International Journal for Housing Science and Its Applications
Publish August 6, 2025. Volume 46, Issue 3 Pages 8189-8202

https,//doi.org/10.70517/ijhsa463700

A study on performance evaluation and improvement of transportation electromechanical systems through big data analysis techniques

Zhiwei Luo[1] and Junyi Li[2,*]
[1] Guangxi Fuhe Expressway Co., Ltd., Hezhou, Guangxi, 542800, China
[2] Guangxi Transportation Science and Technology Group Co., Ltd., Nanning, Guangxi, 530000, China
Corresponding authors: (e-mail: 17630259169@163.com).

Abstract As a core component of urban intelligent transportation infrastructure, the performance of transportation electromechanical system is directly related to the efficiency and safety of transportation operation. In this paper, an improved Canopy-K-means clustering algorithm is proposed to categorize traffic E&M systems, and a performance evaluation method is constructed based on the random forest model. The improved clustering algorithm adopts the “median and maximum distance product method” to determine the initial clustering center, and reduces redundant operations by optimizing the distance calculation. At the same time, a random forest evaluation model is established based on the driving performance index system to scientifically evaluate the performance of the electromechanical system. The experimental results show that the improved Canopy-K-means algorithm achieves an average accuracy of 83.48% on six UCI datasets, which is 5.85% higher than the traditional K-means algorithm; the running time is 169.53ms, which is 35.84% shorter than the traditional algorithm. The random forest model performs well in the evaluation, with an AUC value of 0.951 for the ROC curve and a KS value of 0.8044, which is significantly better than the traditional methods such as logistic regression. The SHAP analysis reveals that the features contributing the most to the evaluation are the absolute maximum of longitudinal acceleration, the mean value of longitudinal velocity, and the standard deviation of the angle of the heading angle from the centerline of the lane. This study provides an effective method for accurate classification and scientific assessment of transportation electromechanical systems.

Index Terms big data analysis, transportation electromechanical systems, Canopy-K-means clustering, random forest, performance evaluation, feature contribution

I. Introduction

With the deepening of reform and opening up, China's national economy continues to develop rapidly through the development of highway grass-roots construction is increasingly expanding and perfect, highway traffic tunnel has become one of the main forms of transportation in China [1]. Transportation electromechanical system is the indispensable foundation of the new infrastructure of transportation, is an important means to improve the efficiency of transportation, but also an important decision-making basis for road emergency command and rescue [2], [3]. As an important part of modern traffic management, the traffic electromechanical system covers signal control, monitoring equipment, communication module and energy management system, etc., which makes the traffic and road data information transmitted in the relevant urban rail transit, and thus ensures the safety and stability of the traffic system [4]-[7].

The construction of highway E&M systems in China has made some progress, for example, highway ETC non-stop electronic toll collection system, tunnel ventilation and lighting and monitoring system and other E&M systems have been widely used [8]. However, the maintenance level of transportation E&M systems has failed to match the development of construction technology. In the operation process of highway transportation, there is a common phenomenon of "reconstruction and light maintenance" and "construction instead of maintenance", which ignores the problems that may arise in the operation process and ignores the health monitoring and evaluation of the electromechanical system in order to rapidly increase the mileage [9]-[11]. At this stage, many highway traffic equipment management in China still rely on manual inspection, recording and statistical analysis, which is cost-effective and inefficient, and the same is true for the inspection and maintenance of traffic electromechanical systems [12], [13]. By establishing a standardized, digital and intelligent performance evaluation model, it can assist managers in assessing the health status of electromechanical systems, improve the efficiency of analysis, and realize the cost reduction and efficiency of highway traffic [14]-[16].

Transportation electromechanical system is an important part of modern intelligent transportation system, and its performance level directly affects the safety, efficiency and comfort of transportation operation. The traditional method of performance evaluation of transportation electromechanical systems mainly relies on manual experience judgment, and there are problems such as strong subjectivity, poor consistency and single evaluation dimension. In recent years, the rapid development of big data technology provides new ideas and methods for the performance evaluation of transportation electromechanical systems. Through the mining and analysis of massive traffic data, the performance of electromechanical systems can be objectively evaluated from multiple dimensions and perspectives, providing a scientific basis for system optimization and improvement.

Existing studies have made some progress in traffic data clustering and performance evaluation. In data clustering, K-means algorithm is widely used because of its simplicity and high efficiency, but it is sensitive to the selection of initial centroids and needs to predetermine the number of clusters.Canopy algorithm does not need to preset the number of initial clusters, and it is fast, but with lower accuracy.DBSCAN algorithm can recognize clusters of arbitrary shapes, but it is sensitive to the selection of parameters and has high computational complexity. In terms of system performance evaluation, traditional methods mostly use linear models or simple classification algorithms, which are difficult to deal with complex nonlinear relationships between features, and the accuracy and interpretability of evaluation results are limited.

The data of transportation electromechanical systems are characterized by high dimensionality, heterogeneity and dynamic changes, and the existing single clustering algorithms and assessment models are difficult to meet the practical needs. How to combine the advantages of different algorithms to construct an efficient and accurate clustering method and how to build an accurate and interpretable performance evaluation model are the key issues to be solved in the current research.

In this study, an improved Canopy-K-means clustering algorithm and a performance evaluation method based on random forest are proposed to address the above problems. Firstly, the traditional Canopy algorithm is improved, and the “median and maximum distance product method” is used to replace the random selection method to determine the initial clustering centers, so as to improve the stability and accuracy of clustering; secondly, the K-means algorithm is optimized, and the distance determination mechanism is introduced to reduce the redundant distance computation, so as to improve the operation efficiency of the algorithm; and then the improved Canopy algorithm is combined with the optimized K-means algorithm to construct the improved K-means algorithm, and the performance evaluation method based on random forests is proposed. means algorithm to construct the improved Canopy-K-means clustering algorithm to realize the accurate classification of traffic electromechanical systems; finally, based on the driving performance feature index system, a random forest assessment model is established to scientifically assess the performance of different categories of traffic electromechanical systems and the contribution of each feature to the assessment results is analyzed by the SHAP method. Through comparative experiments and application validation on the real UCI dataset, the effectiveness and practicability of the proposed method are comprehensively evaluated, and a scientific basis is provided for the enhancement and optimization of the performance of traffic electromechanical systems.

II. Improved Canopy-K-means clustering for classification of transportation electromechanical systems

II. A.K-means clustering

II. A. 1) Basic K-means algorithm

K-means algorithm is the most widely used clustering algorithm who is a clustering technique based on the prototype of creating a single level division of data objects. It attempts to discover clusters of user-specified number (k) [17]. The prototype of K-means algorithm is center of mass defined whose center of mass is the average of all points in the cluster. K-means algorithm is commonly used for objects in n-dimensional continuous space.

II. A. 2) Highlights of the K-means Algorithm

(1) Assigning points to the nearest center of mass

In order to be able to assign data points to the closest center of mass, a proximity measure is needed to give numerical representation to the concept of “closest”. For a given sample $x^{(i)} = \{x_1^{(i)}, x_2^{(i)}, \ldots, x_n^{(i)}\}$ and $x^{(j)} = \{x_1^{(j)}, x_2^{(j)}, \ldots, x_n^{(j)}\}$, where $i, j = 1, 2, \ldots, m$ is the number of samples and n is the feature. The distance measures are ordered distance measure, unordered distance measure and mixed attribute distance measure. The specific measures are as follows:

a) Ordered attribute distance metrics

Minkowski distance, as in the following equation:

$$dist_{mk}(x^{(i)},x^{(j)})=\left(\sum_{u=1}^{n}|\ x_u^{(i)}-x_u^{(j)}\ |^p\right)^{\frac{1}{p}} \tag{1}$$

The Euclidean distance, i.e., $dist_{mk}(x^{(i)},x^{(j)})$ when p = 2, is given in the following equation:

$$dist_{ed}(x^{(i)},x^{(j)})=\|\ x^{(i)}-x^{(j)}\ \|_2=\sqrt{\sum_{u=1}^{n}|\ x_u^{(i)}-x_u^{(j)}\ |^2} \tag{2}$$

The Manhattan distance, i.e., $dist_{mk}(x^{(i)},x^{(j)})$ when p = 1, is given in the following equation:

$$dist_{man}(x^{(i)},x^{(j)})=\|\ x^{(i)}-x^{(j)}\ \|_1=\sum_{u=1}^{n}|\ x_u^{(i)}-x_u^{(j)}\ | \tag{3}$$

b) The disordered attribute distance metric, VDM, is given in the following equation:

$$VDM_p(x_u^{(i)},x_u^{(j)})=\sum_{z=1}^{k}\left|\frac{m_{u,x_u^{(i)},z}}{m_{u,x_u^{(i)}}}-\frac{m_{u,x_u^{(j)},z}}{m_{u,x_u^{(j)}}}\right|^p \tag{4}$$

where $m_u,x_u^{(i)}$ denotes the number of samples taking value $x_u^{(i)}$ on attribute u, $m_u,x_u^{(i)},z^a$ denotes the number of samples taking value $x_u^{(l)}$ on attribute u in the zth sample cluster, and $VDM_p(x_u^{(l)},x_u^{(j)})$ denotes the VDM distance between two discrete values $x_u^{(l)}$ and $x_u^{(j)}$ on attribute u.

c) The mixed attribute distance measure, i.e., a combination of ordered and unordered, is given in the following equation:

$$MinkovDM_p(x_u^{(i)},x_u^{(j)})=\left(\sum_{u=1}^{n_c}|\ x_u^{(i)}-x_u^{(j)}\ |^p-\sum_{u=n_c+1}^{n}VDM_p(x_u^{(i)},x_u^{(j)})\right)^{\frac{1}{p}} \tag{5}$$

which contains n_c ordered attribute, with $n-n_c$ unordered attributes.

(2) Center of mass and objective function

The K-means algorithm step requires updating the center of mass of the clusters, which can vary depending on the clustering objective and with the data proximity measure. Clustering goals are often expressed using an objective function that depends on the proximity of points to each other, or points to the cluster's center of mass. If the objective function and the proximity measure are given, the center of mass can be computed by a mathematical formula. The data studied in this paper are in Euclidean space.

Euclidean space: when the proximity metric is the data with Euclidean distance, the objective function for the quality of its clusters uses the sum of squares of the errors (SSE) as a metric.SSE is defined in the following equation:

$$SSE=\sum_{i=1}^{k}\sum_{x\in c_i}dist_{ed}(c_i,x)^2 \tag{6}$$

where dist_ed is the standard Euclidean distance between two objects in Euclidean space.

The center of mass that minimizes the SSE of a cluster is the mean value. The center of mass of the ith cluster is defined in the following equation:

$$c_i=\frac{1}{m_i}\sum_{x\in c_i}x \tag{7}$$

II. B.Canopy Clustering Algorithm

II. B. 1) Basic Canopy Algorithm

Canopy algorithm is different from traditional clustering algorithms in that its algorithmic similarity measure uses a simple and shorter computation time method, in which similar objects are stored in the same subset called Canopy, and different numbers of Canopy are obtained through a series of computations, and the Canopy may overlap with

each other, but there is no situation in which a certain object doesn't belong to any Canopy, which can be regarded as data preprocessing [18]. The biggest feature of Canopy algorithm is that it does not need to preset the number of initial clusters, and the clustering results obtained by Canopy algorithm have lower accuracy, but its operation speed is much faster than other cluster analysis algorithms, so it can be more convenient to use in practical applications.

II. B. 2) Steps of the Canopy Algorithm

The basic steps of Canopy algorithm are as follows:

(1) Randomly sort the initial dataset to generate a list L=[x1,x2,... ,xm], and without changing the order of the data in L, determine two distance thresholds T_1 and T_2 where $T_1 > T_2$ by cross-validating the tuning parameter or prior knowledge.

(2) Randomly select a sample data P in the data list L, set P as the center of mass of the first Canopy sub-level, and at the same time delete P from L.

(3) Randomly draw a sample data Q in L and compute the distance from Q to the center of mass of each existing Canopy, and choose the smallest value of these distances D. If D ≤ T_1, store Q in the Canopy with the smallest distance from it and attach a weak marker to it; if D ≤ T_2, store Q in the Canopy with the smallest distance from it, attach a strong marker to it and update the centers of all strongly marked samples' center positions are updated as the center of mass of that Canopy and Q is removed from list L. If D > T_1, Q is set as the new Canopy center of mass and Q is removed from list L.

(4) Repeat step 3 until the number of samples in list L is zero.

II. C.Improved Canopy-K-means clustering for data classification

II. C. 1) Principles of Canopy Algorithm Improvement

In order to prevent the random selection of the radius of the Canopy region, t1 and t2, and the initial center point, reference is made to the “Mean and Maximum Distance Product Method”. “Mean and Maximum Distance Product Method”, this paper adopts an improved Canopy algorithm based on “Median and Maximum Distance Product Method” to determine the initial clustering center and improve the accuracy of clustering results. In the past, the average value was chosen to participate in the calculation, but the average value is easily affected by the maximum and minimum values in the data set. Considering the problem of accuracy, this paper chooses to adopt the median instead of the average value.

The operating principle of the "Median and Maximum Distance product method" is as follows:

Input: Dataset $S = \{s_i \mid s_i = (x_1, x_2, ..., x_n), i = 1, 2, 3...n\}$ containing n data objects;

Output: Clustering result set $R = \{r_i, i = 1, 2, 3..., k\}$

Firstly, calculate the data center point C of the data set S. Select the data object farthest from point C as the first clustering center z_1 and add it to the set Z. Next, select the data object z_2 that reaches the maximum value of the product of the distances to points C and z_1 respectively as the second cluster center and add it to the set Z; Then, calculate the remaining data objects x_1 in set S in sequence. Take point C and the data object z_2 with the maximum value of the product of the distances of each point in set Z collected in set S as the next clustering center and add it to set Z. In this way, k initial clustering centers can be obtained successively.

II. C. 2) Principles of K-means algorithm improvement

After the “coarse” clustering operation performed by the Canopy algorithm, the dataset is roughly divided into k clusters, and only needs to be tightened by further “fine” clustering operation performed by the K-means algorithm. However, the traditional k-means algorithm needs to calculate the distance from each data object to the centers of all clusters, which consumes a large amount of unnecessary running time in each iteration. In order to avoid a large number of redundant calculations, this paper proposes an improved method for the traditional k-means algorithm.

Algorithm improvement idea:

For a data point that can be determined as this cluster class, there is no need to calculate its distance to the center of mass of all other cluster classes, thus it can reduce a lot of redundant calculations.

After Canopy clustering the dataset has been roughly divided into k cluster classes;

Definition 1: There exists a distance $d(x, z_i) = Dist_{\min}(i)$ from data object x_1 to its nearest cluster center, where $Dist_{\min}(i)$ denotes that data point x_1 is the largest of all minimum distances;

Definition 2: Calculate the distance $d(x_i, z_i)$ of each data object x_1 to its nearest clustering center, if $d(x_i, z_i) \le Dist_{\min}(i)$ is satisfied, the data point is retained in the initial class cluster; if $d(x_i, z_i) > Dist_{\min}(i)$, the

distance $d(x_i, z_j)$ of the data point to each cluster center $z_j(1 <= j <= k)$ is to be calculated and the data point is assigned to the nearest cluster.

During the computation of the algorithm, some of the data points from each iteration of the algorithm will remain in the old cluster and will not be fully allocated, in other words, these retained data points will not be involved in the computation. In other words, these reserved data points will not be involved in the computation. This saves the total computing time of the algorithm, and thus improves the efficiency of the algorithm.

II. C. 3) Principles of the Improved Canopy-K-means Algorithm

Synthesizing the improved principles of the previous Canopy algorithm as well as the K-means algorithm, the original Canopy-K-means algorithm is further improved as:

Input: dataset containing n data objects $S = s_i \mid s_i = (x_1, x_2, ..., x_n), i = 1, 2, 3...n$;

Output: clustering result set $R = r_i, i = 1, 2, 3..., k$.

(1) For the data set S, the median and maximum distance product method is used to obtain the copy center point set Z, and the elements in Z are the k initial clustering centers $z_1, z_2, ..., z_k$ belonging to R.

(2) Obtain preliminary clustering results by clustering by Canopy algorithm, i.e., divide into k canopy, and perform K-means clustering on the data set S after performing preliminary clustering.

(3) For any data object x_i in the Canopy set, calculate the distance $d(x_i, z_j)(1 <= i <= n), (1 <= j <= k)$ between all k cluster centers $z_j(1 <= j <= k)$ and the data object, and assign the data object to the nearest cluster.

(4) Each iterative computation needs to determine whether the distance $d(x_i, z_i)$ from data point x_i to the center is less than or equal to $Dist_{\min}(i)$. If $d(x_i, z_i) <= Dist_{\min}(i)$ is satisfied, point x_i does not need to compute the distance to each of the other centers, and the data point is directly retained in the initial clusters; if $d(x_i, z_i) > Dist_{\min}(i)$ is satisfied, an iterative computation is performed on it to determine the cluster class to which it belongs.

(5) For each cluster $j(1 <= j <= k)$, recalculate the clustering center.

(6) Repeat steps (3), (4) and (5).

(7) Satisfy the convergence criterion.

(8) Form k new clusters and output the clustering result $R = r_i, i = 1, 2, 3, ..., k$.

II. D.Experimental results and analysis

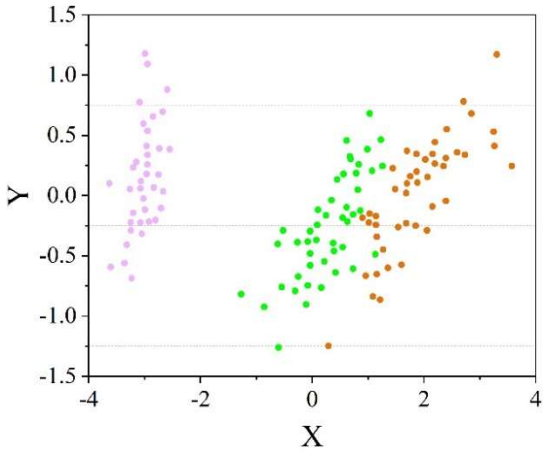
In order to observe the clustering effect of the improved Canopy-K-means algorithm, this experiment first uses 6 real UCI datasets for the comparison of Kmeans clustering, Canopy clustering, Canopy-K-means clustering, and DBSCAN clustering experiments. The datasets include Time, Speed, Distance, Kilometer, Vehicle, Cost. the information of these 6 datasets is shown in Table 1. As can be seen from the table, the real dataset Time is the time data, there are 155 sample data, each data has 6 feature attributes, and the number of real classes is 4; Speed is the speed dataset, containing 183 samples, each data has 17 feature attributes, and the number of real classes is 4; Distance is the driving distance dataset, containing 570 samples, each data has 33 feature attributes with a true class number of 3; Kilometer is the driving speed dataset, containing 109 samples, each with 8 feature attributes, and a true class number of 7; Vehicle is the vehicle dataset, containing 214 sample data, each with 5 features, and a true class number of 4; and Costs is the vehicle toll identification dataset, containing 219 sample data, each with 8 features and the number of true classes is 7.

Table 1: 6 UCI real data set information

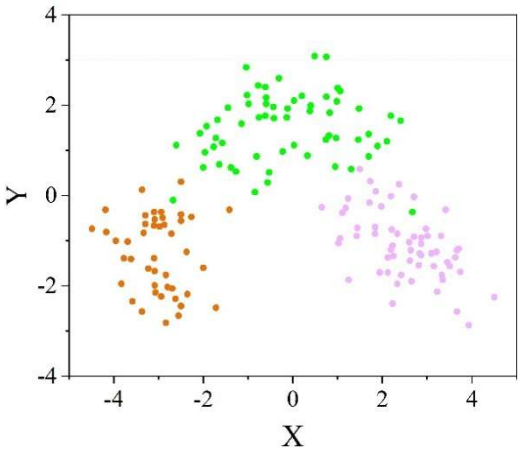
Data set	Sample size	Data feature set	Number of categories
Time	155	6	4
Speed	183	17	4
Distance	570	33	3
Kilometer	109	8	7
Vehicle	214	5	4
Cost	219	8	7

The above datasets were observed by dimensionality reduction using StandardScaler() normalization method and PCA dimensionality reduction method of sklearn library for python. The distribution of the six datasets after dimensionality reduction is shown in Figure 1. As can be seen from the figure, the spatial distribution of the real

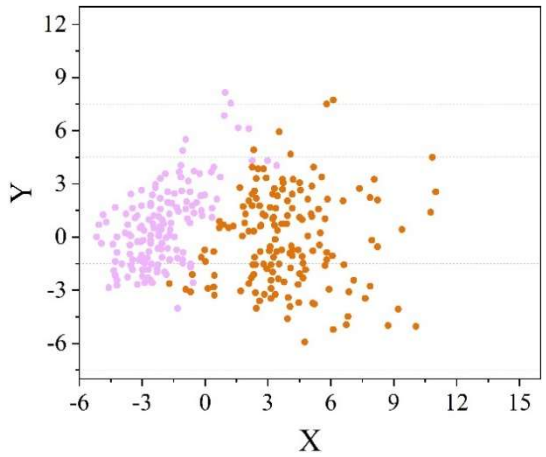
datasets is complex, in which some of the datasets contain overlapping data, such as the Distance dataset, and some of the datasets have multiple outliers, such as the Kilometer and Cost datasets. These two data distributions correspond to the two situations when the correlation between experts' results is low and high in practice, respectively, and are suitable for testing the practical effect of the improved Canopy-K-means algorithm.



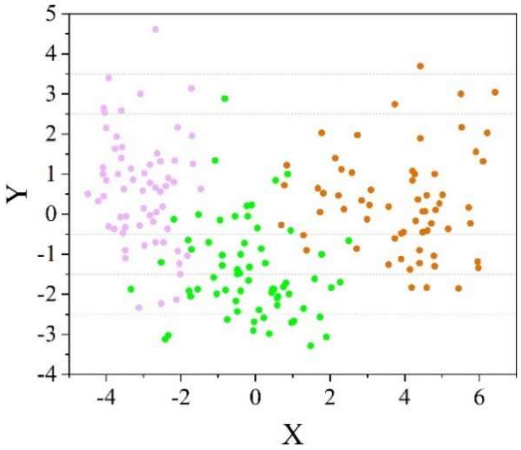
(a) Iris distribution



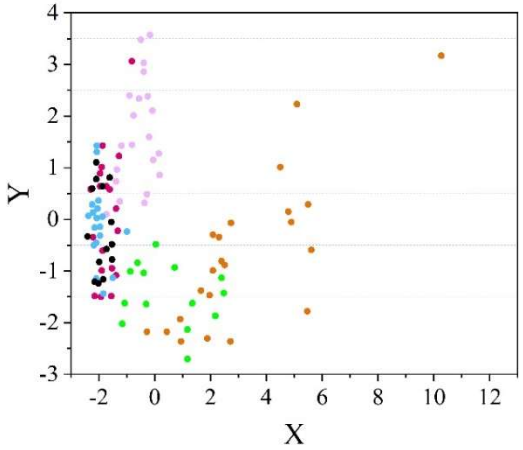
(b) Wine distribution



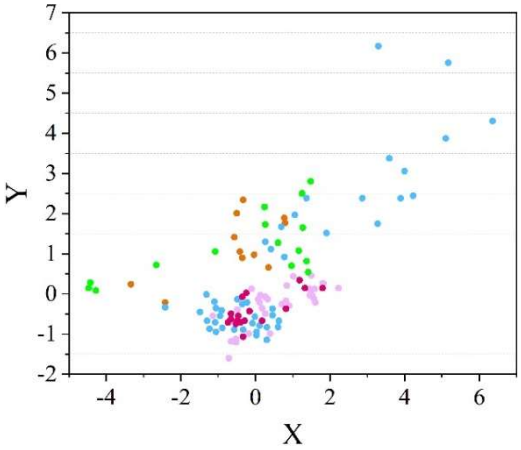
(c) Wdbc distribution



(d) Seeds distribution



(e) Breasttissue distribution



(f) Glass distribution

Figure 1: Six experimental data set two-dimensional distribution

Conventional K-means clustering experiments are carried out on the above datasets first. In order to get the optimal K-means clustering effect, the actual class number of each dataset is selected as the k-value, and k centroids are randomly generated, and the obtained K-means clustering experiment results are shown in Table 2. As can be seen from the table, after selecting the actual number of classes as the k value, the clustering results of K-means algorithm on Time, Speed and Distance are 91.37%, 99.28% and 93.41%, respectively, with high accuracy, while on the two high-dimensional datasets of Kilometer and Cost, the accuracy is 62.54% and 50.27%, which performs effect is average.

Table 2: K-means clustering experiment results

Data set	Class number	Sample size	Accuracy rate (%)	Running time (ms)
Time	4	155	91.37	25.11
Speed	4	183	99.28	18.174
Distance	3	570	93.41	55.209
Kilometer	7	109	55.83	37.15
Vehicle	4	214	62.54	50.78
Cost	7	219	50.27	88.275

Then regular Canop and Canopy-K-means clustering experiments are performed on the six UCI datasets, and the results of the Canopy clustering experiments are shown in Table 3 and the results of the Canopy-K-means clustering experiments are shown in Table 4. As can be seen from Table 3, Canopy algorithm clustering speed is very fast but the accuracy of the clustering results is low. the accuracy of Canopy algorithm on the six data sets are 68.97%, 71.28%, 83.52%, 37.91%, 66.19% and 45.27% respectively, which are not more than 80% but the running time is within 1~2.05ms, which are not more than 3ms. As can be seen from Table 4, the accuracy of Canopy-K-means algorithm on the six datasets is 91.39%, 98.22%, 91.84%, 55.43%, 93.75% and 53.76%, respectively, and the highest accuracy reaches 98.22%, compared with Canopy clustering on Iris, Speed, Distance, Kilometer, Vehicle and Cost datasets, the accuracy is improved by 22.42%, 26.94%, 8.32%, 17.52%, 27.56% and 8.49%, respectively. While the running time is 15.17, 14.22, 53.39, 15.17, 33.33 and 33.28ms respectively, the time consuming performance is not as good as the coarse clustering algorithm.

Table 3: The results of the canopy cluster experiment

Data set	Class number	Sample size	Accuracy rate (%)	Running time (ms)
Time	4	5	68.97	1.0
Speed	4	4	71.28	2.0
Distance	3	4	83.52	2.05
Kilometer	7	7	37.91	1.05
Vehicle	4	5	66.19	1.02
Cost	7	6	45.27	1.0

Table 4: Canopy-K-means clustering experiment results

Data set	Class number	Sample size	Accuracy rate (%)	Running time (ms)
Time	4	4	91.39	16.17
Speed	4	4	98.22	16.22
Distance	3	3	91.84	55.44
Kilometer	7	7	55.43	16.22
Vehicle	4	4	93.75	34.35
Cost	7	7	53.76	34.28

In addition to K-means algorithm and Canopy algorithm, DBSCAN algorithm based on density clustering is selected for clustering experiments on six datasets in this paper. Firstly, DBSCAN clustering on Iris was performed, and in order to select the optimal clustering radius and minimum sample size, several value experiments are needed.The DBSCAN algorithm selects six experimental groups with different clustering radius R on the Iris dataset, and the minimum sample size is incremented by two sample points each time. The clustering radius is incremented by 0.1 between the different experimental groups, and the corresponding accuracy folding statistic graphs of radius and minimum sample size of DBSCAN are plotted as shown in Fig. 2. As can be seen in Fig. 2, there are more

obvious accuracy peaks at radius values of 0.4, 0.6 and 0.9, when the minimum sample size is 4. Among the six groups, the accuracy of the group with a radius of 0.8 has a more stable accuracy performance across sample sizes than the other groups. Therefore, the optimal radius value for DBSCAN clustering on the Iris dataset is 0.8 and the minimum sample size is 4.

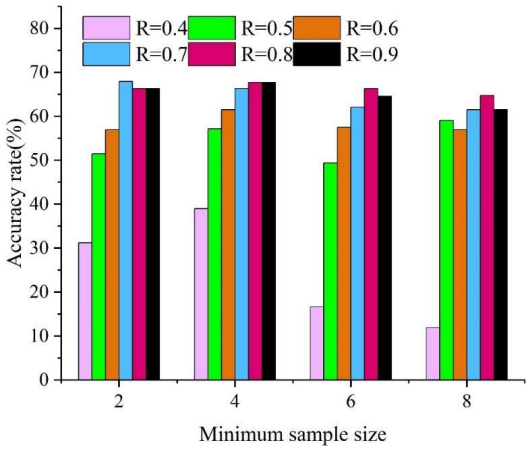


Figure 2: The dbscan clustering performance of different radius and minimum sample size

After the radius and minimum sample size accuracy comparison experiments were conducted on other UCI datasets except Speed using the same method, the optimal radius and minimum sample size values corresponding to each dataset were obtained as shown in Table 5. The experimental results obtained from DBSCAN clustering using the optimal radius and minimum sample size for each dataset in Table 5 are shown in Table 6. As can be seen from Table 6, the number of classes of the clustering result of the DBSCAN algorithm has a large difference from the actual number of classes in some datasets such as Distance and Kilometer, and the performance is unstable. In addition, the DBSCAN algorithm not only needs to determine the cluster radius and minimum sample size before clustering, but also has an accuracy of 69.58%, 59.18%, 63.22%, 23.38%, 59.01%, and 35.49% on the six datasets, with a maximum of 69.58%, which makes the performance of clustering result accuracy less than that of the Canopy algorithm. The running time ranged from 1.001 to 13.4ms, and it took longer in some cases.

Table 5: The optimal radius and minimum sample size of DBSCAN

Data set	Radius	Minimum sample
Time	0.9	5
Speed	2.5	4
Distance	4.9	2
Kilometer	1.8	2
Vehicle	1.2	8
Cost	1.2	3

Table 6: DBSCAN cluster results

Data set	Class number	Sample size	Accuracy rate (%)	Running time (ms)
Time	4	3	69.58	2.005
Speed	4	4	59.18	4.174
Distance	3	8	63.22	13.412
Kilometer	7	13	23.38	1.005
Vehicle	4	3	59.01	1.001
Cost	7	9	35.49	2.005

Finally, the improved Canopy-K-means clustering experiments are performed on the real UCI dataset and the results are shown in Table 7. As can be seen from the table, the accuracy of the improved Canopy-K-means clustering on the six datasets is 92.38%, 98.73%, 93.57%, 58.62%, 92.06%, and 52.74% respectively, with the

highest being 98.73%. The running times were: 17.23, 17.23, 53.14, 18.01, 23.63 and 45.09ms, ranging from 16 to 53ms.

Table 7: Improved Canopy-K-means clustering experiment results

Data set	Class number	Sample size	Accuracy rate (%)	Running time (ms)
Time	4	4	92.38	17.23
Speed	4	4	98.73	17.23
Distance	3	3	93.57	53.14
Kilometer	7	7	58.62	18.01
Vehicle	4	4	92.06	23.63
Cost	7	7	52.74	45.09

The experimental results of the improved Canopy-K-means algorithm on the UCI dataset are statistically compared with the previously obtained experimental results of K-means, Canopy, DBSCAN and Canopy-K-means algorithms, as shown in Table 8. As can be seen, for the total number of classes of clustering results, the total number of classes of clustering results of DBSCAN differs from the actual total number of classes, and the performance of clustering in some datasets is unstable, while the number of classes of clustering results generated by other algorithms differs from the real number of classes to a lesser extent. And the average accuracy of K-means, Canopy-K-means and improved Canopy-K-means algorithms on the six datasets is high, which is above 70%, and the improved Canopy-K-means algorithm has the highest average accuracy, which is 83.48%.The total running time of the K-means algorithm is the highest, which is 264.22ms while the Canopy-K-means algorithm and the improved Canopy-K-means, due to the fact that both use Canopy clustering to optimize the selection of centroids, both have a total running time of around 169.53ms on the six datasets, which is an improvement compared to the K-means algorithm.

Table 8: The different algorithms were compared in the results of the UCI data set

Clustering algorithm	Actual total class number	The total number of clustering results	Average accuracy(%)	Total running time(ms)
K-means	25	25	77.63	264.22
Canopy	25	25	57.94	10.55
DBSCAN	25	25	52.54	24.13
Canopy-K-means	25	25	79.87	163.11
Improved Canopy-K-means	25	25	83.48	169.53

III. Random Forest-based Performance Evaluation of Transportation Electromechanical Systems

III. A. Introduction to the Random Forest Model

III. A. 1) Decision tree model

Decision tree is a classic machine learning method, is the representative of the “divide and conquer” idea, can be applied to classification problems and regression problems, this paper will mainly use classification decision tree. The main process of the decision tree algorithm is as follows:

Algorithm input: training set $D=\{(x_1,y_1),(x_2,y_2),...,(x_m,y_m)\}$, feature set $A=\{a_1,a_2,...,a_d\}$

Algorithm flow:

1: Generate node node;

2: {:} If all the samples in the training set D belong to the same class C, mark node as a leaf node of class C and end the process;

3: If the feature set A is empty or all the features in the training set D take the same value, mark the node using the class with the largest percentage of nodes and end the process;

4: Select the optimal division feature a_* from the feature set A. Using the values of all the samples in A in the a_* feature, generate the corresponding branch. If the sample subset on the child node is empty, mark it as a leaf node, the category is the class with the largest proportion of node, and end the process; if the sample subset is not empty, repeat steps 1~4 for the child node; Algorithm output: a decision tree with node as the root node.

III. A. 2) Optimal branching strategy

The key to the realization of the decision tree algorithm is to choose the optimal branching method for each node, the purpose of branching is to make the child nodes contain only samples of the same category as far as possible, i.e., we hope that the "impurity" of the child nodes is as low as possible. The lower the impurity, the better the classification of the decision tree [19]. As the hierarchical level is deepened, the impurity of the child nodes must be lower than that of the parent nodes, so for each decision tree, the impurity of the leaf nodes must be the lowest.

III. A. 3) Random forest characteristics

The base evaluator of the Random Forest algorithm is the decision tree, and its basic idea is to construct a large number of weak decision tree classifiers in parallel, and take the average or majority vote on the output of each decision tree to output a strongly learned prediction.

In order to improve the accuracy of the final prediction output of the random forest, an important idea is to try to make the base classifiers decision trees independent of each other and maintain the variability, which can be obtained by trying to use as many different training sets as possible for base classifier training. In general, the bagging method can be used to form different training data by random sampling with put-back, i.e., when preparing a training set for a decision tree, one sample is obtained at a time and that sample is put back into the original training set before the next sample is drawn. In this way, the self-help training sets are different from each other each time, which ensures the mutual independence of the decision trees.

III. A. 4) Characterization contributions

Since only one feature is used as a branching benchmark at each node of the decision tree and the difference between the classification situations of the parent and child nodes can be computed, this difference is regarded as the change induced by the feature of the parent node, characterized by the degree of feature contribution. In the random forest model, the increment of probability induced by a feature in all the decision trees is calculated and averaged, which is the predicted contribution of this feature in the random forest.

$$A_{mean_i} = \frac{1}{m}\sum_{k=1}^{m} A_k \tag{8}$$

A_{mean_i} denotes the distribution of the training set samples in node i, m denotes the number of samples in the current node, and A_k denotes the value of the sample of the kth on the current node.

$$LS_{i,Fj} = A_{mean\_C} - A_{mean\_D} \tag{9}$$

$LS_{i,Fj}$ denotes the local increment at node i caused by feature F_j, expressed as the difference between the mean distributions of the child and parent nodes, where C denotes the child node and D denotes the parent node.

In order to determine the overall contribution of a feature to the classification process for a given sample, all local increments caused by that feature need to be calculated and summed.

$$S_{i,Fj} = \frac{1}{T}\sum_{i=1}^{n} LS_{i,Fj} \tag{10}$$

$S_{i,Fj}$ denotes the amount of overall feature contribution of feature F_j to some sample i, T denotes the number of decision trees in the random forest model, and n denotes the number of all nodes.

III. B. Random Forest-based Performance Evaluation Methods

III. B. 1) Hyperparameter selection

(1) Impurity measure

For each decision tree, the information entropy is selected as a measure of impurity in the branching strategy.

(2) Maximum depth of decision tree

The decision tree grows until the impurity is optimal or there are no features available. When no restriction is placed on the growth of the decision tree, the decision tree has a strong tendency to overfit, i.e., it performs well on the training set but poorly on the test set. This is due to the fact that the sample data contains the sample's idiosyncratic noise in addition to the overall characteristics of the data.

(3) Number of decision trees

The number of decision trees of random forest is the number of base evaluators, the more base evaluators, the higher the correctness of the output of the integrated algorithm. Generally speaking, the more the number of

decision trees, the better the effect of the random forest model. When the number of decision trees is increased to a certain level, the accuracy of the Random Forest tends to stop rising or fluctuate, and the benefit of further increasing the number of decision trees decreases.

III. B. 2) Modeling

According to the driving performance characteristic index system established in this paper, the corresponding characteristics are calculated for the driving course of the driving simulator test as shown in Table 9, as a random forest feature, and the driving experience is used to classify the drivers into three categories of skillful, typical, and rusty, which correspond to excellent, good, and average driving performance.

Table 9: Characteristics of driving performance

Serial number	Categories	Characteristic index
H1	Driving adherence	The absolute value of the mean of the lateral migration
H2		The standard deviation of the lateral migration
H3		The absolute value of the mean of the Angle of the heading Angle and the center of the lane
H4		Standard deviation of the Angle of course Angle and lane
H5		Traffic probability
H6		The mean of longitudinal velocity
H7	Sports comfort	Standard deviation of longitudinal velocity
H8		The maximum value of the vertical acceleration
H9		The maximum value of the horizontal acceleration

For the whole random forest model, each driving performance evaluation combines the evaluation results of all 100 decision trees and outputs the categorical expectation as the driving performance evaluation result of the whole random forest model.

III. C. Assessment results and analysis

III. C. 1) Random Forest Algorithm Results and Analysis

Random Forest is an integrated learning method by training multiple decision trees simultaneously and then combining their results to make a final prediction. The Random Forest model ROC graph is shown in Figure 3. In the figure, the performance of the Random Forest model in identifying fraud samples on the test set is visualized, with an AUC value of 0.951. This indicates that the Random Forest has a stronger performance in identifying fraud samples compared to a single decision tree. However, it is important to note that the AUC of the random forest model on the training set is significantly higher than that on the test set, which may exhibit the phenomenon of overfitting.

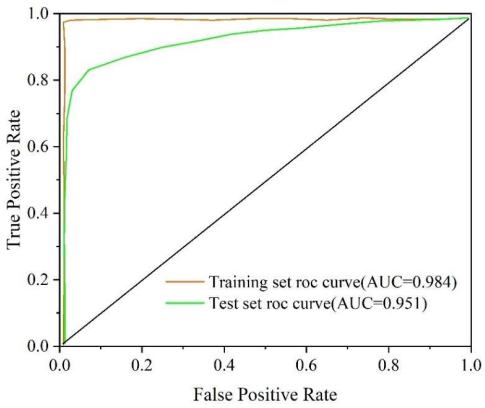


Figure 3: ROC graph of stochastic forest model

The random forest model Lorenz plot is shown in Figure 4. In the figure, the difference between the cumulative distribution function of the model on positive and negative examples is clearly presented, while the KS value is 0.8044. Compared with the Logistic model's 0.719 and the Single Decision Tree's 0.673 as well as the Support Vector Machine's 0.623, the Random Forest's KS value is significantly higher, which further verifies that Random Forest achieves a significant improvement. However, a significant problem with random forests is overfitting. This

means that the model performs well on the training set but has poor generalization ability on unseen data. Overall, the Random Forest also shows significant improvement in its ability to discriminate between positive and negative samples, which is a relatively high improvement over the traditional model as well as the single decision tree model, and the Random Forest is able to better discriminate between fraudulent and non-fraudulent samples.

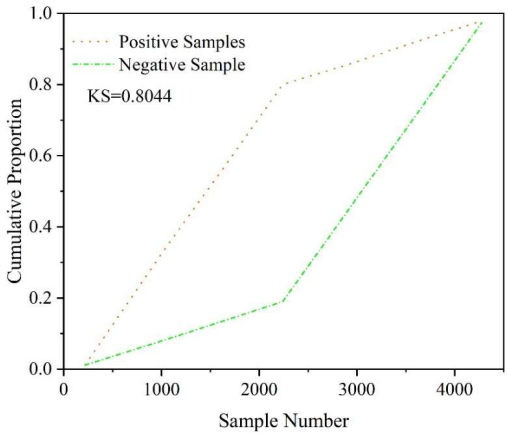


Figure 4: Random forest model lorentz curve

III. C. 2) Characteristic contribution analysis

The Gain, Weight and Cover values of the model features are shown in Table 10. Global attribution analysis is an evaluation of the feature contribution of the entire model, which involves the overall model performance and the association of features. In stochastic models, the feature attribution methods Gain, Cover, and Weight belong to the global attribution methods, which are used to reflect the change in the expected accuracy of the model when a set of features is removed, i.e., the effect of the features on the global accuracy.Gain reflects the effect of the feature variables on the accuracy of the model. Specifically, Gain is measured by measuring how much the splitting of each feature in the construction of the tree improves the accuracy of the model. When a set of features is removed, its effect on the overall accuracy of the model is observed.Cover reflects the extent to which the feature variables cover the observations. It is measured by looking at the number of observed objects associated with the feature. When a set of features is removed, observe its effect on the degree of coverage, i.e., how much the features are associated with the diversity of the sample.Weight reflects how often the feature variable is used in the model. It is measured by counting the number of times the feature variable is used in all decision trees. When a set of features is removed, observe its effect on the overall weight of the model, i.e., how important the features are in the model.

Table 10: Gaines and Cover values for model characteristics

Characteristic index	Gain	Weight	Cover
The absolute value of the mean of the lateral migration	2.78	183.29	151
The standard deviation of the lateral migration	1.83	215.35	157
The absolute value of the mean of the Angle of the heading Angle and the center of the lane	1.71	122.83	124
Standard deviation of the Angle of course Angle and lane	3.69	153.63	117
Traffic probability	2.78	146.35	149
The mean of longitudinal velocity	2.22	125.57	124
Standard deviation of longitudinal velocity	1.64	158.43	154
The maximum value of the vertical acceleration	1.69	118.77	139
The maximum value of the horizontal acceleration	2.65	134.58	178

Feature importance calculation methods, including Gain, Cover and Weight, are usually the expectation of importance obtained over the entire training set. However, there are some problems with this method, such as the effect of changing one feature may directly lead to changes in the importance of other features, which makes it difficult to meet the consistency requirement. To address these issues, this paper introduces the SHAP method.SHAP has several advantages such as consistency, local accuracy, and no effect of missing values. It has better performance in explaining the feature contributions of the integrated model. A positive SHAP value indicates that the corresponding feature has a positive effect on increasing the model output, while a negative SHAP value

indicates that the corresponding feature has a positive effect on decreasing the model output.The absolute value of the SHAP value indicates the degree of absolute effect of the corresponding feature on the model output. The larger the value, the greater the influence of the feature on the model output. The article measures the impact of features on the model based on the average SHAP values of the features, and the absolute values of the average SHAP values of the features are ranked in descending order.

Figure 5 shows the absolute values of the SHAP averages of the top 5 features. It can be seen that the top five features in terms of contribution to the model are "the maximum absolute value of longitudinal acceleration", "the mean value of longitudinal velocity", "the standard deviation of the Angle between the heading Angle and the lane centerline", "the absolute value of the mean value of the Angle between the heading Angle and the lane centerline", and "the absolute value of the mean value of lateral offset".

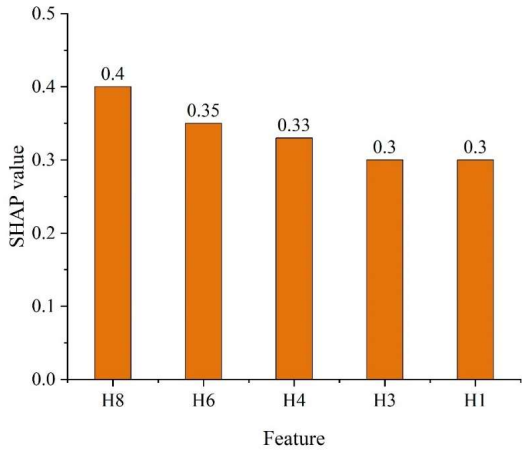


Figure 5: The characteristics of the previous 5 indicators of the index

IV. Conclusion

The following conclusions are drawn from the study of performance evaluation and improvement of transportation E&M systems:

The improved Canopy-K-means clustering algorithm performs well in classifying transportation E&M systems. Experiments show that the algorithm achieves an average accuracy of 83.48% on six UCI datasets, which is a significant improvement over the 77.63% of the traditional K-means algorithm and the 57.94% of the Canopy algorithm. The total running time of the algorithm is 169.53ms, which is 94.69ms lower than the 264.22ms of the traditional K-means algorithm, and the efficiency is significantly improved. The strategy of determining the initial clustering centers by “median and maximum distance product method” effectively solves the instability problem caused by random selection in the traditional algorithm.

The performance evaluation model of transportation electromechanical system based on random forest has high accuracy and interpretability. The AUC value of the model ROC curve reaches 0.951, and the KS value is 0.8044, which is significantly better than the logistic regression model (0.719) and the decision tree model (0.673). The absolute value of the mean value of the angle between the heading angle and the centerline of the lane, and the absolute value of the mean value of the lateral offset. These characteristics mainly reflect the driver's maneuvering stability and driving path accuracy, which are of great significance in evaluating the performance of traffic electromechanical systems.

In summary, the improved Canopy-K-means clustering algorithm and the performance evaluation method based on random forest proposed in this paper provide an effective tool for the accurate classification and scientific evaluation of traffic electromechanical systems, and have important theoretical and practical value for the construction and development of intelligent transportation systems.

References

[1] Zhou, Y., Tong, C., & Wang, Y. (2022). Road construction, economic growth, and poverty alleviation in China. Growth and Change, 53(3), 1306-1332.

[2] Lin, S., Wang, Y., Jia, L., & Zhang, H. (2018). Reliability assessment of complex electromechanical systems: A network perspective. Quality and Reliability Engineering International, 34(5), 772-790.

[3] Jain, N. K., Saini, R. K., & Mittal, P. (2019). A review on traffic monitoring system techniques. Soft computing: Theories and applications: Proceedings of SoCTA 2017, 569-577.

[4] Day, C. M., & Li, H. (2016). Implementation of automated traffic signal performance measures. Institute of Transportation Engineers. ITE Journal, 86(8), 26.
[5] Zhang, Y., & Yang, B. (2020). Traffic flow detection using thermopile array sensor. IEEE Sensors Journal, 20(10), 5155-5164.
[6] Gao, Y., & Chen, X. N. (2019). Comparative Research of Electromechanical Design Schemes for Highway Tunnels between China and Eastern Europe. Journal of Highway and Transportation Research and Development (English Edition), 13(3), 70-79.
[7] Dai, X., Su, G., Tian, W., & Cheng, L. (2025). Research on the Operation, Maintenance, and Parameters of Expressway Mechanical and Electrical Equipment Based on Markov Prediction. Applied Sciences, 15(7), 3628.
[8] Lin, S., Luo, M., Niu, J., & Xu, H. (2022). Research on the Reliability of a Core Control Unit of Highway Electromechanical Equipment Based on Virtual Sensor Data. Sensors, 22(20), 7755.
[9] Liu, L., Tian, W., Dai, X., & Song, L. (2025). Research on Resource Consumption Standards for Highway Electromechanical Equipment Based on Monte Carlo Model. Sustainability, 17(10), 4640.
[10] Qiu, T., Huang, S., Cai, Z., & Sun, L. (2024, July). Enhancing expressway electromechanical device maintenance through a quantitative AHP-based decision-making model. In Third International Conference on Electronic Information Engineering and Data Processing (EIEDP 2024) (Vol. 13184, pp. 318-323). SPIE.
[11] Zhang, J., Liu, L., & Yang, H. (2023). An overview of intelligent construction and maintenance technology for highway subgrade engineering. Intelligent Transportation Infrastructure, 2, liad019.
[12] Huang, Y., Gao, J., Wang, L., Zhu, J., & Li, W. (2025). A Framework of Life-Cycle Infrastructure Digitalization for Highway Asset Management. Sustainability, 17(3), 907.
[13] He, X., Dong, G., & Yang, Z. (2018, August). A Novel Prediction Model of Expressway Electromechanical Equipment Life. In 2018 International Conference on Advanced Mechatronic Systems (ICAMechS) (pp. 246-250). IEEE.
[14] Yu, Q., Qin, Y., Liu, P., & Ren, G. (2018). A panel data model-based multi-factor predictive model of highway electromechanical equipment faults. IEEE Transactions on Intelligent Transportation Systems, 19(9), 3039-3045.
[15] Zhang, J., Zhu, L., Li, W., & Chen, H. (2024, December). Exploration of electromechanical digital development innovation and application enabled by business data of highway electromechanical maintenance platform. In Eighth International Conference on Traffic Engineering and Transportation System (ICTETS 2024) (Vol. 13421, pp. 913-919). SPIE.
[16] Lin, L., Jierui, Z., Jian, G., Yuqi, G., Hengyu, L., & Sheng, Y. (2024, August). Exploration of digital maintenance for highway electromechanical equipment based on digital twin. In Ninth International Conference on Electromechanical Control Technology and Transportation (ICECTT 2024) (Vol. 13251, pp. 533-538). SPIE.
[17] Duygu Selin Turan & Burak Ordin. (2025). The incremental SMOTE: A new approach based on the incremental k-means algorithm for solving imbalanced data set problem. Information Sciences,711,122103-122103.
[18] Dongliang Xia,Feifei Ning & Weina He. (2020). Research on Parallel Adaptive Canopy-K-Means Clustering Algorithm for Big Data Mining Based on Cloud Platform. Journal of Grid Computing,18(prepublish),1-11.
[19] Huyen Lynh Duong & Hai Nam Tran. (2025). Empowering Die Selection in V-Bending: Insights from Decision Tree Algorithms. Advances in Science and Technology,161,73-81.