

A K-means clustering method with feature learning for unbalanced vehicle fault diagnosis

Bo Wang, Guanwei Wang, Youwei Wang, Zhengzheng Lou,
Shizhe Hu and Yangdong Ye

School of Information Engineering, Zhengzhou University, Zhengzhou, China

Abstract

Purpose – Vehicle fault diagnosis is a key factor in ensuring the safe and efficient operation of the railway system. Due to the numerous vehicle categories and different fault mechanisms, there is an unbalanced fault category problem. Most of the current methods to solve this problem have complex algorithm structures, low efficiency and require prior knowledge. This study aims to propose a new method which has a simple structure and does not require any prior knowledge to achieve a fast diagnosis of unbalanced vehicle faults.

Design/methodology/approach – This study proposes a novel K-means with feature learning based on the feature learning K-means-improved cluster-centers selection (FKM-ICS) method, which includes the ICS and the FKM. Specifically, this study defines cluster centers approximation to select the initialized cluster centers in the ICS. This study uses improved term frequency-inverse document frequency to measure and adjust the feature word weights in each cluster, retaining the top τ feature words with the highest weight in each cluster and perform the clustering process again in the FKM. With the FKM-ICS method, clustering performance for unbalanced vehicle fault diagnosis can be significantly enhanced.

Findings – This study finds that the FKM-ICS can achieve a fast diagnosis of vehicle faults on the vehicle fault text (VFT) data set from a railway station in the 2017 (VFT) data set. The experimental results on VFT indicate the proposed method in this paper, outperforms several state-of-the-art methods.

Originality/value – This is the first effort to address the vehicle fault diagnostic problem and the proposed method performs effectively and efficiently. The ICS enables the FKM-ICS method to exclude the effect of outliers, solves the disadvantages of the fault text data contained a certain amount of noisy data, which effectively enhanced the method stability. The FKM enhances the distribution of feature words that discriminate between different fault categories and reduces the number of feature words to make the FKM-ICS method faster and better cluster for unbalanced vehicle fault diagnostic.

Keywords K-means, Cluster-centers selection method, Feature learning K-means, Unbalanced vehicle faults diagnosis

Paper type Research paper



1. Introduction

With the rapid increase in the number of railway transport vehicles, inertial fault frequently occurs, which dramatically reduces the overall efficiency of vehicle operating (Yang *et al.*, 2018). Therefore, to figure out what is the main factor leading to the vehicle faults is necessary, which can make the subsequent repair work well be done and formulate corresponding procedures in this mode. According to the research (Li *et al.*, 2017), the main causes of vehicle faults are human, machine, environmental and management factors. The four factors are the major causes of vehicle faults and even passenger safety problems. However, this is an unbalanced fault category data that the vast majority of vehicle faults are caused by the same factors, while only a small number of faults are caused by the remaining factors. In the age of intelligent railways, the use of text mining and other machine learning algorithms to achieve intelligent diagnostic of unbalanced vehicle fault text (VFT) data is the current urgent technical method (Li *et al.*, 2021).

1.1 Related work

Recently, vehicle fault diagnosis has attracted a lot of attention, and the topic is to perform the fault diagnostic and find fault-causing factors in machine learning methods. Zheng *et al.* (2007) implemented the text classification model based on rough set and fuzzy clustering theory to improve the adaptive performance and classification capability of text data classification system for vehicle faults without the unbalanced fault category data problem. Gao *et al.* (2020) proposed a K-fold cross-validation + stacking classification model to implement a single-level classification model for multi-level classification of signaling equipment faults and to solve the sample unbalance problem. However, it is not very efficient in practice. Hu and Meng (2020) applied the Hadoop platform, adopted an adaptive synthetic – gradient boosting decision tree decision tree algorithm based on the rolling stock fault diagnosis method to compensate for the defects brought about by the imbalance of data distribution to the fault diagnosis. However, the cost of data annotation is too high. Frequently it is hard to get strong supervised information such as full-truth labels. Zhao and Tang (2009) fused the fault diagnosis expert system of an unmanned vehicle based on back propagation neural networks. However, the unbalanced fault category data problem was not involved. Zhang *et al.* (2020) proposed the expert system for fault diagnosis of station signal control equipment, which was susceptible to outliers and less robust in practical application.

1.2 Contributions

The feature learning K-means-improved cluster-centers selection (FKM-ICS) method is proposed in this paper, to address the fault imbalance problem in vehicle fault diagnosis. The data unbalance in vehicle fault diagnosis means that the vast majority of vehicle fault is caused by the same factors (Yang *et al.*, 2018), while only a small number of faults are caused by the remaining factors. To solve the unbalance problem in vehicle fault diagnosis, the ICS is used to eliminate the influence of isolated points and improve the stability of the method, the FKM divides different fault categories by enhancing the discrimination among different feature word clusters. In ICS, owing to a random setting at the initial clustering centers of the traditional K-means algorithm, the cluster results are unstable and sensitive to outliers when faced with high-dimensional and sparse VFT data (Zhao and Xu, 2015). Therefore, we propose a method for initial cluster center selection based on density and distance based on the density peak method idea. In particular, the method combines the sample density (SD) with the distance between the selected cluster centers (CD) (Wei *et al.*, 2020) to define a cluster center approximation φ . The initial cluster centers are selected by φ to obtain the initial input parameters of the K-means algorithm. This process excludes the effect of outliers, solves the disadvantages of the traditional

K-means with poor noise immunity effectively and improves the stability of the algorithm. In the FKM, owing to the traditional term frequency-inverse document frequency (TF-IDF) method (Qin and Li, 2013) treats the text set as a whole and does not consider the unbalanced fault word distribution among clusters in the VFT data set, so some obscure words are often mistaken as fault keywords. Therefore, we propose to use the probability ratio of occurrence of feature words among clusters instead of the frequency ratio of occurrence of feature words among clusters to solve the above problem. The improved TF-IDF feature learning method fully considers the influence of the fault word distribution on the unbalanced fault categories, which is not only simpler to implement but also can achieve a better clustering performance.

The major contributions in this paper are as follows:

- We propose a novel FKM-ICS method to solve the unbalanced faults in vehicle fault diagnosis, which can well address the high-dimensional sparse unbalanced vehicle faults text data clustering.
- Using the FKM to generate different feature word representations for faults caused by different factors, which not only increases the distinction between fault categories but also greatly reduces the number of feature words, makes subsequent vehicle fault diagnosis simpler and more efficient.
- The ICS eliminates the effect of outliers by selecting the initial cluster centers and can make subsequent vehicle fault diagnosis more efficient and stable.

The remaining part of the paper is organized as follows: Section 2 revisits the traditional K-means algorithm. Section 3 introduces the FKM-ICS method. The experiments on the VFT data set are presented in Section 4. We conclude this paper in Section 5.

2. Revisit: the K-means algorithm

The basic idea of the traditional K-means algorithm (MacQueen, 1967) is: after inputting the number of clusters K , first select K samples randomly from the data set $X = \{x_1, x_2, \dots, x_n\}$ as the initial cluster centers, then calculate the distance from each sample to the K initial cluster center, classify the samples according to the principle of minimum distance to form K clusters, and then calculate the average value of the clusters to get the new cluster center. Keep repeating the above process until the center of the cluster stops changing or the algorithm ends after the iterations attain the set value.

The objective function is as follows:

$$\Phi = \sum_{k=1}^K \sum_{i=1}^n \sum_{j=1}^m |x_{ij} - u_{kj}|^2 \quad (1)$$

where K is the number of clusters, n is the number of samples, m is the feature dimension and $u_i = \{u_{i1}, u_{i2}, \dots, u_{im}\}$ is the i_{th} cluster center.

Given a set of data points and the objective function, the K-means algorithm partitions X into K clusters $\{C_i\}_{i=1}^K$ and K cluster centers set $U = \{u_1, u_2, \dots, u_K\}$

$$d_e(x_i, x_j) = \sqrt{\sum_{p=1}^m (x_{ip} - x_{jp})^2} \quad (2)$$

where $x_i = \{x_{i1}, x_{i2}, \dots, x_{im}\}$ and $x_j = \{x_{j1}, x_{j2}, \dots, x_{jm}\}$.

The cluster center u_i is the mean value of all samples belonging to the cluster. That is,

$$u_i = \frac{1}{|c_i|} \sum_{j \in c_i} x_j \quad (3)$$

where $|c_i|$ is the number of samples in i_{th} cluster.

3. The feature learning K-means based on the improved cluster-centers selection method

3.1 The proposed feature learning K-means-improved cluster-centers selection method

To improve the clustering accuracy of the traditional K-means algorithm, we propose the FKM-ICS method to achieve a fast diagnosis of vehicle faults, which includes the ICS method and the FKM method. Specifically, we first define a cluster centers approximation based on the SD and the distance between the samples and the selected cluster centers in the ICS. The improved TF-IDF is used to measure the importance of feature words for each cluster and adjust the weights of feature words in different clusters, keep the first τ feature words of each cluster with the highest weights and perform the cluster process again in the FKM. With the improved FKM-ICS, clustering performance for unbalanced vehicle fault diagnosis can be significantly enhanced. We find that the FKM-ICS can achieve a fast diagnosis of vehicle faults on the VFT data set. The experimental results on VFT show that the methods proposed in this paper, outperform several state-of-the-art methods. The results of VFT show that the method proposed in this paper outperforms some recent methods.

3.2 The ICS: an improved cluster-centers selection method

K-means algorithm is sensitive to the initial cluster centers and different initial cluster centers correspond to different cluster results (Wang and Shi, 2017). When faced with high-dimensional and sparse locomotive fault text data because the initial cluster center is a random set, the cluster result is unstable and easy to fall into a local optimal and the result is easily affected by noise points.

Based on the idea of density peak and the previous research experience, a clustering center selection method based on density and distance is proposed. Among them, this method combines the SD and the distance between the clusters to define a cluster center approximation φ . Use φ to select the initial cluster center and to obtain the initial input parameters of the K-means algorithm. This process eliminates the influence of isolated points, effectively solves the shortcomings of the classical K-means algorithm in terms of poor noise immunity and the tendency to fall into a local optimum and improves the stability of the algorithm.

Definition 1: The average sample distance $MeanDist(X)$ formula of data set X :

$$MeanDist(X) = \frac{2}{n(n-1)} \sum_i^n \sum_{j=i+1}^n d(x_i, x_j) \quad (4)$$

where $d(x_i, x_j) = \sum_{p=1}^m |x_{ip} - x_{jp}|$ is Manhattan distance between samples x_i and x_j , and the distances referred to later in the paper are Manhattan distances.

Definition 2: The density of sample points x_i :

$$\rho(i) = \sum_{j=1}^n f \left[d(x_i, x_j) - \text{MeanDist}(X) \right] \quad (5)$$

where the function $f(x) = \begin{cases} 1, & x < 0 \\ 0, & x \geq 0 \end{cases}$

A visual explanation of the defined SD is shown in [Figure 1](#).

From Definition 2, $\rho(i)$ means the number of sample points contained within a circle with sample point x_i as the center and $\text{MeanDist}(X)$ as the radius.

Definition 3: The distance $\bar{d}(i)$ between samples and clusters, which represents the average distance from sample x_i to center C of all identified clusters ([Huang et al., 2019](#)). The mathematical expression is as follows:

$$\bar{d}(i) = \frac{1}{|C|} \sum_{c_j \in C} d(x_i, c_j) \quad (6)$$

where $|C|$ is the number of selected clusters.

Definition 4: The cluster center approximation φ_i for x_i is calculated as follows:

$$\varphi_i = \rho_i \cdot \exp(\bar{d}(i)) \quad (7)$$

From Definition 4, it can be seen that the larger ρ_i (the more samples around x_i), the higher the cluster center approximation and the larger $\bar{d}(i)$ (the larger the inter-cluster distance),

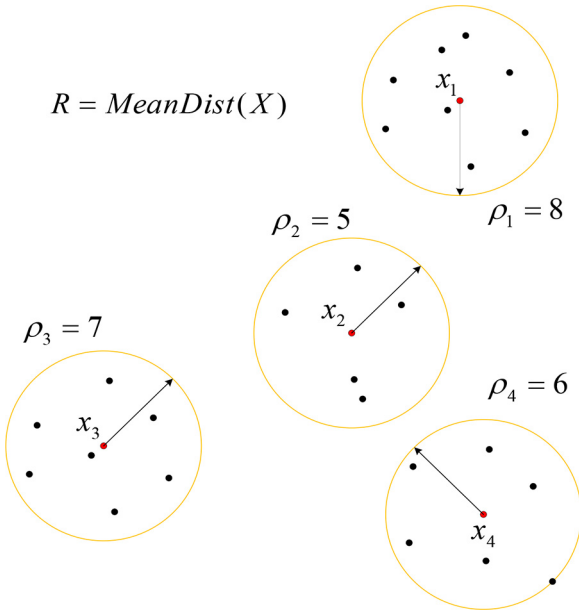


Figure 1.
Density of sample
points

the higher the cluster center approximation. We summarize the above initial cluster center selection process in Algorithm 1.

Algorithm 1 The ICS method.

Input: A data set $X = \{x_1, x_2, \dots, x_n\}$, number of clusters K .

Output: The initial cluster centers $U = \{u_1, u_2, \dots, u_K\}$.

Method:

1: compute the density of all samples in the data set;

2: the first initial cluster center u_1 is chosen as the sample with the highest density in X . then it is added to the set U of cluster centers, where $U = \{u_1\}$ and then all points in X that are less than $MeanDist(X)$ away from point c_1 are removed;

3: select x_i with the largest φ_i as the next initial cluster center, noting it as u_2 and add u_2 to the set U , where $U = \{u_1, u_2\}$, similar to step 2 and remove all samples in X that are less than $MeanDist(X)$ from u_2 ;

4: execute step 2, until K initial cluster centers are selected.

The first cluster center is chosen to be the largest dense point, which can effectively avoid the influence of isolated samples and noise. The later cluster center selection relies on the value of φ_i , from the definition of φ_i equation (8), we can see that the later selected cluster centers are larger density, a large distance between the clusters, which not only eliminates the possibility of randomly selecting the initial clustering center to select noise and isolated samples such as the traditional K-means algorithm but also consistent with the density peaks cluster algorithm: the cluster center of a data set is surrounded by data points of low local density that are located at larger distances from other points of high local density.

3.3 The FKM: a feature learning K-means method

The K-means cluster algorithm based on the improved TF-IDF feature learning is proposed in this paper, which introduces the importance of the sample features to the cluster into the objective function (Zhou et al., 2018) and fully considers the influence of the sample feature distribution characteristics on the unbalanced data, better cluster effect can be achieved and simple implementation.

3.3.1 The traditional term frequency-inverse document frequency method. The TF-IDF is a statistical method used to evaluate the importance of a feature word. It is based on the idea that the weight of a feature word in a text is the frequency it appears in the text and the number of texts that contain it. The *TF* is the occurrence frequency of feature word t in text x and *IDF* is related to the ratio of the number of all texts in the data set to the number of all texts containing the feature word t .

The common calculation is as follows:

$$TF = \frac{m}{M} \quad (8)$$

where m denotes the number of times of the feature word t in text x and M denotes the total number of the feature words t in text x .

$$DF = \log\left(\frac{N}{n+1}\right) \quad (9)$$

where N is the total number of texts in the data set, n is the total number of texts containing the feature word t , the reason to add 1, is to avoid the denominator is 0 (that is, all texts do not contain t).

$$TF - IDF = TF \times IDF \quad (10)$$

It is found that the traditional text feature weight representation method TF-IDF has some deficiencies: the TF-IDF treats the text set as a whole and does not take into account the uneven distribution of text features among clusters. In particular, the calculation of IDF has obvious defects in text cluster: in [equation \(10\)](#), $n_1 \gg n_2$, n_1 means the number of texts containing feature word t in cluster c_1 and n_2 means the number of texts containing feature word t in other clusters. When $n_1 \gg n_2$, In the case where N is certain, the value of IDF is small. However, the reality is that the feature word t occurs much more frequently in cluster c_1 than in other clusters and t should have a better ability to distinguish, but here it is the opposite of the desired result.

3.3.2 The improved term frequency-inverse document frequency method. According to the traditional TF-IDF function, the IDF of some obscure words is often higher, so these obscure words are often mistaken as text keywords. This is because IDF only considers the relationship between feature words and the number of texts in which they occur and ignores the distribution of feature words within a cluster and across clusters. (In other words, if a feature word occurs in large numbers in only a single text within a cluster and in few occurrences in most of the other texts within the cluster, then it is not excluded that these individual texts are special cases within that cluster, and therefore, such a feature word is not representative.). The above problem can be solved by replacing the ratio of the number of occurrences of feature words between clusters with the ratio of the probability of their occurrence to calculate the IDF.

Definition 5: According to the properties of The Law of Large Numbers, it is assumed that which words are used by the author of a certain type of text in writing that type of text is a random event, so $P(t_i)$ can be used to denote the probability of the occurrence of the feature word t in cluster c_i and $count(t_i)$ to denote the number of times t occurs in c_i . Finally, $P(t_i)$ can be formalized as follows:

$$P(t_i) = \frac{count(t_i)}{count(t_1) + count(t_2) + \dots + count(t_k)} \quad (11)$$

Definition 6: The feature word of a cluster of texts should represent well the feature information of the texts, denoted by $P(t_i)$ for the probability of occurrence of the word t in the cluster c_i , and $Q(t_i)$ denotes the sum of the probabilities that t occurs in clusters besides c_i . Then:

$$Q(t_i) = \sum_{j=1, j \neq i}^k P(t_j) \quad (12)$$

Then the IDF can be redefined as:

$$IDF = \log \left(\frac{P(t_i)}{P(t_i) + Q(t_i)} \right) \quad (13)$$

In $\frac{P(t_i)}{P(t_i) + Q(t_i)}$, When $P(t_i)$ is very larger, the absolute value of IDF but small, then take the opposite of it, according to the characteristics of the log function, the independent variable to be greater than 0, IDF to be positive and to avoid the denominator of 0 (that is, all texts do not contain the word), and finally, get the corrected IDF:

$$IDF = -\log \left(1 - \frac{P(t_i)}{P(t_i) + Q(t_i)} \right) = \log \left(1 + \frac{P(t_i)}{Q(t_i) + 0.001} \right) \quad (14)$$

Let $TF = P(t_i)$ denote the probability of feature word t occurring in a certain cluster of texts, consistent with the meaning denoted by [equation \(12\)](#). Finally, the improved TF-IDF formula is obtained as follows:

$$TF - IDF = P(t_i) \times \log \left(1 + \frac{P(t_i)}{Q(t_i) + 0.001} \right) \quad (15)$$

As compared to the traditional TF-IDF, the proposed method can not only reduce the weights of obscure words, reduce the influence of high-dimensional sparsity of sample features on the experimental results but also preserve the relevant information of the sample more completely. In addition, it also takes into account the differences in the representation of feature words among different clusters, so that the information of feature words can be extracted more effectively.

3.4 Final objective function of feature learning K-means method

$$\min \Gamma(X) = \sum_{i=1}^N \sum_{k=1}^K \frac{N}{|c_k|} d_w(x_i, u_k) \quad (16)$$

where $|c_k|$ is the number of samples in cluster c_k , $d_w(x_i, x_j)$ is the weighted Manhattan distance between x_i and x_j . The formula is as follows:

$$d_w(x_i, x_j) = \sum_{x_i \in c_k, p=1}^m w_{kp} |x_{ip} - x_{jp}| \quad (17)$$

where w_{kp} is the feature word weight for cluster c_k . The improved TF-IDF calculates the feature words weight in cluster c_k as follows:

$$w_{kp} = \frac{g(TF - IDF_{kp})}{\sum_{j=1}^m g(TF - IDF_{kp})} \tag{18}$$

where function $g(TF - IDF_{kp}) = \begin{cases} TF - IDF_{kp}, & rank(TF - IDF_{kp}) \geq \tau \\ 0 & , rank(TF - IDF_{kp}) < \tau \end{cases}$, remove feature words with a TF-IDF rank lower than τ . We summarize the FKM-ICS in Algorithm 2.

Algorithm 2 The FKM-ICS method.

Input: A data set $X = \{x_1, x_2, \dots, x_n\}$, number of clusters K and number of dimensions to be retained τ .

Output: Global partition $C = \{c_1, c_2, \dots, c_K\}$

Method:

- 1: initialize the cluster centers and obtain the K initialized cluster centers;
- 2: calculate the distance of all samples from these K cluster centers and assign them to the nearest cluster according to the minimum distance principle;
- 3: calculate the feature words weights for each cluster according to [equation \(19\)](#);
- 4: normalize the weights, perform K-means cluster based on the new weights, Execute step 5 if convergent, 3 otherwise;
- 5: output the global partition.

4. Experiments

To verify the validity of the FKM-ICS algorithm in vehicle fault analysis, this paper tests it on the VFT data set. The experimental environment configuration is shown in [Table 1](#).

4.1 Data set

The VFT data set is derived from a railway station’s 2017 VFT data, which is mainly formed by field personnel through natural language records and contains a total of 9,263

Table 1.
Experimental
environment

Environment	Configuration
Language	MATLAB
Tool	MATLAB R2017a
CPU	Intel Xeon E5-2699 v3
RAM	128GB

texts. The main information of those texts is that the occurrence of a fault, the description of a fault causes and the fault categories. Figure 2 shows the distribution of VFT data set fault categories.

From Figure 2, the faults are mainly in relationship person and relationship devices and the faults for relationship machines and not applicable are less, with an unbalanced ratio of 1:30, which is a typical data unbalance problem.

The VFT structured processing firstly needs to implement the partitioning of the fault text. Mainstream word segmentation techniques mainly include Chinese word segmentation based on dictionary matching, Chinese word segmentation based on word statistical model, Chinese word segmentation based on word annotation and Chinese word segmentation based on deep learning. In this paper, we use jieba.cut word segmentation tool to implement VFT segmentation using a general dictionary and custom domain dictionary (Su *et al.*, 2006). In addition, the vector representation of the text fault data is obtained by the bag of words model (Zhong *et al.*, 2018).

4.2 Baseline methods

In this paper, five popular cluster methods: K-means (KM), sequential information bottleneck [(sIB), Seldin *et al.*, 2001], spectral clustering [(SC), Ding *et al.*, 2014], normalized cuts [(Ncut), Shi and Malik, 2001] and automated variable weighting K-means [(WKM), Zhexue *et al.*, 2005] are used to compare with the FKM-ICS.

KM: KM is a popular partition-based clustering algorithm that divides n data objects into k clusters so that there is a high similarity within clusters and a low similarity between clusters. sIB: sIB is a method of data analysis based on information theory. It treats the extraction of data patterns as a data compression process, by compressing data objects into a “bottleneck variable” and maximizing to preserve the information in data, to obtain the inner patterns implicit in the data objects.

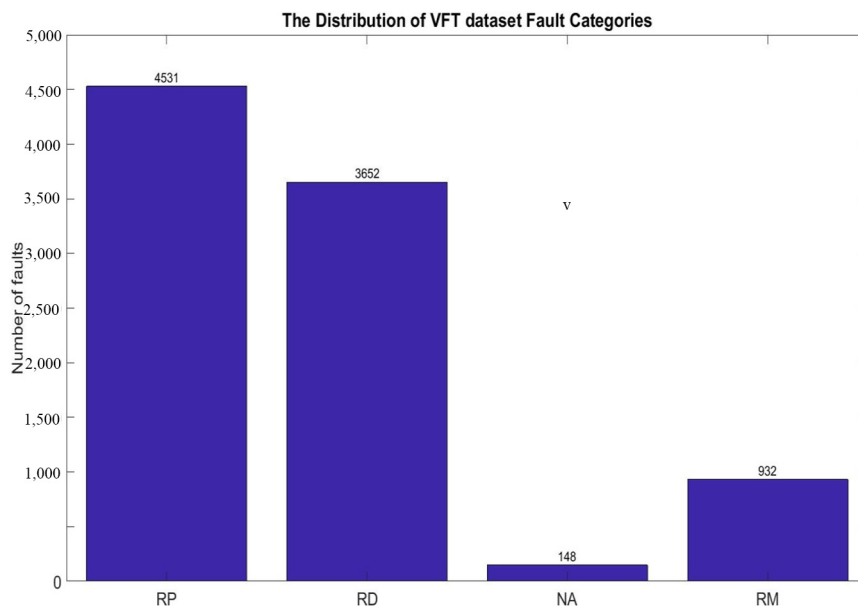


Figure 2.
Distribution of VFT
data set fault
categories

SC: SC converts the clustering problem into a multiple partitioning problem of an undirected weighted graph. Cutting the graph consisting of all data subjects so that the sum of edge weights between different subgraphs is lower and the sum of edge weights within subgraphs is higher, to achieve the clustering effect.

Ncut: Ncut is based on the SC of cutting a graph by considering the connection between clusters relative to the density of each cluster. Generally, the clustering performance of Ncut is better than that of SC.

WKM: WKM is a feature weighting method based on KM and proposed by Zhexue Huang, which reduces the effect of the noise dimension and improves the robustness of KM.

4.3 Evaluation indicators

In this paper, Acc and F1-score indicators are used to evaluate the validity of the cluster results. Acc is the most commonly used metric in cluster problems and it calculates the ratio of the number of correctly clustered predictions to the total number of predictions, with a maximum of 1 and a minimum of 0. That is,

$$Acc = \frac{n_correct}{n_total} \quad (19)$$

where $n_correct$ indicates the number of samples that were correctly clustered, n_total indicates the total number of samples.

F1-score is the weighted average of accuracy and recall, with a maximum of 1 and a minimum of 0. The formula is as follows:

$$F1 - score = \frac{2 * precision * recall}{precision + recall} \quad (20)$$

where *precision* is represented by the proportion of true frame in the cluster results, which reflects the model's ability to discriminate between negative samples. *recall* uses the proportion of the true samples to the detected frame, which reflects the model's ability to identify the positive samples. The F1-score is a combination of accuracy and recall and the higher F1-score, the greater robustness of the algorithm.

4.4 Experimental settings

For all baseline methods, we refer to the parameter settings of the original paper. The optimal cluster results are then searched in the parameter space. Note that we only report the best cluster results. For the proposed FKM-ICS algorithm, one parameter τ needs to be set, as shown in the hyper-parameter sensitivity test experiments. The following experiments were repeated 30 times and averaged.

4.5 Experimental results

4.5.1 Comparative experiment. Table 2 shows that the FKM-ICS algorithm achieves the best results among all the baseline methods on the VFT data set. The experimental results show the FKM-ICS algorithm can realize the vehicle fault diagnostic and has good diagnostic performance. The algorithm improves Acc by 4.58% and F1-score by 8.01% compared to sIB. The time cost was reduced by 12.59s compared to KM.

4.5.2 Ablation experiments. To verify the contribution of ICS and FKM to the algorithm, the ablation experiments were conducted on the VFT data set. In ICS+KM, the clustering

centers selected by ICS are used as the initial cluster center for the KM algorithm. In the FKM the initial clustering centers are K samples randomly selected from the data set. The results of the ablation experiments are shown in Table 3.

Table 3 shows the effects of ICS and FKM on the algorithm and the VFT data set. As shown, the ICS can effectively improve the stability of the algorithm; the FKM can significantly improve the clustering performance and the combination of both can achieve a better fault diagnosis performance.

4.5.3 Hyper-parameter sensitivity test experiments. To study the sensitivity of the proposed FKM-ICS to the parameters τ by varying its values. The parameter τ is the number of feature words retained in the clustering process. Normally, the larger number of retained feature words, the better performance of clustering results, and finally,

Method	Acc	F1-score	Time(s)
KM	$42.1 \pm (1.23)$	$42.23 \pm (1.35)$	$23.43 \pm (6.22)$
NC	$37.62 \pm (0.44)$	$39.75 \pm (0.62)$	$41.85 \pm (13.6)$
sIB	$44.36 \pm (1.34)$	$49.64 \pm (1.75)$	$56.23 \pm (5.24)$
SC	$35.33 \pm (0.13)$	$37.84 \pm (0.32)$	$67.25 \pm (3.4)$
WKM	$37.06 \pm (2.08)$	$37.07 \pm (2.55)$	$59.61 \pm (8.35)$
FKM-ICS	$48.94 \pm (0)$	$57.65 \pm (0)$	$10.82 \pm (1.63)$

Table 2.
Experimental results
of the methods on the
VFT data set

Method	Acc	F1-score	Time(s)
KM	$42.1 \pm (1.23)$	$42.23 \pm (1.35)$	$23.43 \pm (6.22)$
ICS+KM	$46.24 \pm (0)$	$47.16 \pm (0)$	$16.25 \pm (1.89)$
FKM	$47.5 \pm (2.18)$	$54.57 \pm (2.32)$	$11.39 \pm (1.31)$
FKM-ICS	$48.94 \pm (0)$	$57.65 \pm (0)$	$17.82 \pm (1.63)$

Table 3.
Results of ablation
experiments on the
VFT data set

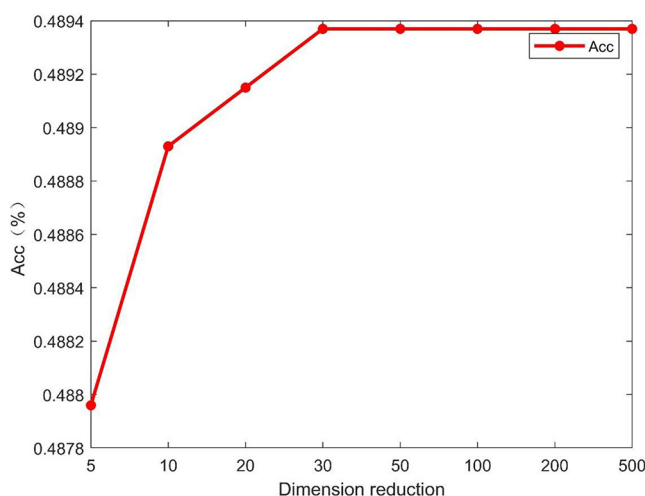


Figure 3.
Dimension sensitivity
test of FKM-ICS
algorithm on VFT
data set: Acc

convergence. We set τ from the grids {10, 50, 100, 200, 300, 500} to test the algorithm performance. The Acc and F1-score are shown in Figures 3 and 4.

Figure 3 shows the best performance of 48.94% obtained by the FKM-ICS on Acc when parameters τ is 30 and Figure 4 shows that when parameters τ is 50, the FKM-ICS obtains the best result on F1-score 57.65%. For the evaluation indicators in Figures 3 and 4, as parameters τ increase, the clustering performance of the FKM-ICS on the VFT data set gradually increases until convergence, proving the stability and convergence of the FKM-ICS method.

4.5.4 Convergence analysis. We empirically study the convergence property of our FKM-ICS algorithm. Figure 5 shows the objective value of the proposed method on the VFT data

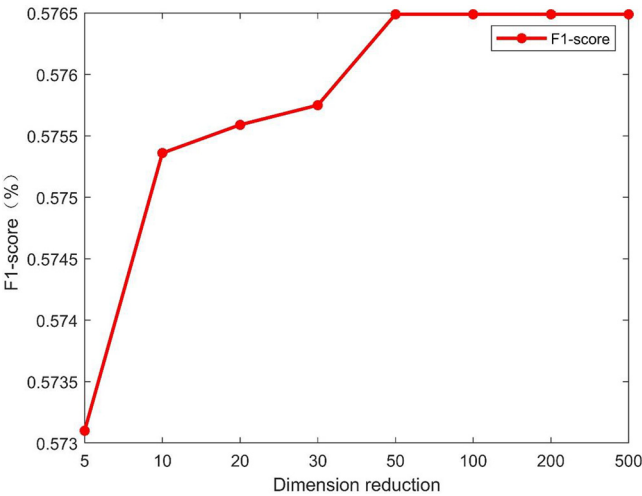


Figure 4.
Dimension sensitivity
test of FKM-ICS
algorithm on VFT
data set: F1-score

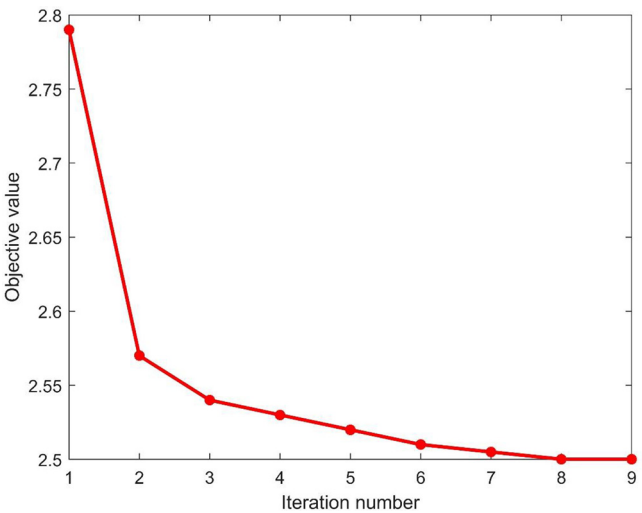


Figure 5.
Convergence of FKM-
ICS algorithm on
VFT data set

set. It is observed that the objective value increases monotonically with each iteration and finally, converges to an optimal solution within at most eight iterations.

5. Conclusion

In this paper, we propose the FKM-ICS algorithm to achieve the diagnosis of unbalanced VFT data in a more economical and faster way. The ICS enables the FKM-ICS method to exclude the effect of outliers, effectively solving the disadvantages of the fault text data contains a certain amount of noise data and improving the stability of the method. The FKM enhances the distribution of feature words that discriminate between different fault categories and reducing the number of feature words enables the FKM-ICS method faster and better cluster for unbalanced vehicle fault diagnostic. The experiments on the unbalanced VFT data of a railway station verify the Acc and validity of the proposed method and provide a new idea and solution for the intelligent classification of vehicle faults. In the future, we intend to expand the application of this method to other segments of the railway domain.

References

- Ding, S., Jia, S. and Shi, Z. (2014), "Spectral clustering algorithm based on adaptive nystrom sampling for big data analysis", *Journal of Software*, Vol. 25 No. 9, pp. 2037-2049.
- Gao, F., Li, P., Zhang, M., Wang, Z. and Zhao, J. (2020), "Research on multi-level classification of high-speed railway signal equipment fault based on text mining", *Computer Measurement and Control*, Vol. 28 No. 7, pp. 59-63.
- Hu, W. and Meng, J. (2020), "Research on fault diagnosis method of electric multiple units based on big data", *Industrial Control Computer*, Vol. 33 No. 6, pp. 31-32.
- Huang, X., Wang, C., Xiong, L. and Ceng, H. (2019), "A weighting K-meansclustering approach by integrating Intra-Cluster and Inter-Cluster distances", *Chinese Journal of Computers*, Vol. 42 No. 12, pp. 2836-2848.
- Li, Y., Liu, F., Wang, S. and Yin, J. (2021), "Multiscale symbolic Lempel-Ziv: an effective feature extraction approach for fault diagnosis of railway vehicle systems", *IEEE Transactions on Industrial Informatics*, Vol. 17 No. 1, pp. 199-208.
- Li, J., Wang, J., Xu, N. and Zhou, Z. (2017), "Analysis of safety risk factors for metro construction based on text mining method", *Tunnel Construction*, No. 2, pp. 160-166.
- MacQueen, J.B. (1967), "Some methods for classification and analysis of multivariate observations", *Mathematics*.
- Qin, S. and Li, F. (2013), "Improved TF-IDF method in text classification", *New Technology of Library and Information Service*, No. 10, pp. 27-30.
- Seldin, Y., Bejerano, G. and Tishby, N. (2001), "Unsupervised sequence segmentation by a mixture of switching variable memory Markov sources".
- Shi, J. and Malik, J. (2001), "Normalized cuts and image segmentation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22 No. 8, pp. 888-905.
- Su, J., Zhang, B. and Xu, X. (2006), "Advances in machine learning based text categorization", *Journal of Software*, Vol. 17 No. 9, pp. 1848-1859.
- Tao, X., Hao, S., Zhang, D. and Xu, P. (2013), "Overview of classification algorithms for unbalanced data", *Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition)*, Vol. 1, pp. 101-110.
- Wang, H. and Shi, Y. (2017), "K-meansclustering algorithm based on initial center optimization and feature weighted", *Computer Science*, Vol. 44 No. 11A, pp. 457-459.

- Wei, W., Tang, Z. and Liu, G. (2020), "PBK-meansalgorithm based on distance and density", *Computer Engineering*, Vol. 46 No. 9, pp. 68-75.
- Xie, H., Zhang, L., Lim, P., Yu, Y., Liu, Y., Liu, H. and Walters, J. (2019), "Improving K-meansclustering with enhanced firefly algorithms", *Applied Soft Computing Journal Computing*, Vol. 84.
- Yang, L., Li, P., Xue, R., Ma, X., Wu, Y. and Zhou, D. (2018), "Intelligent classification of faults of railway signal equipment based on imbalanced text data mining", *Journal of the China Railway Society*, Vol. 40 No. 2, pp. 59-66.
- Zhang, S., Zhao, H. and Wu, Y. (2020), "DCMN+: dual co-maching network for multi-choice reading coprehension", *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, New York, pp. 9563-9570.
- Zhao, Y. and Xu, T. (2015), "Text mining based fault diagnosis for vehicle on-board equipment of high speed railway signal system", *Journal of the China Railway Society*, Vol. 37 No. 8, pp. 53-59.
- Zhao, Z. and Tang, H. (2009), "Overview on Chinese segmentation algorithm", *Computer Knowledge and Technology*, Vol. 10 No. 1, pp. 2605-2607.
- Zheng, L., Wang, H. and Liu, L. (2007), "Research and implementation of text classification system based on theories of rough set and fuzzy clustering", *Journal of the China Railway Society*, Vol. 29 No. 1, pp. 45-49.
- Zhexue, H., Ng, M., Rong, H. and Li, Z. (2005), "Automated variable weighting in K-meanstype clustering", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 27 No. 5, pp. 657-668.
- Zhong, Z., Tang, T. and Wang, F. (2018), "Research on fault feature extraction and diagnosis of railway switches based on PLSA and SVM", *Journal of the China Railway Society*, Vol. 40 No. 7, pp. 80-87.
- Zhou, Y., Shu, X. and Wang, L. (2018), "A clustering algorithm based on cluster feature weighting fuzzy compactness and separation", *Information and Control*, Vol. 47 No. 6, pp. 722-729.

Further reading

- An, J., Yan, Z. and Zhai, J. (2015), "K-meansclustering algorithm based on distance threshold and weighted sample", *Microelectronics and Computer*, Vol. 32 No. 8, pp. 135-138.
- Li, J. and Dang, J. (2013), "Study on electric fault diagnosis model based on MAS", *Journal of the China Railway Society*, Vol. 35 No. 2, pp. 72-80.
- Lloyd, S. (1982), "Least squares quantization in PCM", *IEEE Transactions on Information Theory*, Vol. 28 No. 2, pp. 129-137.
- Singh, D. and Singh, B. (2019), "Investigating the impact of data normalization on classification performance", *Applied Soft Computing*, Vol. 97, p. 105524.
- Wang, Z., Lli, J. and Song, Y. (2020), "An improved K-meansalgorithm based on distance and weight", *Computer Engineering and Applications*.
- Yang, Z. and Wang, H. (2020), "Improved density peak clustering algorithm based on weighted K-nearest neighbor", *Application Research of Computers*, Vol. 37 No. 3, pp. 667-671.
- Yang, J., Ma, Y., Zhang, X., Li, S. and Zhang, Y. (2017), "An initialization method based on hybrid distance for K-meansalgorithm", *Neural Comput*, Vol. 29 No. 11.

Corresponding author

Yangdong Ye can be contacted at: yeyd@zzu.edu.cn.

For instructions on how to order reprints of this article, please visit our website:

www.emeraldgroupublishing.com/licensing/reprints.htm

Or contact us for further details: permissions@emeraldinsight.com