting application analysis, the results are unique due to factors such as equipment wear and personnel changes in the manufacturing production process. However, the combination of application analysis results and performance testing results can indicate that the research results are effective and reliable. The research method has a lower number of association rules at runtime under different confidence and support levels, and has lower data throughput during computation, indicating that the research method can complete the analysis with a smaller amount of calculation methods and data at runtime. The research method reduces data processing time as the scale node increases from 25 to 125. Additionally, it maintains a lower data error rate while processing data volume, thus indicating the capability of balancing speed and accuracy when analyzing big data in manufacturing production. The research method can analyze the processing processes and time nodes during production with lower errors, and can generate better product and processing time optimization plans. The manufacturing production big data analysis system designed by the research institute has good operational performance. The research method of using big data technology for manufacturing production analysis can reduce the occurrence of accidental situations caused by limited data volume in the analysis results, which is beneficial for the management of manufacturing enterprise managers.

5. Conclusion

Analyzing manufacturing production can improve the production efficiency of the manufacturing industry. This study suggests a big data analysis system for complex manufacturing production data, using set CA. During the process, the K-means algorithm is first used for production measurement data analysis, the Apriori algorithm is used to search for association rules, and the improvement rate is introduced for frequent itemset relationship judgment. Afterwards, it used matrix encoding to optimize the GA, and finally analyzed the effectiveness of the research method. The experiment showcases that when testing the number of association rules, the number of association rules in the research method with a confidence level of 0.006 is 182. When conducting calculation time testing, the research method achieved a calculation time of 634s and 318s when the data scale nodes reached 125 in two different frameworks. During the data throughput testing, the research method achieved a data throughput of 224360 threads/s when the number of threads reached 8 in the Identity scenario. In the analysis error test, the data error rate of the research method in temporary data tends to stabilize when the data volume reaches about 24M, fluctuating around 1.6%, which is lower than other methods. In the analysis of operation number, the maximum error in the nodal analysis of the time of two processing areas is 1 node. When conducting processing time optimization simulation, the optimization results of the research method are superior to other optimization methods. This indicates that the research method can effectively conduct production analysis in the manufacturing industry, resulting in analysis results with high validity and enabling better operational strategies. And the research method has a more concise data volume at runtime, reducing the model's running burden and shortening calculation time. Research methods can be utilized to analyze the production quality of manufacturing enterprises by calculating crucial parameters in the production process, promptly detecting product quality anomalies, and enhancing the product qualification rates. It is also possible to analyze the sensor data of production equipment to identify inefficient links and limiting bottlenecks in the production process, thereby improving the production and manufacturing efficiency of the enterprise. In addition, it can also analyze the energy and material consumption during the operation of the enterprise, providing reference information for enterprise planning. In future development, automation control systems can be combined to achieve more efficient management of enterprises, lessening operational stress on enterprises amidst high societal demand. However, the research mainly focuses on mechanical manufacturing enterprises when designing methods, and also selects mechanical manufacturing enterprises for application analysis. The application effect in manufacturing enterprises such as light textile industry and manual manufacturing may not be satisfactory. In subsequent experiments, the optimization will be carried out based on the characteristics of other types of manufacturing enterprises, and the experimental scope will be expanded to optimize the method and improve its applicability.

Author Contributions

Conceptualization, Y.L., Z.Z., S.J. and Y.D.; methodology, Y.L., Z.Z., S.J. and Y.D.; software, Y.L., Z.Z., S.J. and Y.D.; validation, Y.L. and Y.D.; formal analysis, Y.L. and S.J.; investigation, Y.L., Z.Z., S.J. and Y.D.; writing—original draft preparation, Y.L., Z.Z., S.J. and Y.D.; writing—review and editing, Y.L., Z.Z., S.J. and Y.D.; visualization, Z.Z. and Y.D.; supervision, Y.D. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Data Availability Statement

Data sharing is not applicable to this article.

Conflicts of Interest

The author declare that have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The research is supported by Ministry of Education now project of research of humanitarian company division, research on the location and optimization of manufacturing enterprise process bottleneck cycle under rational cost constraint (Project No. 18YJC630256).

References

[1] Jamwal, A., Agrawal, R., Sharma, M., & Kumar, V. (2021). Review on multi-criteria decision analysis in sustainable manufacturing decision making. *International Journal of Sustainable Engineering*, 14(3), 202-225. https://doi.org/10.1080/19397038.2020.1866708.

[2] Calzavara, M., Battini, D., Bogataj, D., Sgarbossa, F., & Zennaro, I. (2020). Ageing workforce management in manufacturing systems: state of the art and future research agenda. *International Journal of Production Research*, 58(3), 729-747. https://doi.org/10.1080/00207543.2019.1600759.

[3] Babubudjnauth, A., & Seetanah, B. (2021). An empirical analysis of the impacts of real exchange rate on GDP, manufacturing output, and services sector in Mauritius. *International Journal of Finance & Economics*, 26(2), 1657-1669. https://doi.org/10.1002/ijfe.1869.

[4] Chen, D. J. I. Z. (2021). Automatic vehicle license plate detection using K-means clustering algorithm and CNN. *Journal of Electrical Engineering and Automation*, 3(1), 15-23. https://doi.org/10.36548/jeea.2021.1.002.

[5] Meena, N., & Singh, B. (2021). Firefly optimization based hierarchical clustering algorithm in wireless sensor network. *Journal of Discrete Mathematical Sciences and Cryptography*, 24(6), 1717-1725. https://doi.org/10.1080/09720529.2021.1880146.

[6] Silva, A. F., Marins, F. A. S., Dias, E. X., & Ushizima, C. A. (2021). Improving manufacturing cycle efficiency through new multiple criteria data envelopment analysis models: An application in green and lean manufacturing processes. *Production Planning & Control*, 32(2), 104-120. https://doi.org/10.1080/09537287.2020.1713413.

[7] Wang, W., Zhang, Y., Gu, J., & Wang, J. (2021). A proactive manufacturing resources assignment method based on production performance prediction for the smart factory. *IEEE Transactions on Industrial Informatics*, 18(1), 46-55. https://doi.org/10.1109/TII.2021.3073404.

[8] Paul, S. K., & Chowdhury, P. (2021). A production recovery plan in manufacturing supply chains for a high-demand item during COVID-19. *International Journal of Physical Distribution & Logistics Management*, 51(2), 104-125. https://doi.org/10.1108/IJPDLM-04-2020-0127.

[9] Zhou, G., Zhang, C., Li, Z., Ding, K., & Wang, C. (2020). Knowledge-driven digital twin manufacturing cell towards intelligent manufacturing. *International Journal of Production Research*, 58(4), 1034-1051. https://doi.org/10.1080/00207543.2019.1607978.

[10] Oliveira, E., Miguéis, V. L., & Borges, J. L. (2022). On the influence of overlap in automatic root cause analysis in manufacturing. *International Journal of Production Research*, 60(21), 6491-6507. https://doi.org/10.1080/00207543.2021.1992680.

[11] Wang, H. Y., Wang, J. S., & Zhu, L. F. (2021). A new validity function of FCM clustering algorithm based on intra-class compactness and inter-class separation. *Journal of Intelligent & Fuzzy Systems*, 40(6), 12411-12432. https://doi.org/10.3233/JIFS-210555.

[12] Tang, C., Wang, H., Wang, Z., Zeng, X., Yan, H., & Xiao, Y. (2021). An improved OPTICS clustering algorithm for discovering clusters with uneven densities. *Intelligent Data Analysis*, 25(6), 1453-1471. https://doi.org/10.3233/IDA-205497.

[13] Dalmaijer, E. S., Nord, C. L., & Astle, D. E. (2022). Statistical power for cluster analysis. *BMC Bioinformatics*, 23(1), 1-28. https://doi.org/10.1186/s12859-022-04675-1.

[14] Zhang, Y., Zhang, Y., & Zhang, R. (2020). Text information classification method based on secondly fuzzy clustering algorithm. *Journal of Intelligent & Fuzzy Systems*, 38(6), 7743-7754. https://doi.org/10.3233/JIFS-179844.

[15] Yang, Z., Xu, P., Yang, Y., & Kang, B. (2021). Noise robust intuitionistic fuzzy c-means clustering algorithm incorporating local information. *IET Image Processing*, 15(3), 805-817. https://doi.org/10.1049/ipr2.12064.

[16] Calzavara, M., Battini, D., Bogataj, D., Sgarbossa, F., & Zennaro, I. (2020). Ageing workforce management in manufacturing systems: state of the art and future research agenda. *International Journal of Production Research*, 58(3), 729-747. https://doi.org/10.1080/00207543.2019.1600759.

[17] Barma, M., & Modibbo, U. M. (2022). Multiobjective mathematical optimization model for municipal solid waste management with economic analysis of reuse/recycling recovered waste materials. *Journal of Computational and Cognitive Engineering*, 1(3), 122-137. https://doi.org/10.47852/bonviewJCCE149145.

[18] Ghazal, T. M. (2021). Performances of K-means clustering algorithm with different distance metrics. *Intelligent Automation & Soft Computing*, 30(2), 735-742. https://doi.org/10.32604/iasc.2021.019067.

[19] Kaur, A., & Kumar, Y. (2022). A multi-objective vibrating particle system algorithm for data clustering. *Pattern Analysis and Applications*, 25(1), 209-239. https://doi.org/10.1007/s10044-021-01052-1.

[20] Arun, R., & Balamurugan, R. (2020). Distributed Entropy energy-efficient clustering algorithm for cluster head selection (DEEEC). *Journal of Intelligent & Fuzzy Systems*, 39(6), 8139-8147. https://doi.org/10.3233/JIFS-189135.

[21] Ran, X., Xi, Y., Lu, Y., Wang, X., & Lu, Z. (2023). Comprehensive survey on hierarchical clustering algorithms and the recent developments. *Artificial Intelligence Review*, 56(8), 8219-8264. https://doi.org/10.1007/s10462-022-10366-3.

[22] Gao, X., Zhang, Y., Wang, H., Sun, Y., Zhao, F., & Zhang, X. (2023). A modified fuzzy clustering algorithm based on dynamic relatedness model for image segmentation. *The Visual Computer*, 39(4), 1583-1596. https://doi.org/10.1007/s00371-022-02430-4.

[23] Sabir, Z., Ali, M. R., & Sadat, R. (2023). Gudermannian neural networks using the optimization procedures of genetic algorithm and active set approach for the three-species food chain nonlinear model. *Journal of Ambient Intelligence and Humanized Computing*, 14(7), 8913-8922. https://doi.org/10.1007/s00371-022-02430-4.

[24] Gülmez, B. (2023). A novel deep neural network model based on Xception and genetic algorithm for detection of COVID-19 from X-ray images. *Annals of Operations Research*, 328(1), 617-641. https://doi.org/10.1007/s10479-022-05151-y.

[25] Pajak, M., Brus, G., & Szmyd, J. S. (2023). Genetic algorithm-based strategy for the steam reformer optimization. *International Journal of Hydrogen Energy*, 48(31), 11652-11665. https://doi.org/10.1016/j.ijhydene.2021.10.046.

[26] Chen, X., Ding, Y., Cory, C. A., Hu, Y., Wu, K. J., & Feng, X. (2021). A decision support model for subcontractor selection using a hybrid approach of QFD and AHP-improved grey correlation analysis. *Engineering, Construction and Architectural Management*, 28(6), 1780-1806. https://doi.org/10.1108/ECAM-12-2019-0715.