SMART AGRICULTURE DATA MINING AND GRAIN HARVESTER DATA ANALYSIS BASED ON CLUSTER ANALYSIS ALGORITHM

/

基于聚类分析算法的智慧农业数据挖掘谷物收获机数据分析研究

Yujing HE*, Xinran SHANG, Zehe LIU, Chang WEI, Ruiqiang JI, Hengbin ZHANG
College of Mechanical & Electrical Engineering, Henan Agricultural University, Zhengzhou, 450046, China
E-mail: heyujinghn@outlook.com
Corresponding author: Yujing He

DOI: https://doi.org/10.35633/inmateh-78-17

Keywords: *Smart agriculture, Grain harvester, K-means clustering, Apriori algorithm, Data analysis*

ABSTRACT

With the speedy development of information technology, smart agriculture has become the key to the transformation and upgrading of modern agriculture. To improve the precision and practicality of data analysis, a data analysis model for grain harvesters based on a combination of K-means and Apriori is designed. This model collects grain harvester data in real-time, uses the K-means for preliminary clustering, and integrates Apriori algorithms to dynamically adjust clustering centers to improve clustering accuracy. At the same time, the model introduces ResNet to extract image features of grain harvesters, thereby enhancing the comprehensiveness of data analysis. Comparative experiments show that the algorithm has a high adjusted Rand index of 0.92, an F-value of 0.89, a convergence time of only 12.4 seconds, and a clustering accuracy of 95% for agricultural machinery databases. The analysis of actual operational data of grain harvesters shows that their faults are concentrated in the transmission device, accounting for up to 45%. When operating under high load, the work efficiency drops sharply, and when the rated power exceeds 30%, the work efficiency is only 72%. When working in a wet and muddy environment, the failure rate reaches 42.8%. From the above results, the data analysis model for grain harvesters based on the combination of K-means and Apriori algorithm proposed in the study can perform cluster analysis on the data of grain harvesters, laying a solid foundation for the sustainable development of smart agriculture.

摘要

随着信息技术的飞速发展，智慧农业已成为现代农业转型升级的关键。为了提高数据分析的精度和实用性，设计了一种基于 K-means 和 Apriori 相结合的粮食收割机数据分析模型。该模型实时采集粮食收割机数据，采用 K-means 进行初步聚类，并结合 Apriori 算法动态调整聚类中心，提高聚类精度。同时，该模型引入 ResNet 提取粮食收割机图像特征，增强了数据分析的全面性。对比实验表明，该算法调整后的 Rand 指数为 0.92，f 值为 0.89，收敛时间仅为 12.4 秒，对农机数据库的聚类准确率达到 95%。对粮食收割机实际运行数据的分析表明，其故障集中在传动装置上，占比高达 45%。在高负荷下运行时，工作效率急剧下降，当额定功率超过 30%时，工作效率仅为 72%。在潮湿泥泞的环境下工作时，故障率达到 42.8%。从以上结果可以看出，本研究提出的基于 K-means 和 Apriori 算法相结合的粮食收割机数据分析模型可以对粮食收割机数据进行聚类分析，为智慧农业的可持续发展奠定坚实的基础。

INTRODUCTION

As a country with a long history and an important position in agriculture, the development of smart agriculture in China is not only a key path to improve agricultural production efficiency, but also an important means to achieve rural revitalization. The grain harvester, as one of the key equipment in agricultural production, takes an irreplaceable part in the process of intelligent transformation. In depth mining and analysis of data from grain harvesters is greatly significant for the development of smart agriculture (*Myhailovych et al., 2023*). The data generated by grain harvesters during actual operation has the characteristics of high dimensionality, nonlinearity, and dynamic changes. How to effectively mine and analyze grain harvester data has become a key and difficult research topic (*Ding et al., 2023*). In traditional data analysis methods for grain harvesters, the exploration of complex relationships between data is not deep enough and relies heavily on manual judgment. The accuracy and real-time performance of grain harvester data analysis cannot meet the requirements of real-time decision-making in smart agriculture (*Geng et al. 2023*).

To solve this problem, numerous scholars have carried out research on it. For example, the Wang team designed a smart control algorithm based on multi-sensor data fusion and model predictive control technology to address the problem of difficulty in accurately and stably controlling the height of the cutting table in grain combine harvesters due to terrain. The experiment showed that its control accuracy reached 91.25%-91.5%, with zero ground contact times and good stability (*Wang et al., 2024*). *Xie et al. (2023)* proposed a novel method for the synchronous detection of multiple grain lodging features using 3D point cloud data acquired from a depth camera, with the aim of reducing high harvest losses caused by grain lodging during combine harvester operations. The detection experiment showed that the max error of lodging degree was less than 9.0 centimeters, and the mini detection error was less than 5.0 centimeters. The *Guo et al. (2025)* team proposed a method of enhancing threshing performance monitoring using digital twins to address the difficulty of directly monitoring the threshing performance of combine harvesters. Comparative experimental results showed that the fragmentation rate was cut down by 2.08% and the working speed was increased by 1.12 km/h. The *Cui et al. (2024)* team proposed an automatic unloading method based on stereo vision to solve the issues of low unloading efficiency and dependence on manual labor in traditional track driven rice combine harvesters. Comparative experimental results showed that the MAE between the depth of the unloading truck frame and the actual value was 0.014 m, and the RMSE was 0.017 m. *Zhang et al. (2023)* proposed improved deep separable convolutional neural network V3+ and YOLOv4 methods to address the problem of difficult extraction of small and dense objects in dynamic rice flow images of combine harvesters. The experiments showed that the extraction accuracy was promoted by more than 4.01%, and the error and evaluation schedule were excellent.

However, in the above studies, scholars mostly focus on single problems or specific scenarios, lacking systematic optimization of the overall data mining process, and have not comprehensively solved the complex problems under the synergistic influence of multiple factors in smart agriculture, which still needs to be optimized. K-Means Clustering Algorithm (K-means) is a classic unsupervised learning algorithm that iteratively updates cluster centers to divide data into K clusters, achieving effective data classification (*Ikotun et al., 2023*). The Apriori algorithm (Apriori) mines frequent itemsets to identify association rules within datasets, thereby revealing the intrinsic relationships among data items (*Zhang and Zhang, 2023*). These two algorithms have been broadly applied in various fields. For example, *Kumar et al. (2024)* team proposed an innovative framework that includes data-driven clustering estimation and robust initialization for the manual selection of cluster numbers and sensitivity to initial centroids in the K-means. Results showed that the Adjusted Rand Index (ARI) increased by 15%, and the algorithm performance was enhanced. The *Zubair et al. (2024)* team proposed an effective method for finding the optimal initial centroid in the K-means to address the problem of difficult determination of the initial centroid. Experiments denoted that it outperformed traditional k-means++ and random initialization methods in terms of computation time and iteration times. The *Javidan et al. (2023)* team proposed a method combining a new image processing algorithm to address the problems of time-consuming diagnosis of plant diseases and difficulty in distinguishing similar symptoms. The experiment showed that the accuracy of grape leaf disease classification reached 98.97%, and the processing time was shorter than that of deep learning models. *Hassan et al. (2023)* proposed using Apriori to predict suicide behavior through association rule analysis in response to the preventable public health issue of suicide. The experiment showed that Apriori found eight key rules in the data, with a support rate of 0.25 and a confidence level of 0.90. The *Song and He (2023)* team proposed an intelligent tourism recommendation system based on artificial intelligence and the Internet of Things to solve the problem of tourists' difficulty in obtaining high-value data from massive tourism information. The experiment showed that its Apriori had an accuracy of 94.3%, which was better than traditional algorithms. *Vlăduț de et al. (2024)* addressed the issue of reduced rainfall and lower crop yields due to severe climate change over the past 20 years by proposing the adoption of conservation and ecological organic farming practices (no-till, surface mulching with crop residues, crop rotation, etc.). Their review of 425 global studies concluded that soil quality and health significantly improved, while these practices could mitigate or enhance crop yields under adverse conditions. Sun et al. (2025) addressed the issues of insufficient functionality and visibility in the interface of a smart agricultural management cloud platform by proposing a UI design evaluation model integrating the Analytic Hierarchy Process and fuzzy comprehensive evaluation. Based on user heatmaps, three solution sets were optimized, resulting in an improvement in platform scores from 80.524 to 86.927, validating the method's effectiveness and its potential to enhance user experience. Chen et al. (2023) addressed the issue of the lack of accurate models in discrete element simulation and proposed using alfalfa stems with high moisture content during the budding stage as the object.

With the help of EDEM, Hertz Mindlin (no slip) and Hertz Mindlin with bonding contact models were used to calibrate physical and bonding parameters such as Poisson's ratio, shear modulus, collision recovery coefficient, static/rolling friction coefficient, normal/tangential contact stiffness, critical normal/tangential stress, bonding radius, etc. through Plackett Burman, Steepest Ascent, and Box Behnken experiments. The simulated angle of repose had a relative error of 0.52% compared to the physical experiment, and the simulated shear failure force had a relative error of 0.86% compared to the physical experiment. This indicates that the calibrated parameters can truly reflect the physical and bonding parameters of alfalfa stems. Physical and mechanical properties. *Chen et al. (2024)* proposed using EDEM based Hertz Mindlin (no slip) contact model to address the issue of lack of accurate models in the development of grass harvesting and crushing machinery. They calibrated the recovery coefficient, static/rolling friction and other contact parameters using Plackett Burman, Steepest Ascent and Box Behnken experiments, and verified them through angle of repose simulation. The relative error between the simulated and measured angles of repose was 0.48%, indicating that the calibrated parameters can truly reflect the physical characteristics of alfalfa stems during the bud stage, providing a reliable model and reference for discrete element simulation and grass machinery design.

In summary, the K-means algorithm requires manual or evaluation methods to determine the number of clusters, but in practice, it is difficult to accurately judge the rationality of K, which can lead to unstable clustering effects and affect the accuracy of data analysis. Although the Apriori can reveal data correlations, it requires setting minimum support and confidence thresholds. Improper thresholds can easily miss important rules and affect the depth and breadth of data mining. In response to the above issues, this study proposes a hybrid algorithm that combines the advantages of K-means and Apriori. By adaptively adjusting the number of clusters and thresholds, it optimizes data classification and association rule mining. The novelty of the research lies in using the Apriori's association rules to dynamically adjust the clustering centers of K-means, achieving collaborative optimization of data clustering and association rule mining. Residual Networks (ResNets) are employed to extract harvester image features, and the feature maps are fused with numerical data to provide more accurate support for grain harvester data analysis.

MATERIALS AND METHODS

EXPERIMENTAL DESIGN

In order to systematically verify the superiority of the K-Apriori hybrid algorithm proposed in this study and the effectiveness of the data analysis model for grain harvesters it constructs, a multi-level and comparative experimental scheme was designed. The core of the experiment revolves around two main parts: one is the performance evaluation of the algorithm itself, and the other is the efficacy verification of the complete model built based on this algorithm in specific application scenarios.

This experiment aims to isolate the application scenarios and objectively evaluate the clustering performance, efficiency, and stability of the K-Apriori algorithm on a standard dataset. In the selection of benchmark algorithms for comparison, three representative clustering algorithms were chosen as the benchmark: 1) Mean Shift (MS): a density-based non-parametric algorithm that does not require specifying the number of clusters in advance; 2) DBSCAN: capable of discovering clusters of any shape and identifying noise points; 3) Agglomerative Hierarchical Clustering (AHC): generating a cluster tree diagram through hierarchical decomposition. These algorithms were selected to cover different technical routes such as parametric and non-parametric, density-based and distance-based.

Regarding the experimental data set, in addition to using the harvesters' data collected in this study, to ensure the universality of the evaluation, the experiment also introduced two publicly available high-dimensional datasets from the UCI machine learning library (such as "Wine" and "Iris") as supplementary test sets to verify the performance of the algorithm in different data distributions.

In terms of parameter settings and optimization, for K-Apriori, the initial number of clusters K in the K-means part was determined through the "elbow rule" combined with the silhouette coefficient in the pre-experiment. The minimum support and minimum confidence (optimized through grid search on the validation set to balance the quantity and quality of rules) in the Apriori part were set. For the key parameters of the benchmark algorithms, grid search or heuristic methods based on data distribution were used for optimization settings to ensure that all algorithms run under their relatively optimal or reasonable parameters.

Regarding the performance evaluation process, the K-Apriori, MS, DBSCAN, and AHC algorithms were run on the preprocessed data set separately. Each experiment was run 30 times, and their clustering results were recorded. During the evaluation, the true labels of the data or unsupervised internal indicators were used to calculate the adjusted Rand Index (ARI) and F value to measure the clustering accuracy.

At the same time, the average running time from the start to convergence of each algorithm was recorded to evaluate the computational efficiency, and the normalized mutual information (NMI) between the clustering results of different runs was calculated to evaluate the stability of the algorithm. Additionally, the memory usage during the algorithm's execution process was monitored, and the theoretical value of its computational complexity was analyzed.

The comprehensive verification experiment of the grain harvester data analysis model aims to evaluate the efficacy of the complete analysis framework built based on the K-Apriori algorithm in actual application scenarios. This experiment takes the preprocessed dataset of the harvesters, which contains multi-dimensional time series data and synchronous image data, as the input, ensuring that the model strictly follows its established architecture. In the core analysis module, the K-Apriori algorithm first clusters the numerical operating condition parameters to divide into typical working mode clusters such as "normal operation" and "high-load operation"; then, the Apriori module mines frequent item sets and association rules within and between clusters, aiming to discover potential relationships between specific parameter combinations and fault types. To quantitatively evaluate the improvement in model performance brought about by introducing ResNet to extract image features, an ablation experiment was designed: by comparing the baseline model that only uses numerical data with the complete model that integrates image features, the accuracy differences between the two on specific recognition tasks (such as identifying blockage faults and judging crop lodging) were tested on the same test set, thereby clearly identifying the information gain brought by the image features. The evaluation of the model output is multi-dimensional. Firstly, using the labeled information of equipment models or service years, the accuracy of clustering the harvesters by the model was calculated to verify its pattern recognition ability. Secondly, the fault association rules mined by the model were cross-validated with historical maintenance records to analyze whether the support and confidence of the rules were consistent with the actual statistical situation. Finally, the conclusions about the relationship between load, environment and operation efficiency derived by the model were compared with the data from engineering manuals or the experience of domain experts to verify the rationality of its analysis results. In addition, typical clusters from the clustering results were selected for in-depth case analysis to demonstrate how the model reveals the coupling relationships among multiple factors (such as service years, operating environment, maintenance history) through association rules, thereby generating targeted decision recommendations beyond the analysis of a single indicator. Through this series of comprehensive verifications, the aim is to fully prove that this data analysis model can provide precise, in-depth and actionable decision support for equipment health management in smart agriculture.

K-mean algorithm optimized based on Apriori

Against the backdrop of speedy advancement of smart agriculture, the intelligence of agricultural machinery has become a key direction for improving agricultural production efficiency (*Gheisari et al., 2023*). As a core equipment in agricultural production, grain harvesters accumulate massive amounts of data through sensors, global positioning systems, and other devices during operation. To analyze grain harvester data more accurately, data mining technology plays a key role in improving agricultural production efficiency. Traditional data mining techniques face challenges in dealing with the complexity and diversity of massive data. The K-means algorithm significantly enhances the efficiency and accuracy of data processing by optimizing the selection and iteration process of clustering centers. Therefore, the K-means is introduced in the study, and its structural diagram is denoted in Fig. 1.

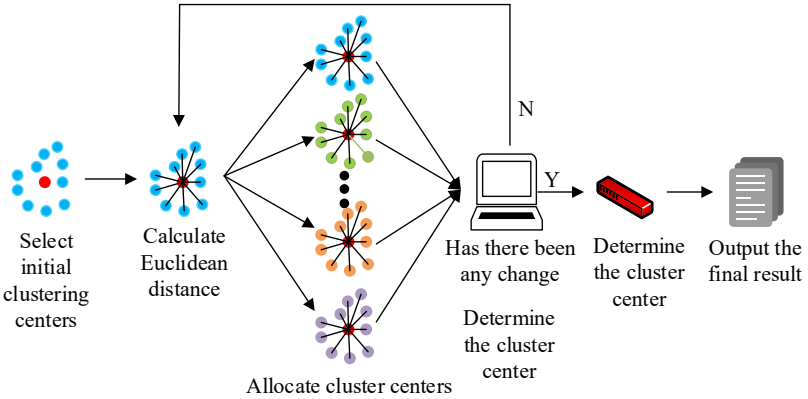


Fig. 1 - Structure diagram of K-means

As shown in Fig. 1, the K-means algorithm consists of four steps: selecting initial cluster centers for data, calculating the Euclidean distance between samples and cluster centers, determining cluster centers, and assigning data to the final clustering results. It randomly selects k initial cluster centers in the dataset, and the choice of initial centers directly affects the clustering effect. By calculating the Euclidean distance between each sample and the initial center, the samples are assigned to the nearest cluster center. The Euclidean distance calculation is shown in equation (1).

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

In equation (1), d means the Euclidean distance, x_i and y_i denote the coordinates of the sample point and cluster center, respectively. Each data point is labeled with a cluster label c_i and the cluster center is recalculated using equation (2).

$$\mu_j = \frac{1}{|C_j|}\sum_{x_i \in C_j} x_i \tag{2}$$

In equation (2), μ_j is the cluster center, C_j denotes the member of the j-th cluster, and $|C_j|$ denotes the sample size of the j-th cluster. Compare the cluster centers before and after, identify if there have been any changes, and continue iterating until the cluster centers stabilize. Finally, all samples are assigned to a determined cluster, and clustering is completed through multiple iterations. The algorithm gradually optimizes the clustering results to ensure that each sample point belongs reasonably. The final clustering center is stable and outputs accurate clustering results. The computational complexity of K-means algorithm increases with the increase of sample size, and tends to generate clusters of similar size and regular shape, which can lead to poor performance in handling imbalanced data. The Apriori reduces computational complexity through pruning strategies and efficiently processes large datasets by mining frequent itemsets to discover association rules. Therefore, the study introduces the Apriori to optimize the K-means, and its structural flowchart is shown in Fig. 2.

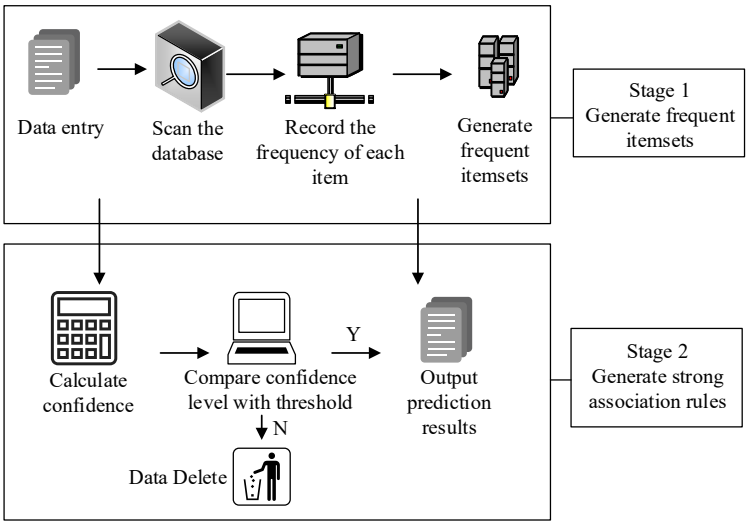


Fig. 2 - Flowchart of Apriori structure

(Source from:https://www.reshot.com/free-svg-icons/item/calculator-CYBQ3F87ZV/)

As shown in Fig. 2, the Apriori is composed of two stages. The first stage is to generate a frequent itemset, which scans the database, records the frequency of each item, and filters out itemsets with support greater than the threshold, denoted as frequent itemsets. A candidate set is generated from frequent itemsets, and pruned according to Apriori properties. Candidate itemsets that do not meet the conditions are removed, and candidate itemsets that meet the conditions are retained. The second stage is to generate strong association rules, by calculating the confidence level, filtering out rules with confidence levels higher than the threshold, and finally outputting effective association rules to achieve efficient data mining. The confidence calculation formula is denoted in equation (3).

$$Confidence(A \rightarrow B) = \frac{Support(A \cup B)}{Support(A)} \tag{3}$$

In equation (3), $Support(A \cup B)$ represents the support for both itemsets A and B, while $Support(A)$ means the support for itemset A. Although the simplicity and ease of implementation of the Apriori make it highly valuable in the field of association rule mining, frequent itemset generation and pruning processes consume a significant amount of computational resources, leading to performance bottlenecks when dealing with a large number of itemsets. Therefore, the study combines K-means with Apriori to form the K-Apriori, and the flowchart of the algorithm structure is shown in Fig. 3.

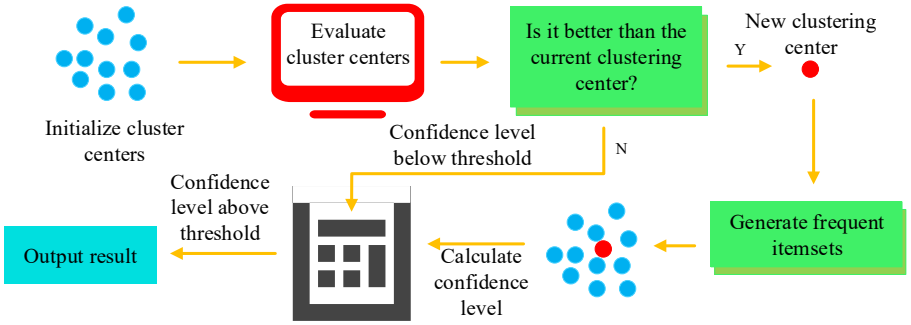


Fig. 3 K- Apriori structure flow chart

As shown in Fig. 3, the K-Apriori introduces the Apriori on the basis of traditional K-means. The algorithm first utilizes K-means to determine the initial cluster center, clusters similar data into several clusters, reduces data dimensionality, and reduces the number of candidate sets. Subsequently, the Apriori is independently employed within each cluster to mine frequent itemsets and generate association rules. Based on the distribution characteristics of frequent itemsets within the cluster, the algorithm recalculates the cluster centers to make the data distribution more compact, thereby enhancing the accuracy and efficiency of association rule mining. The K-Apriori achieves adaptive adjustment of data structure by iteratively optimizing the clustering center and frequent itemset mining process, thereby improving the convergence speed of rule mining while reducing computational complexity.

Construction of data analysis model for grain harvester

The K-Apriori algorithm can quickly identify frequent itemsets in different data clusters and extract association rules based on these frequent itemsets, thereby optimizing the efficiency and accuracy of data mining. However, traditional grain harvester data analysis is usually limited to single indicator evaluation, lacking a global perspective, and overly relying on manual experience judgment, making it difficult to accurately grasp the mutual influence between multiple factors in complex operating environments. The K-Apriori algorithm achieves comprehensive and specific data analysis by collecting data features of different operating states of grain harvesters. Therefore, this study used the K-Apriori to construct a data analysis model for grain harvesters, achieving automated collection and deep clustering analysis of operational data. The model structure is shown in Fig. 4.

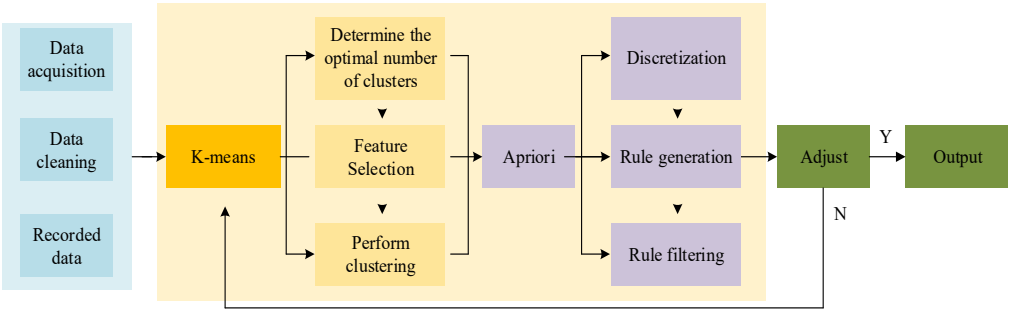


Fig. 4 - Structure diagram of data analysis model for grain harvester

As shown in Fig. 4, the data analysis model of the harvester based on K-Apriori algorithm is divided into four modules: data acquisition and preprocessing, core algorithm analysis module, data analysis module, and result application and decision support. The data acquisition module records the real-time operation status images and data of the harvester, and the preprocessing module cleans the noisy data to ensure data quality. The data is standardized to eliminate the influence of dimensionality, as denoted in equation (4).

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{4}$$

In equation (4), x' denotes the normalized value, $x_{\min}$ denotes the mini value of the data, and $x_{\max}$ means the max value of the data. The core algorithm analysis module deeply analyzes the data through the K-Apriori algorithm, divides the harvester data into different clusters, and calculates the minimum sum of squared errors within each cluster. The calculation formula is denoted in equation (5).

$$J = \sum_{i=1}^{K} \sum_{x \in C_i} || x - \mu_i ||^2 \tag{5}$$

In equation (5), K means the amount of clusters, C_i denotes the point set of the i-th cluster, x refers to the data point belonging to C_j, and μ_i denotes the centroid of the i-th cluster.

The data is input into the data analysis module to analyze the performance indicators and operating status of the harvester, determine the current status of the grain harvester, and optimize its parameters. Finally, the optimized parameters are fed back to the actual task through result application and decision support module.

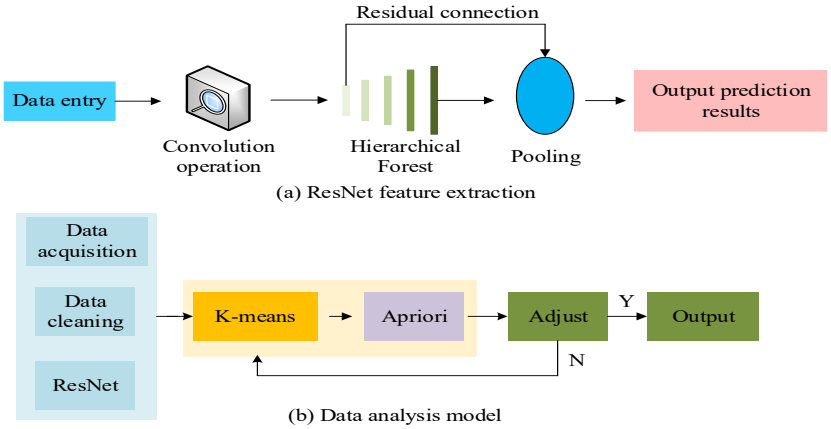


Fig. 5 - Fault prediction model

However, in the process of data processing, pure data cannot fully reflect the status of the grain harvester, and it needs to be analyzed in conjunction with images. Therefore, the study introduces ResNet to optimize the data analysis model, extracts features from images through ResNet, and then analyzes the features. The flowchart is shown in Fig. 5.

As shown in Fig. 5 (a), ResNet performs convolution operation after data input to capture effective features in the image, and its feature extraction formula is shown in equation (6).

$$\begin{cases} H_{output} = \left\lfloor \dfrac{Input_h - Filter_h + 2 \times padding}{stride} + 1 \right\rfloor \\ W_{output} = \left\lfloor \dfrac{Input_w - Filter_w + 2 \times padding}{stride} + 1 \right\rfloor \end{cases} \tag{6}$$

In equation (6), H_{output} and W_{output} represent the height and width of the output feature map, $Input_h$ and $Input_w$ represent the height and width of the input feature map, $Filter_n$ and $Filter_w$ refers to the height and width of the convolution kernel, $padding$ represents the fill value, and $stride$ represents the stride. After convolution operation, residual connections are used to fuse the input features with the convolved features, enhancing the feature extraction ability. The residual connection formula is denoted in equation (7).

$$a^{(l+2)} = g(Z^{(l+2)} + a^{(l)}) \tag{7}$$

In equation (7), $a^{(l)}$ represents the activation output of the l-th layer, $Z^{(l+2)}$ represents the result after 2-layer network transformation, g represents the activation function, and $a^{(l+2)}$ represents the final output of the residual block. The feature fusion formula is shown in equation (8).

$$F_o = F_{in} \times (M_h + M_v) \tag{8}$$

In equation (8), F_{in} is the feature fusion map, M_h and M_v are the horizontal and vertical attention maps, respectively, and F_o denotes the input feature map. Finally, the feature map of the grain harvester is output.

These feature maps capture key information in grain harvester operation images, such as grain distribution density, mechanical component wear status, or visual features of environmental obstacles. As shown in Fig. 5 (b), the extracted feature vectors are then fused with the preprocessed numerical data, and the fusion formula is shown in equation (8).

$$F(x) = \alpha(f_1(x) + f_2(x) + \cdots f_n(x)) + \beta(f_1'(x) + f_2'(x) + \cdots f_n'(x)) \quad (9)$$

In equation (9), $F(x)$ is the fused image, α is the numerical data weight, β is the image data weight, $f_n(x)$ is the numerical data, n is [1, n], and $f_n'(x)$ is the image data. This matrix is input into the data analysis module and subjected to deep processing using the K-Apriori algorithm. The fused analysis results are used to real-time determine the status of the harvester, generate parameter optimization suggestions, and provide feedback to the control system through the decision support module.

In summary, the grain harvester data analysis model uses the K-Apriori algorithm to identify frequent itemsets during the operation of the grain harvester, mine the association rules between the operating parameters of each component of the harvester, and perform collaborative analysis with the image features extracted by the ResNet to comprehensively analyze the grain harvester data.

RESULTS

Performance testing of K-Apriori algorithm

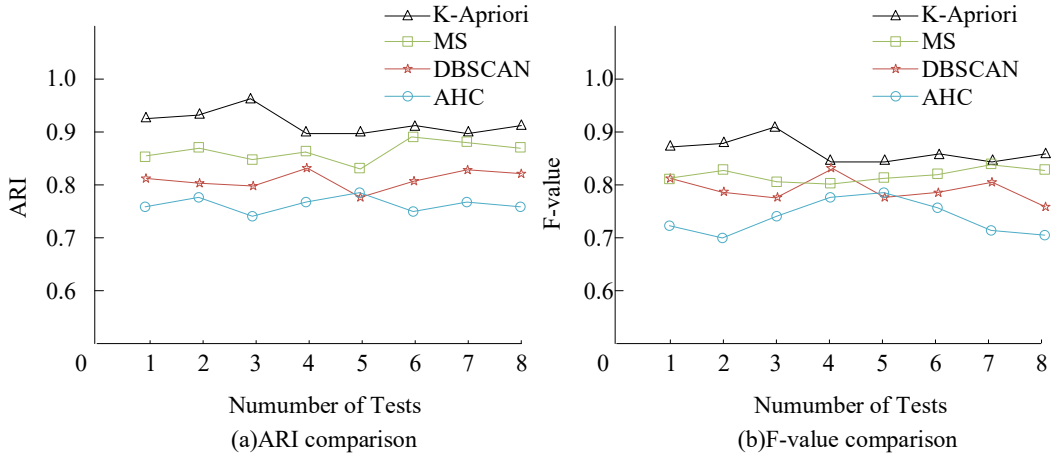
To verify the clustering effectiveness of K-means and Apriori, a comparative experiment was conducted between K-Apriori algorithm and Mean Shift (MS) algorithm, Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm, and Agglomerative Hierarchical Clustering (AHC) algorithm. The environment configuration during the experiment is denoted in Table 1.

Table 1

Experimental environment configuration

Environment	Index	Type
Hardware environment	OS	Windows10
	Processing element	Intel Core i7-13700KF
	EMS memory	16GB
Software environment	Global mapper version	Global mapper 24 (64-bit)
	Python Version	Python 3.6
	Crawler framework	Pandas2.1.0

This study carefully selected some public data disclosed by multiple agricultural machinery enterprises in Heilongjiang Province during the research process, and combined it with rich data resources obtained through efficient data crawling technology to ensure the accuracy and reliability of the experimental findings. Four algorithms were tested on the same dataset, and the ARI values and F-values were compared. The results are shown in Fig. 6.



(a)ARI comparison (b)F-value comparison

Fig. 6 - Comparison of ARI and F-values for different algorithms

As shown in Fig. 6, the K-Apriori algorithm performed excellently in clustering performance evaluation, with an ARI of 0.92 and an F-value of 0.89, both of which significantly surpass other comparative algorithms, demonstrating its outstanding ability in clustering accuracy and consistency. In contrast, the ARI value of the MS algorithm was 0.85 and the F-value was 0.82. Although its performance was decent, there was still a certain gap compared to the K-Apriori algorithm. The ARI value of DBSCAN algorithm was 0.81, and the F-

value was 0.78, indicating relatively weak performance but still within an acceptable range. The ARI value of AHC algorithm was 0.75, and the F-value was 0.72, further highlighting its shortcomings in clustering performance. By comparing and analyzing these data, the significant superiority of the K-Apriori in clustering performance could be more clearly verified. Further comprehensive evaluation of the convergence speed and stability of the algorithm shows that the K-Apriori has demonstrated high efficiency and stability in multiple experiments. Further comparison of the convergence speed and stability of the algorithm is shown in Fig. 7.

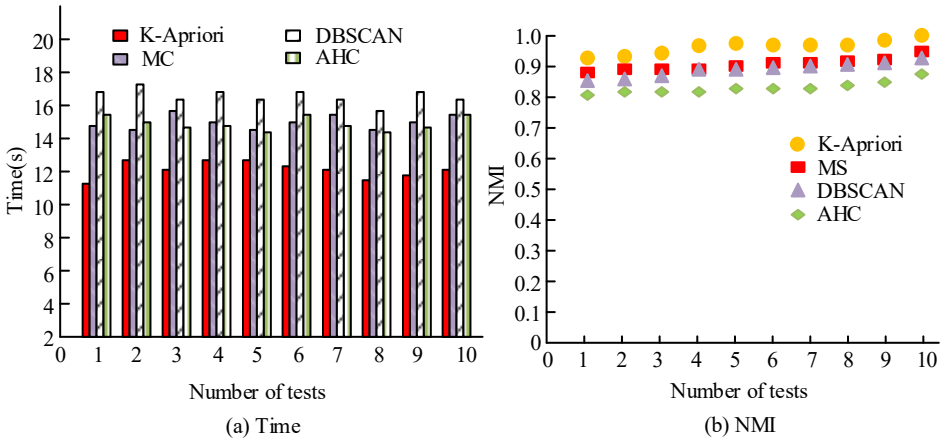


Fig. 7 - Comparison of convergence speed of four algorithms with NMI

According to Fig. 7 (a), the average convergence time of K-Apriori, MS, DBSCAN, and AHC algorithms reached 12.4 seconds, 14.3 seconds, 17.4 seconds, and 15.2 seconds, respectively. Among them, the K-Apriori algorithm had a significant advantage in convergence speed, and its efficient performance provided strong support for processing large-scale data, significantly improving data processing efficiency. According to Fig. 7 (b), in the stability testing phase, the Normalized Mutual Information (NMI) values of each algorithm were recorded as 0.95, 0.88, 0.84, and 0.81, respectively. The K-Apriori algorithm also performed excellently in this metric, not only leading in numerical values, but also ensuring reliability and consistency in data processing, further verifying its stability and credibility in practical applications. In addition, a comparison was made on the memory consumption and computational complexity of the algorithm, as shown in Fig. 8.

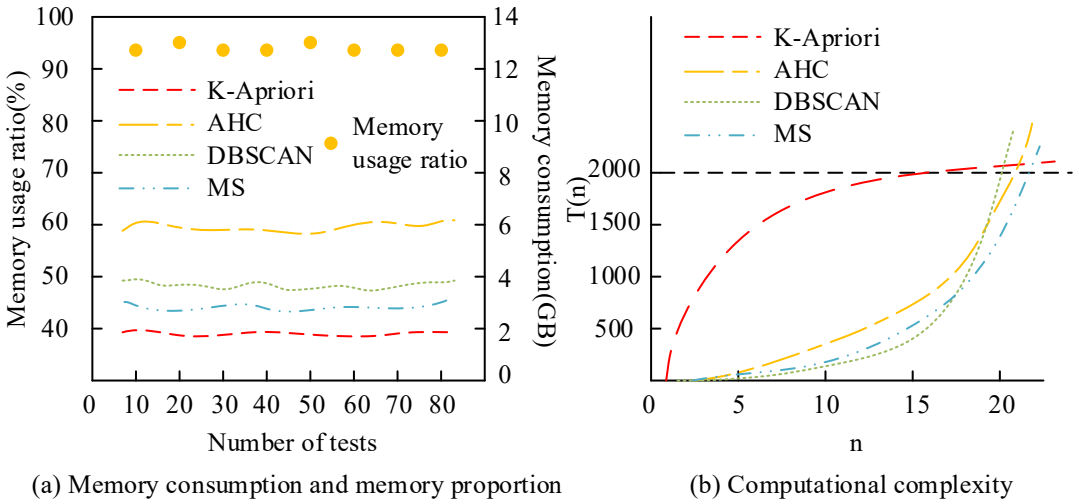


Fig. 8 - Comparison of memory consumption and computational complexity of four algorithms

As shown in Fig. 8 (a), the K-Apriori algorithm performed particularly well in terms of memory consumption, occupying only 2GB of memory resources. Compared to the 3GB of memory required by the MS algorithm, the 4GB of memory required by the DBSCAN algorithm, and the 6GB of memory required by the AHC algorithm, its memory usage was significantly reduced. As shown in Fig. 8 (b), the K-Apriori algorithm also exhibited significant advantages in computational complexity, with a computational complexity of only O ($n\log_2 n$). Compared to other algorithms that generally achieve O (n^2) complexity, the K-Apriori algorithm had a qualitative leap in computational efficiency. This feature fully demonstrates the dual superiority of K-Apriori algorithm in resource utilization and computational efficiency, making it more competitive in practical applications.

Application verification of data analysis model for grain harvester

This study selected the grain harvester data from the above-mentioned database and analyzed the grain harvester data using this model. The results are shown in Fig. 9.

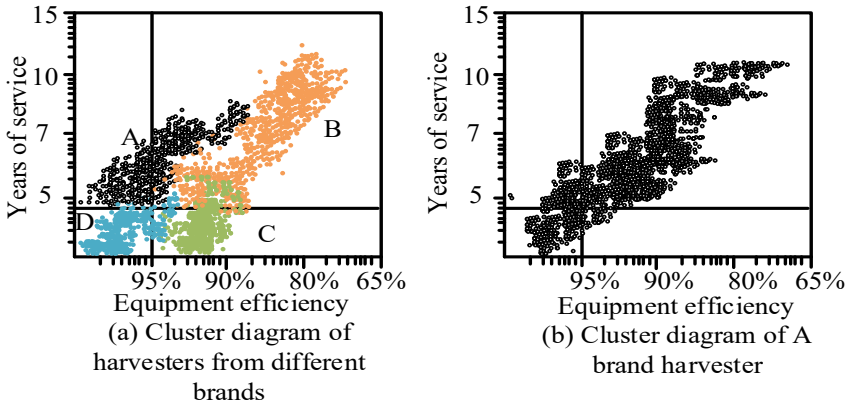


Fig. 9 - Cluster diagram of grain harvester

As shown in Fig. 9 (a), the data analysis model of grain harvesters conducted detailed clustering analysis on the types and working years of grain harvesters through in-depth mining and analysis. This process not only accurately distinguished different models and types of grain harvesters, but also performed clear clustering based on their service life, and the accuracy of clustering reaches 95%. Through this clustering analysis, the model successfully identified the differences in performance among different grain harvesters, and also revealed the trend of the impact on equipment efficiency with increasing working years. As shown in Fig. 9 (b), the data analysis model for grain harvesters classified the harvesters of brand A with an accuracy of 97% in clustering, verifying the applicability and accuracy of the model on brand specific data. This analysis result provides a scientific reference for equipment selection and maintenance strategies in agricultural production, enabling agricultural producers to choose and use grain harvesters more reasonably, thereby improving agricultural production efficiency and yield. In addition, the data analysis model of the grain harvester was used to analyze the operating indicators of the grain harvester, and the results are shown in Fig. 10.

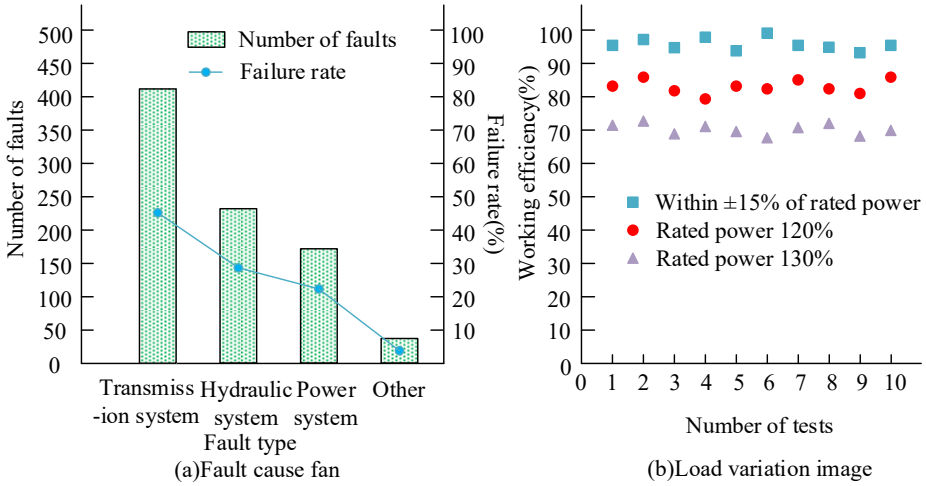


Fig. 10 - Fault analysis of grain harvester

According to Fig. 10 (a), the statistical analysis of the frequency of faults in the grain harvester data analysis model showed that the faults of the grain harvester were mainly concentrated in the transmission system, hydraulic system, and electrical system. Among them, the transmission system had the highest proportion of faults, reaching 45%, hydraulic system faults accounted for 28%, electrical system faults accounted for 23%, and the remaining 4% were caused by equipment aging and other external environmental factors. Based on this, targeted maintenance recommendations could be provided for the maintenance of grain harvesters, thereby reducing equipment downtime and improving overall operational efficiency. As shown in Fig. 10 (b), the monitoring of load changes by the grain harvester data analysis model showed that when the load fluctuation range was within ± 15% of the rated value, the equipment operation status remained stable and the operating efficiency was maintained at over 98%. When the peak load exceeded 120% of the rated value, the operating efficiency dropped to 85%, and when the load replication exceeded 130% of the rated

value, the operating efficiency of the grain harvester dropped significantly to 72%, accompanied by a higher risk of failure. To verify the universality and discriminative ability of the constructed data analysis model, the study selected four mainstream grain harvester brands with high market share and representative technological generations in the target area, and referred to them with anonymous letters A, B, C, and D in the report. Further correlation analysis was conducted on the operation and failure rate of the grain harvester, and the outcomes are denoted in Table 2.

Table 2

Path length and task completion rate

Harvester brand	A	B	C	D
New machine failure rate	3.1	3.5	2.5	4.8
After five years of use	16.8	18.9	17.6	21.8
Dry and flat	26.2	28.3	25.1	29.2
Muddy and damp	43.3	44.1	42.8	49.2
Working continuously for eight hours	47.2	48.2	43.2	50.7

According to Table 2, the failure rate of grain harvesters was relatively low during the new machine stage, with failure rates of all brands below 5%. However, as the service life increased, the failure rate gradually increased, especially after 5 years of use, the failure rate significantly increased, with some brands even reaching over 20%. The failure rate of grain harvesters varied significantly in different operating environments, with the lowest failure rate being 25.1% in dry and flat plots and over 42.8% in wet and muddy plots, indicating that the operating environment has a significant impact on equipment performance. When the harvester operated continuously for more than 8 hours, the failure rate further increased to 43.2%, indicating that long-term high-intensity operation causes significant equipment damage. It is recommended that agricultural producers arrange their work time reasonably, avoid excessive use, and strengthen equipment protection in wet and muddy environments to reduce the risk of malfunctions and ensure production efficiency.

DISCUSSION AND CONCLUSIONS

This study is based on a grain harvester data analysis model using the K-Apriori algorithm. The performance of the K-Apriori was analyzed, and the ARI value of the algorithm reached 0.92, with an F-value of 0.89. This outcome is similar to the research outcomes of *Huseynov et al. (2024)*, but the ARI value of Huseynov R's algorithm was only 0.85, with an F-value of 0.82. After analysis, the possible reason for this result is that the dataset size is different. This study used a larger dataset and optimized the feature extraction method to enable the model to more accurately capture fault patterns. In addition, the ARI value of *Panwar and Nanda (2024)* team's research on the same dataset was only 0.87, and the F-value was 0.84. The performance bottleneck of their algorithm lied in the failure to fully explore the correlations between data, resulting in insufficient model generalization ability. The Apriori association rule algorithm used in this study was capable of mining deeper fault patterns from data, revealing the relationships between various sets, and providing a more comprehensive and accurate perspective for data analysis of grain harvesters. Meanwhile, the clustering performance of the grain harvester data analysis model based on K-Apriori was better, with a clustering accuracy of 95%. The above results are roughly the same as those of the *Chen et al. (2022)* team. However, the clustering accuracy of the *Chen et al. (2022)* team was only 90%, which is lower than this study. This may be due to the introduction of the ResNet structure in this study to optimize the feature extraction process and enhance the accuracy and stability of clustering.

This model conducted an in-depth analysis of the causes of faults in grain harvesters, with faults mainly concentrated in the transmission system, hydraulic system, and electrical system. Among them, transmission system faults accounted for the highest proportion, reaching 45%. The application of this algorithm not only revealed the distribution pattern of faults, but also provided targeted maintenance strategies, effectively extending equipment life and improving job stability. Compared with *Xiaohui et al. (2022)*, the data analysis model of the grain harvester in this study improved in data mining depth and clustering accuracy, especially in predicting transmission system faults more accurately. It could accurately identify fault types and frequencies, providing scientific and reasonable preventive measures for agricultural producers. The data analysis model conducted a correlation analysis between the working conditions of the harvester and the failure rate, and found that the working hours were positively correlated with the failure rate.

The failure rate was less than 5% in the new machine stage, but as the usage time increased, the failure rate gradually increased, reaching over 20%. Moreover, different work environments also have a significant impact on the failure rate. In damp and muddy environments, the failure rate was 15% higher than in dry environments, reaching 40%. This demonstrated the significant impact of environmental factors on equipment performance, further validating the importance of refined data analysis in agricultural machinery management.

In summary, the grain harvester data analysis model based on K-Apriori algorithm proposed in this study not only deeply explores and comprehensively analyzes harvester data, but also achieves high-precision and high-efficiency data analysis by integrating K-Apriori and ResNet algorithms. However, this study also has limitations, such as a limited sample size that did not cover all operating environments, and the model's ability to analyze grain harvester data needs to be improved. Future research needs to expand the data scope and optimize algorithms to further strengthen the universality and accuracy of the model, providing more comprehensive guarantees for agricultural production.

ACKNOWLEDGEMENT

The research is supported by The Key Scientific and Technological Project of Henan Province Department of China, Research on key technology and equipment of peanut intelligent agricultural machine for dust removal, (No.: 232102211087).

REFERENCES

[1] Chen M, Jin C, Ni Y. (2022). Online field performance evaluation system of a grain combine harvester (谷物联合收割机在线田间性能评估系统) [J]. *Computers and Electronics in Agriculture*, 198: 107047. DOI: 10.1016/j.compag.2022.107047

[2] Chen T., Yi S.J., Li Y.F., Tao G.X., Qu S.M., Li R. (2023). Establishment and Parameter Calibration of Discrete Element Model for Alfalfa Stems at Budding Stage (苜蓿芽期茎秆离散元模型的建立及参数校准) [J]. *Transactions of the Chinese Society for Agricultural Machinery*, 54(5): 91-100. DOI: 10.6041/j.issn.1000-1298.2023.05.009

[3] Chen T., Yi S.J., Li Y.F., Tao G.X., Mao X. (2024). Calibration and test of contact parameters for alfalfa stalk at primary florescence based on discrete element method (基于离散元法的苜蓿茎秆初花期接触参数校准与测试) [J]. *PloS One*, 19(8): e0303064. DOI: 10.1371/journal.pone.0303064

[4] Cui Z., Hu J., Yu Y., Cao G., Zhang H., Chai X., Xu L. (2024). Automatic grain unloading method for track-driven rice combine harvesters based on stereo vision (基于立体视觉的履带式水稻联合收割机自动卸粮方法) [J]. *Computers and Electronics in Agriculture*, 220(8): 108917-108924. DOI: 10.1016/j.compag.2024.108917

[5] Ding F., Zhang W., Luo X. (2023). Gain self-adjusting single neuron PID control method and experiments for longitudinal relative position of harvester and transport vehicle (收割机与运输车辆纵向相对位置的增益自调整单神经元 PID 控制方法及实验) [J]. *Computers and Electronics in Agriculture*, 213(12): 108215-108221. DOI: 10.1016/j.compag.2023.108215

[6] Geng S., Tan J., Li L., Miao Y., Wang Y. (2023). Legumes can increase the yield of subsequent wheat with or without grain harvesting compared to Gramineae crops: A meta-analysis (与禾本科作物相比，豆类作物在收获谷物与否的情况下都能提高后续小麦的产量：一项荟萃分析) [J]. *European Journal of Agronomy*, 142(4): 126643-126649. DOI: 10.1016/j.eja.2022.126643

[7] Gheisari M., Hamidpour H., Liu Y., Saedi P., Raza A., Jalili A., Rokhsati H., Amin R. (2023). Data Mining Techniques for Web Mining: A Survey. *Artificial Intelligence and Applications*, 1(1): 3-10. DOI: 0000-0002-5643-0021

[8] Guo D., Du Y., Wang L., Zhang W., Sun T., Wu Z. (2025). Digital twin for monitoring threshing performance of combine harvesters (用于监测联合收割机脱粒性能的数字孪生模型) [J]. *Measurement.* 239(15): 115411-115421. DOI: 10.1016/j.measurement.2024.115411

[9] Hassan M.M., Karim A., Mollick S., Azam S., Ignatious E., Al Haque A.F. (2023). An Apriori algorithm-based association rule analysis to detect human suicidal behaviour. *Procedia Computer Science*. 219(12): 1279-1288. DOI: 10.1016/j.procs.2023.01.412

[10] Huseynov R., Aliyeva N., Bezpalov V., Syromyatnikov D. (2024). Cluster analysis as a tool for improving the performance of agricultural enterprises in the agro-industrial sector. *Environment, Development and Sustainability*, 26(2): 4119-4132. DOI:10.1007/s10668-022-02873-8

[11] Ikotun A.M., Ezugwu A.E., Abualigah L., Abuhaija B., Heming J. (2023). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, 622(13): 178-210. DOI: 10.1016/j.ins.2022.11.139

[12] Javidan S.M., Banakar A., Vakilian K.A, Ampatzidis Y. (2023). Diagnosis of grape leaf diseases using automatic K-means clustering and machine learning. *Smart Agricultural Technology*. 3(1): 100081-100089. DOI: 10.1016/j.atech.2022.100081

[13] Kumar A., Kumar A., Mallipeddi R., Lee D.G. (2024). High-density cluster core-based k-means clustering with an unknown number of clusters. *Applied Soft Computing*. 155(12): 111419-111432. DOI: 10.1016/j.asoc.2024.111419

[14] Myhailovych Y., Rogovskii I., Korobko M., Berezova L. (2023). Experimental studies of vibration load of synchronous threaded connections of grain harvester combines. *Engineering for Rural Development*. 22(12): 908-914. DOI: 10.22616/ERDev.2023.22.TF179

[15] Panwar A., Nanda S.J. (2024). Distributed enhanced multi-objective evolutionary algorithm based on decomposition for cluster analysis in wireless sensor network. *Journal of Network and Computer Applications*, 232(13): 104032-104040. DOI: 10.1016/j.jnca.2024.104032

[16] Song Y., He Y. (2023). Toward an intelligent tourism recommendation system based on artificial intelligence and IoT using Apriori algorithm (基于人工智能和物联网的智能旅游推荐系统：采用 Apriori 算法) [J]. *Soft Computing*. 27(24): 19159-19177. DOI: 10.1007/s00500-023-09330-2

[17] Sun J., Qi B., Sun X., Liu Y., Ren Y., Ma T., Zhang B., Ren Y. (2025). Evaluation and optimization of agricultural management cloud platform based on AHP/FCE (基于 AHP/FCE 的农业管理云平台的评估与优化) [J]. *INMATEH - Agricultural Engineering*. 75(1): 366-375, DOI: 10.35633/inmateh-75-31

[18] Wang Q., Meng Z., Wen C., Qin W.C., Wang F., Zhang A.Q., Yin Y.X. (2024). Grain combine harvester header profiling control system development and testing (谷物联合收割机割台轮廓控制系统的开发与测试) [J]. *Computers and Electronics in Agriculture*, 223(5): 109082-109092. DOI: 10.1016/j.compag.2024.109082

[19] Vlăduț N.V., Atanasov A., Ungureanu N., Ivașcu L.V., Cioca L.I., Popa L.D., et al. (2024). Trends in the development of conservation/ecological agriculture in the context of current climate change–a review. *INMATEH Agric. Eng*, 74(3): 980-1032. DOI: 10.35633/inmateh-74-71

[20] Yang X., Zhang G., Yao J.., Lian J., Wang X., Lv D., Deng Y., Zhang A. (2022). Reliability analysis of grain combine harvesters based on data mining technology (基于数据挖掘技术的谷物联合收割机可靠性分析) [J]. *INMATEH-Agricultural Engineering*, 67(2): 12-23. DOI: 10.35633/inmateh-67-21

[21] Xie B., Wang J., Jiang H., Zhao S., Liu J., Jin Y., Li Y. (2023). Multi-feature detection of in-field grain lodging for adaptive low-loss control of combine harvesters (针对联合收割机适应性低损耗控制的田间谷物倒伏多特征检测) [J]. *Computers and Electronics in Agriculture*, 208(12): 107772-107782. DOI: 10.1016/j.compag.2023.107772

[22] Zhang Q., Hu J., Xu L., Cai Q., Yu X., Liu P. (2023). Impurity/breakage assessment of vehicle-mounted dynamic rice grain flow on combine harvester based on improved Deeplabv3+ and YOLOv4 (基于改进的 Deeplabv3+ 和 YOLOv4 的车载动态水稻谷物流量联合收割机杂质/破损评估) [J]. *IEEE Access*, 11(3): 49273-49288. DOI: 10.1109/ACCESS.2023.3276450

[23] Zhang X., Zhang J. (2023). Analysis and research on library user behavior based on apriori algorithm (基于 Apriori 算法的图书馆用户行为分析与研究) [J]. *Measurement: Sensors*. 27(12): 100802-100812. DOI: 10.1016/j.measen.2023.100802

[24] Zubair M., Iqbal M.A., Shil A., Chowdhury M.J., Moni M.A., Sarker I.H. (2024). An improved K-means clustering algorithm towards an efficient data-driven modeling. *Annals of Data Science*. 11(5): 1525-1544. DOI: 10.1007/s40745-022-00428-2