Journal of Network Intelligence ©2023 ISSN 2414-8105
Taiwan Ubiquitous Information Volume **8**, Number **4**, November **2023**

Text Mining and Sentiment Classification Based on Differential Evolutionary Clustering in Cloud Computing

Yi Wang*

Department of Information Engineering
Guangzhou Institute of Technology
Guangzhou 510075, P. R. China
70374@gzvtc.edu.cn

Yan Sun

Computational Institute of Aerodynamics of
China Aerodynamics Research and Development Center,
Mianyang 621000, P. R. China
y.sun@cardc.cn

Hui-Ru Cao

Department of Information Engineering
Guangzhou Institute of Technology
Guangzhou 510075, P. R. China
xiaocao0924@163.com

Xiang-Zhen Zhou

Faculty Information Science and Technology
National University of Malaysia
Selangor 43600, Malaysia
zhouxiangzhen@shengda.edu.cn

*Corresponding author: Yi Wang

Received April 3, 2023, revised June 25, 2023, accepted August 2, 2023.

Abstract. *Various office software based on cloud computing platforms have gradually become an important technical tool for scientific management, thus improving the quality and efficiency of enterprise management. These application systems generate a large amount of text data during operation and are stored in the relevant databases. Existing management systems can only perform simple queries on this data and are therefore unable to carry out intelligent analysis of human behavioural characteristics. The intelligent management system needs to perform data mining on human behavioural characteristics, so as to uncover the hidden features of the massive data. To address the above problems, an intelligent text mining and sentiment classification system based on K-means clustering analysis is constructed. First, the behavioural feature data of enterprise employees are pre-processed. Then, the K-means clustering algorithm is used to analyse the massive amount of historical data generated in the course of daily work in order to evaluate the employees. Secondly, an adaptive differential evolutionary algorithm based on dynamic subpopulation is proposed in order to overcome the shortcomings of the traditional K-means algorithm, and it is used to improve the K-means algorithm. Finally, the improved K-means algorithm was implemented in the Spark platform and the employee behavioural trait evaluation metrics were constructed. The experimental results show that the improved K-means algorithm effectively improves the clustering quality and convergence speed, thus effectively obtaining accurate and comprehensive employee sentiment assessment results.*

Keywords: Text mining; Sentiment classification; Cluster analysis; K-means algorithm; Adaptive differential evolutionary algorithm

1. Introduction.

As companies become larger and larger, the management of employee information becomes more and more difficult. As a result, the digitalisation of employee work management has become an inevitable trend for the future. Digital management systems are currently the dominant technology. These applications have greatly improved the efficiency of employee information management and saved manpower and time costs [1-4]. As a result, digital management systems have been widely adopted by various enterprises. However, most of these application systems only have record-keeping functions and query functions. After years of operation, these application systems have accumulated a huge amount of raw data. At present, managers are unable to effectively tap into the potential value of the raw data.

Employees generate a lot of information records in the course of their daily work, including basic employee information, employee achievement information, web access log information, canteen consumption information, etc.. All these informations are stored in the form of data in the database. These data imply a lot of important information about employees, such as their life trajectory, work situation and behavioural characteristics [5-8]. By analysing the potential connections between these massive amounts of raw data, the characteristics and patterns of employee behaviour can be uncovered, thus providing the necessary basis for decision-making in employee management. In addition, through the analysis of emotions in employee commentary information, early warnings can be given on abnormal employee behaviour or speech, thus enabling the monitoring of online public opinion. Therefore, how to mine and analyse these data, so as to provide the necessary data support for enterprise development, has become a more popular research direction.

Currently, more and more companies are using big data mining technology for their employees. They digitise employees' personal information by analysing data relating to their behavioural characteristics, such as the Employee Behaviour Analysis System at Lakeland University in Canada [9]. Companies assess the regularity of employees' learning and life based on their recent behavioural characteristics. If there are any abnormalities in the employee's life, the company will alert the employee via email. Managers can also use the system to get feedback on any abnormalities in their employees, which can support decision making in employee management. Marist College in New York State, USA, has partnered with Business Data Analytics to develop an open source academic analytics programme [10] that predicts which employees may not be able to complete their work tasks successfully. Employees are predicted by collecting their work habits, such as online reading time, messages, and hours worked. If problems are identified, teachers will help employees in time in order to improve productivity. Ithaca College uses the IBM Statistical Analysis System to collect data on employees' social networks [11]. Data mining methods are used to analyse employee online work data in order to develop sound management strategies, which has greatly improved work efficiency. The above analysis shows that data mining has a good practical application in digital management systems.

Data mining is also known as knowledge discovery [12, 13], which means "mining knowledge from data". Using machine learning, statistical learning and other techniques, data mining can sort out high-value models or data from massive amounts of data. In recent years, with the rapid development of computer and Internet technology, people's work and life have produced many forms of data, such as text, images, audio, video and so on. In addition, the amount of storage of this data is becoming increasingly large. How to extract hidden and valuable information from these massive data accurately and effectively has become a hot research problem in the field of computer science recently [14-18]. Data mining can be seen as a result of the natural evolution of information technology. Prediction and description are the two goals of data mining. Prediction means predicting the implicitly valuable fields and variables based on certain fields or variables of the information stored in the database. Description means describing data into comprehensible patterns.

The objects of data mining are data in various forms, including structured and unstructured data. Structured data refers to data records in database management systems. Unstructured data, on the other hand, refers to data stored in text documents, including web data on the web, etc. The diversity of structured forms has led to a variety of analysis algorithms for data mining. Whether the choice of algorithm is reasonable or not directly affects the result of data mining, so the choice of mining algorithm is the most important part of data mining research.

1.1. **Related Work.**

Clustering analysis has received increasing attention as one of the most important methods in Big Data mining. Clustering is the process of dividing data objects into subsets. Each subset is a cluster. Common applications of clustering techniques include outlier detection and customer group classification. Among them, K-means clustering algorithm is a common division-based algorithm in cluster analysis, which has the characteristics of simple implementation and scalability. Currently, K-means clustering algorithms are widely used in fields such as intelligence retrieval, machine learning, expert systems (relying on past rules of thumb) and pattern recognition.

Lin et al. [19] proposed a K-means clustering algorithm, which was applied in big data image retrieval and obtained high retrieval accuracy. However, the K-means clustering algorithm has a large dependence on the initial parameter settings, and its clustering results are less robust and easily fall into local optimum solutions.

Therefore, many researchers have attempted to improve the K-means clustering algorithm. Douzas et al. [20] proposed a unique density-based K-means optimization method, and the improved clustering results are more stable and have improved accuracy. Lohrmann and Luukka [21] established an initial center of mass using min-max similarity, thus improving the traditional K-means clustering algorithm. Avanija and Ramar [22] improved the typical K-means algorithm using the particle swarm algorithm and applied it in network intrusion detection. The method was able to solve the problem of random selection of initial clustering centres. In addition, with the "explosive" growth of data volume, the traditional model of K-means clustering algorithm can no longer adapt to the mining of large-scale data sets. Therefore, Cuomo et al. [23] proposed a GPU-assisted parallelised kernel K-means clustering algorithm, which significantly reduces the time consuming clustering algorithm. All these methods overcome the problems of the K-means clustering algorithm to a certain extent, but they still need to continue to be optimised and improved.

1.2. Motivation and contribution.

In order to solve these problems and thus improve the clustering stability and speed of the K-means clustering algorithm, this paper proposes an adaptive differential evolution algorithm based on dynamic subpopulations and introduces it into K-means clustering. Differential evolution algorithm is a new heuristic algorithm based on population intelligence, which can calculate the difference information among individuals in a population and complete the evolution of the population according to the fitness function. The differential evolution algorithm is a global optimisation algorithm with good robustness. The differential evolutionary algorithm is a good solution to the problem that K-means clustering algorithms tend to fall into the local optimum trap. Experimental results show that the proposed method improves the accuracy and stability of clustering results better than the traditional algorithm. In addition, the improved K-means algorithm was implemented in the cloud computing Spark platform in order to significantly improve the operational efficiency. The aim of this study is to analyse the massive amount of data in the digital management system through the improved K-means algorithm in order to build a text mining system based on the Spark platform. Companies can use this system to assess employees and thus enable early warning of abnormal employee behaviour.

The main innovations and contributions of this paper include:

(1) An adaptive differential evolution algorithm based on dynamic subpopulations is proposed for the initial cluster centre selection problem of the K-means cluster analysis algorithm. The whole population is divided into 2 sub-populations by means of individual fitness functions, and the 2 sub-populations are updated dynamically according to different variation strategies and parameters respectively, which improves the probability of obtaining the global optimum. K-means clustering is improved using the proposed adaptive differential evolution algorithm.

(2) An improved K-means clustering algorithm was applied to cluster and analyse information on employees' consumption levels, work performance, life patterns and social networks within the company. The improved K-means algorithm was implemented in the cloud computing Spark platform in order to significantly improve operational efficiency.

2. Pre-processing of employee behavioural characteristics data.

2.1. Data Extraction.

Data pre-processing is a prerequisite for data mining analysis. The data after data pre-processing is directly related to the analysis results. Therefore, data pre-processing is the most important step of data mining work. Generally speaking, data pre-processing takes up about 60% of the total work time, which shows the importance of data pre-processing work.

Data extraction is the process of transferring data from the data source to the target database, as shown in Figure 1. There are two main types of data extraction: full extraction and incremental extraction. Full extraction extracts the data from the data source to the target database in its original form. Incremental extraction will only extract data for insert, update or delete. As you can see, full extraction is relatively simple, however, incremental extraction is more widely used.

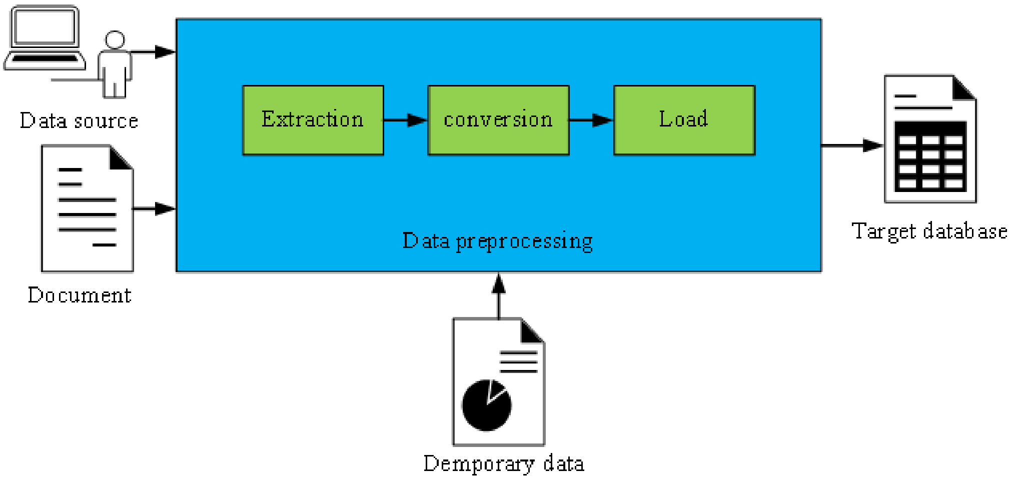

Figure 1. Data extraction process

In this study, Oracle ODI tool is used for data extraction. Oracle ODI is a data extraction/data conversion tool based on E-LT(Extract, Load & Transform) concept. Oracle ODI is mainly used for data cleaning. Data pre-processing methods mainly include data cleaning and data integration. Data cleaning is the process of finding and correcting errors in data files, including data consistency, invalid values and missing values.

2.2. **Data cleaning.** In the database of a digital management system, we need to first understand the basics of the data and see which data are unreasonable. Then, cleaning is performed by common methods of data governance, such as filling in missing data, eliminating anomalous data and smoothing noisy data [24]. In addition, inconsistent data needs to be corrected. Typically, the principle of data cleansing is shown in Figure 2. Generally speaking, missing values are cleaned up manually. For example, if the

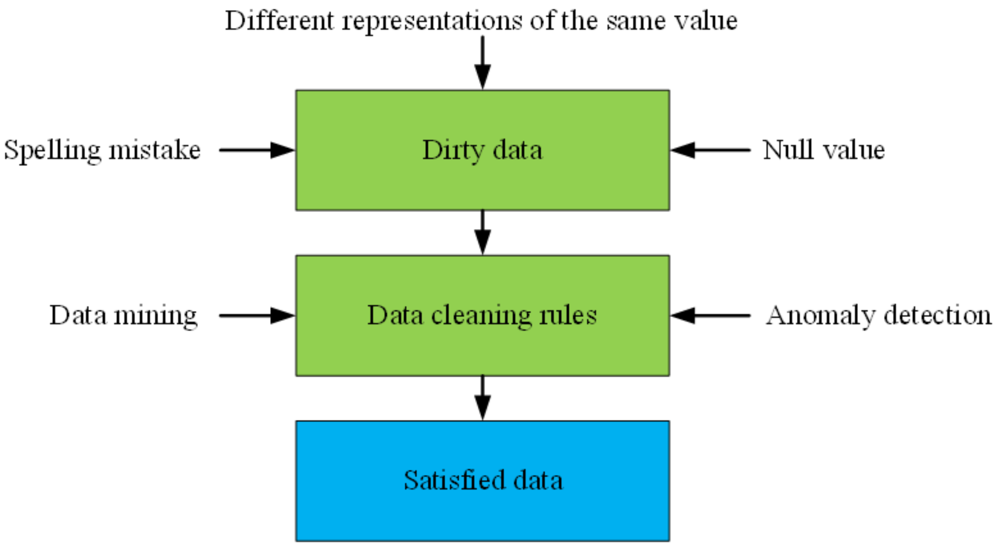

Figure 2. Principle of data cleansing

performance of an employee in the database is empty, then we can ignore the entire data record for that student or use the average value to fill in the missing values. As there are no uniform standards for the various sub-applications in the digital management system, it is possible that the data extracted is duplicated or incorrect. This noisy data can then affect the analysis results. For example, the employee performance data sheet has records of partially missing grades. Such missing values are usually the result of employees missing tests or cheating. We need to remove the IDs corresponding to the missing values or fill in the average of the grades. The employee appraisal performance is shown in Table 1. 201610003 and 201610006 have two IDs with missing attribute values and therefore need to be deleted.

Table 1. Employee appraisal performance

ID	Work Number	Sector	Work Projects	DKP Performance	Assessment Level
201610001	0304711	0304	Digital Circuits	85	2
201610002	0304712	0304	Digital Circuits	76	2
201610003	0304713	0304	Digital Circuits		2
201610004	0304714	0304	Digital Circuits	4	2
201610005	0304715	0304	Digital Circuits	78	2
201610006	0304716	0304	Digital Circuits		2
201610007	0304717	0304	Digital circuits	78	2
201610008	0304718	0304	Digital circuits	87	2
201610009	0304719	0304	Digital Circuits	89	2

2.3. **Data integration.** Data integration is an important step in forming a data warehouse, as the individual applications have their own databases. For example, the database for the finance department is sql server 2010, while the database for the personnel department is Oracle.

Therefore, we need to convert data from different databases into a uniform data format. After removing redundant and erroneous data, fields with the same data meaning are merged so that the data can be imported uniformly into a certain database. Through data integration, a regular data warehouse can be constructed, thus providing a good basis for subsequent data mining. For example, in this study, data from employee appraisal performance (Table 2) needs to be integrated with data from the personnel system (Table 3). First, the two tables are linked using ryid. Then, as all other fields are different, the two tables can be directly merged into one table.

Table 2. Employee appraisal performance record form

Serial number	Listings	Data type	Null
1	ryid	int	No
2	employee xh	int	No
3	employee cj	int	Yes
4	employee kc	varchar	No
5	employee xf	int	No
6	employee sfzh	int	No

3. An improved K-means clustering algorithm based on adaptive differential evolution.

Table 3. Personnel system records table

Serial number	Listings	Data type	Null
1	Local_ID	int	No
2	Card_ID	int	Yes
3	SWiID	int	No
4	PosID	tinyint	No
5	moneyXF	Uinyint	No
6	moneyRemain	int	No
7	DateXF	datetime	No
8	timeXF	datetime	No
9	Operator	smallint	No
10	protocalType	tinyint	Yes
11	ryid	int	No

3.1. **The basic idea of K-means algorithm.** As a distance-based divisional clustering algorithm, the K-means clustering algorithm has the advantages of a simple algorithm structure, high operational efficiency and a wide range of applicability. K-means clustering algorithm generally achieves optimization through the objective function [25]. The objective function is a process of calculating the sum of squares of errors.

$$E = \sum_{j=1}^{K} \sum_{x \in C_j} \|x - m_j\|^2 \tag{1}$$

where E is the clustering criterion function, K is the total number of clusters, C_j is the cluster in the cluster, x is a cluster target in the cluster and m_j is the average size of the cluster. In general, the smaller the value of E, the better the clustering effect. Conversely, the smaller the value of E, the worse the quality of the clustering.

The main disadvantages of the K-means clustering algorithm are the high dependence on the initial parameters and the tendency of the algorithm to produce local traps. To address this drawback, the K-means clustering algorithm is improved in this paper by using a differential evolutionary algorithm.

3.2. **Differential evolutionary algorithms.** Differential evolution algorithm is a new heuristic algorithm based on population intelligence, which can calculate the information of differences between individuals in a population and complete the evolution of the population according to the fitness function. The differential evolution algorithm is a global optimization algorithm with good robustness [26].

Let X denotes the initial population, N denotes the size of the population, $X_i(t)(t = 1, 2, ..., N)$ denotes the evolved individuals in the current population, and t denotes the number of iterations of the evolutionary process. The process of variation of individuals in a population operates as shown below:

$$V_i(t) = (v_{i1}(t), v_{i2}(t), ..., v_{iD}(t)) = X_{p1}(t) + F\left(X_{p2}(t) - X_{p3}(t)\right) \tag{2}$$

where F is the scaling weight, $F = 0.6$. D is the dimension of the individuals in the population, that is, the variant individuals $V_i(t)$ consist of D components.

The evolved individuals in the population $X_i(t)$ need to be crossed over with $V_i(t)$ to produce the competing individuals $U_i(t) = (u_{i1}(t), u_{i2}(t), ..., u_{iD}(t))$ (consisting of D components). The calculation of the j-th component of the competing individuals $U_i(t)$

is shown below:

$$u_{ij}(t) = \begin{cases} v_{ij}(t)\ randj(0,1) \le C_R\, or\, j \ne z \\ x_{ij}(t)\, randj(0,1) > C_R\, and\, j \ne z \end{cases} \tag{3}$$

where z is a random integer and $z \in \{1, 2, ..., D\}$. $C_R \in [0, 1]$ is the crossover probability. The crossover probability C_R usually takes a value between 0.3 and 0.9.

The fitness function compares competing and evolving individuals in order to renew the population on merit.

$$X_i(t+1) = \begin{cases} U_i(t)\ \ IF\ f(U_i(t)) \le f(X_i(t)) \\ X_i(t)\ \ IF\ f(U_i(t)) > f(X_i(t)) \end{cases} \tag{4}$$

3.3. **The proposed adaptive differential evolution algorithm.** Due to the use of a simple greedy selection strategy, during the evolutionary process of traditional differential evolution algorithms, the differences between individuals in the population become progressively smaller, leading to a gradual reduction in the overall diversity. This phenomenon can easily lead to the concentration of most individuals in the population around a local extremum.

When this happens, no matter how much mutation, crossover and selection is performed, no new individuals can be generated, resulting in the new population all being very similar to the old one. Therefore, this paper divides the whole population into 2 sub-populations by means of the fitness function of individuals, and dynamically updates each of the 2 sub-populations according to different variation strategies and parameters to improve the probability of obtaining the global optimum.

First, the population is divided according to the parameter δ . For example, when $\delta = 0.4, 40\%$ of the individuals in the original population form a subpopulation X' and the rest form a subpopulation X''. At the end of each iteration of evolution, the parameter δ needs to be updated.

$$\delta = \delta_{\min} + rand \cdot (\delta_{\max} - \delta_{\min}) \tag{5}$$

The 2 subpopulations were updated dynamically according to different mutation strategies and parameters. The variation strategy used for the individual $X_i(t)$ in the subpopulation X' is shown below:

$$V_i(t) = X_i(t) + F_i(t) \cdot (X_{p1}(t) - X_{p2}(t))\ + F_i(t) \cdot (X_{p3}(t) - X_{p4}(t)) \tag{6}$$

where $F_i(t)$ is the scaling weight corresponding to the individual $X_i(t)$ [27]. The variation strategy used for individuals $X_i(t)$ in the subpopulation X'' is shown below:

$$V_i(t) = X_i(t) + F_i(t) \cdot (X_{\delta best} - X_i(t)) + F_i(t) \cdot (X_{p5}(t) - X_{p6}(t)) \tag{7}$$

where $X_{\delta best}$ is a random selection of individuals from the subpopulation X'.

For individuals $X_i(t)$ in subpopulations X' and X'', if the offspring generated by $F_i(t)$ are able to enter the next generation population, then the scaling weights corresponding to the next generation individuals $X_i(t+1)$ can be generated according to equation (8). Otherwise, the scaling weights corresponding to the next generation of individuals $X_i(t{+}1)$ are generated according to equation (9).

$$F_i(t+1) = \frac{1}{2}\left[\frac{1}{2} \times \frac{fit_i - fit_b}{fit_w - fit_b} + \frac{1}{2} \times F_i(t)\right] \tag{8}$$

$$F_i(t+1) = \frac{1}{2}\left[\frac{1}{2} \times \frac{fit_i - fit_b}{fit_w - fit_b} + \frac{1}{2} \times mean_A(S_F)\right] \tag{9}$$

where fit_i is the fitness function of the current individual, fit_b is the best fitness function of the individuals in the current population, fit_w is the worst fitness function of the individuals in the current population, and $mean_A(S_F)$ is the arithmetic mean of all objects in the set S_F [28].

The dynamic update method according to the above scaling weights can fully take into account the survival of the offspring of different individuals. Dynamic updating of each of the 2 subpopulations according to different variation strategies and scaling parameters can effectively increase the probability of achieving the global optimum.

3.4. **Flow of improved K-means clustering algorithm.** The flow of the K-means clustering algorithm based on adaptive differential evolution is shown below:

Input: number of clusters K, crossover probability C_R, scaling weights F, initial data set X and population size N .

Output: Best clustering results.

Step 1: Coding of randomly selected cluster centres in the dataset using real number coding. When the number of evolutionary generations is $t = 0$, the sample i in the initial population is coded in the following way:

$$X_i(0) = (c_{i1}, c_{i2}, ..., c_{iK}) \quad i = 1, 2, ..., N \tag{10}$$

where c_{ij} shows the cluster centre j of the individual i and $j = 1, 2, ..., K$.

Step 2: Calculate the fitness function $f(X_i(t))$ for the individual $X_i(t)$. The fitness function is calculated as shown below:

$$f(c_i, c_i, ..., c_i) = \sum_{i=1}^{K} \sum_{x \in w_i} ||x - m_i|| \tag{11}$$

where m_i is the centre of the cluster w_i.

Step 3: Perform a mutation operation on an individual in the current population $X_i(t)$ to obtain a mutated individual $V_i(t) = (v_{i1}(t), v_{i2}(t), ..., v_{iK}(t))$.

$$v_{ij}(t) = x_{aj}(t) + F\left(x_{bj}(t) - x_{cj}(t)\right) \; j = 1, 2, ..., K \tag{12}$$

where a, b, c are a randomly generated integer and $a, b, c \in \{1, 2, ..., N\}$.

Step 4: Perform crossover operations on competing individuals $U_i(t)$.

Step 5: Use equations (6) and (7) to mutate individuals in the two sub-populations, respectively.

Step 6: When the maximum number of iterations is reached, output the optimal solution. Otherwise $t = t + 1$, skip to Step 2.

Step 7: Using the output of the adaptive differential evolution algorithm as the initial clustering centre of K-means, the similarity of all data objects to the centre is calculated using the Euclidean distance. Based on the nearest neighbour principle, the data objects are classified into the corresponding clusters.

Step 8: Output the clustering results.

Therefore, the flow of the K-mean clustering algorithm based on adaptive differential evolution is shown in Figure 3.

4. Improved K-means based text mining system under cloud computing.

4.1. **Overall system design.** In order to realize the clustering analysis of enterprise employee behavior characteristics data, an intelligent text sentiment classification system based on the improved K-means clustering algorithm was constructed, as shown in Figure 4. Full and incremental extraction of the large amount of historical data in the digital management system is carried out through the Oracle ODI tool and the data is preprocessed to obtain the desired data. The system has a user GUI interface to represent the data mining content. The data mining engine mines the data warehouse using an

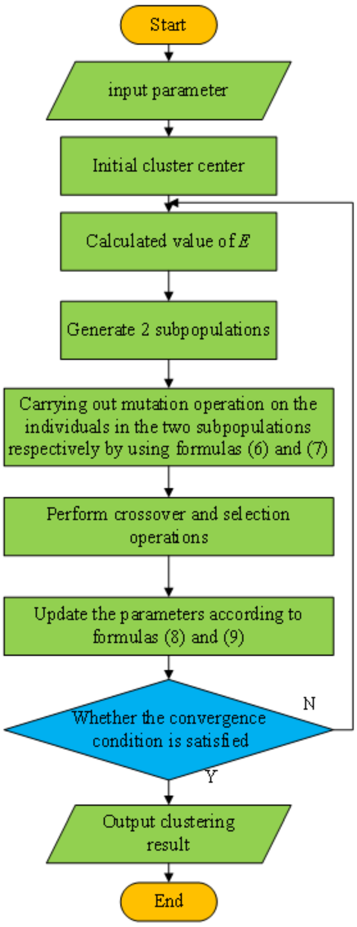
Start
input parameter
Initial cluster center
Calculated value of E
Generate 2 subpopulations
Carrying out mutation operation on the individuals in the two subpopulations respectively by using formulas (6) and (7)
Perform crossover and selection operations
Update the parameters according to formulas (8) and (9)
Whether the convergence condition is satisfied
N
Y
Output clustering result
End

Figure 3. Process of adaptive differential evolution based K-means clustering algorithm

improved K-means clustering algorithm. In addition, the models for data mining are optimised through a knowledge base in order to improve the accuracy of data mining.

By quantifying the life and learning behaviour of employees within the company, the behavioural characteristics of employees can be analysed in several dimensions (consumption levels, working hours and appraisal performance). As a result, the system assigns different weights to employee behavioural characteristics.

$$\sum_{i=1}^{m} w_i = 1 \tag{13}$$

where w_i indicates the weight of employee's behavioural characteristic i. The employee behavioural characteristics indicators are shown in Table 4.

4.2. **Implementation of the Platform Architecture.** Spark is an open source big data computing platform. spark has important applications in the field of big data parallel computing.

In this paper, Spark is used to achieve fast and efficient clustering mining. Spark uses a Master-Slave architecture based on in-memory parallelisation. The Master node is responsible for controlling the cluster, while the Distributed node is responsible for running the job tasks, and the Actuator starts the execution process of specific tasks. An intelligent employee sentiment assessment system was built on a Spark cluster of four machines. the hardware and software configuration of the four nodes (one master and

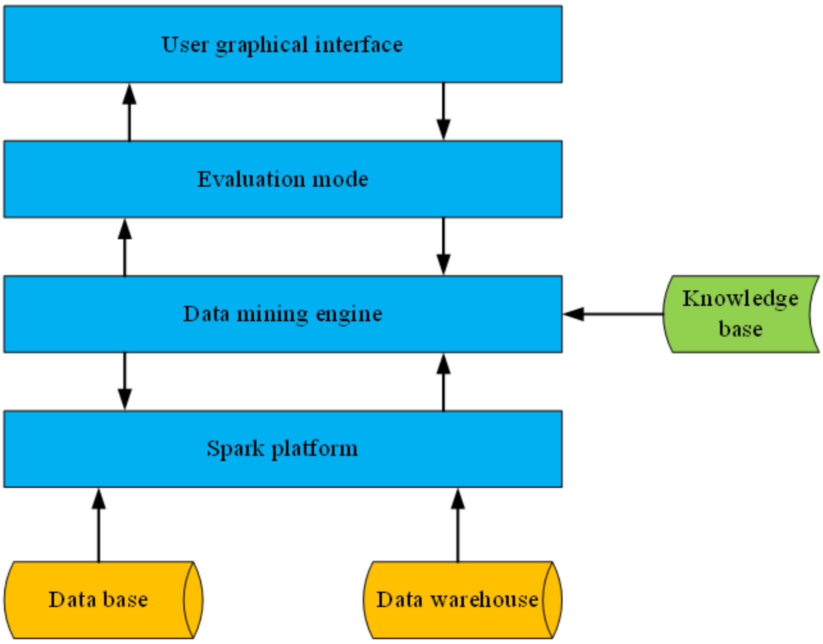

Figure 4. Framework of intelligent text sentiment classification system

Table 4. Indicators of employee behavioural characteristics

Employee behavioural characteristics	**Symbols**	**Weighting**	**Data type**	**Range of values**
Employee gender	C1	w1	Enumerated	M/F
Years of work	C2	w2	Enumerated	1/2/3/4
Affiliation Code	C3	w3	Numerical	1-13
Consumption	C4	w4	Enumerated	Low/Medium/High
Appraisal of performance	C5	w5	Numerical	0-100
Access control access	C6	w6	Numerical	0-500

three slave nodes) is shown in Table 5. All service nodes communicate with each other via 1000M fiber. All service nodes were installed with Spark version 1.2.1 and JDK version 1.7. The ip addresses of the four nodes were 214.102.61.2, 214.102.61.3, 214.102.61.4 and 214.102.61.5.

Table 5. Software and hardware configuration parameters of the node

Node type	**Hardware**	**Software environment**
master	Intel(R) Core(TM) i7-5600M CPU @2.60GHZ, 16GB RAM	Cent OS 6.7(x86_64), Jdk1.7.0_67, Hadoop-1.2.1, Spark-2.1.1
slave1	Intel (R) Core (TM) i7-5600M CPU @2.60GHZ, 8GB RAM	Jdk1.7.0_67, hadoop-2.7.3
slave2	Intel(R) Core(TM) i7-5600M CPU @2.60GHZ, 8GB RAM	Jdk1.7.0_67, hadoop-2.7.3
slave3	Intel(R) Core(TM) i7-5600M CPU @2.60GHZ, 8GB RAM	Jdk1.7.0_67, hadoop-2.7.3

5. Experimental results and analysis.

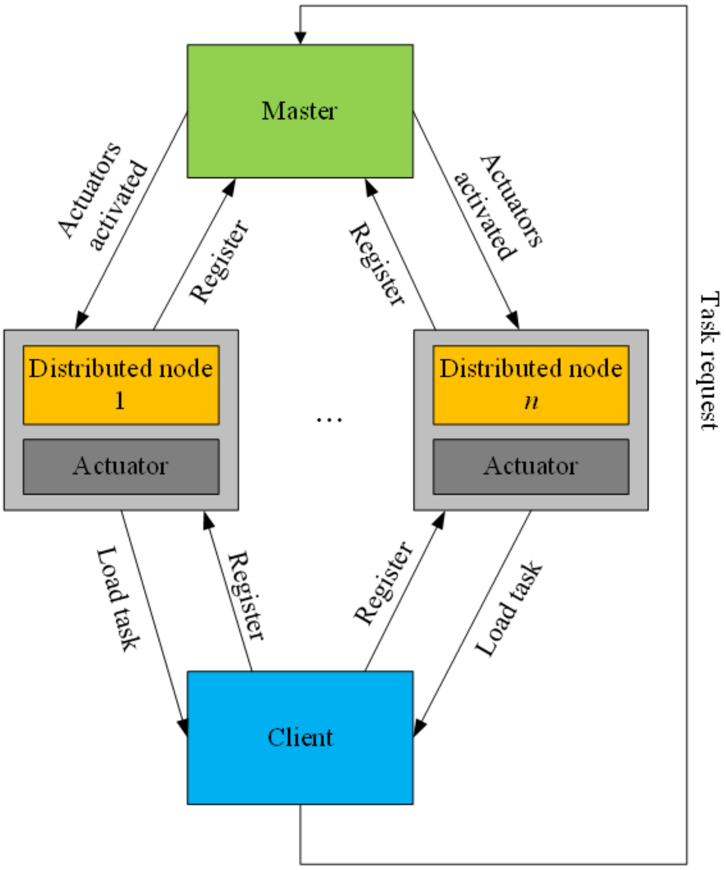

Figure 5. Spark task runflow

5.1. **Testing of standard machine learning datasets.** The experiment is divided into three parts. The first experiment analyses the performance of an improved K-means clustering algorithm using a standard machine learning dataset. The second experiment tested the feasibility of the designed system using real employee behavioural profile data from a digital management system. The third experiment uses employee social network data to validate the improved K-means clustering algorithm's ability to analyse online public opinion.

First, a standard dataset was used to test the improved K-means clustering algorithm and to compare it with the K-means clustering algorithm and the K-means clustering algorithm based on differential evolution [29]. The datasets used were three datasets from the UCI database, as shown in Table 6. The parameter settings associated with the algorithms are shown in Table 7.

Table 6. Experimental data set parameters

	IRIS	**Glass**	**Vowel**
Sample size	150	214	990
Number of sample attributes	4	9	10
Number of types	3	6	11

For the IRIS, Glass and Vowel datasets, the experimental results for the three clustering algorithms are shown in Table 8, Table 9 and Table 10 below respectively. Figure 6 shows the convergence characteristics for 20 dimensions.

Table 7. Experimental parameters

Parameters	**Numerical values**
Population size	40
$\delta_{\max}$	0.4
$\delta_{\min}$	0.2
Initial scaling weights F	0.6
Crossover probabilities C_R	0.5
Dimension D	20
The maximum number of iterations	20,000

Table 8. Experimental results for the IRIS dataset

	K-means	**Differential evolution based K-means**	**Improving K-means**
Minimum intra-class distance	87.7284	85.2816	82.2974
Maximum intra-class distance	157.9271	89.1268	81.0263
Average intra-class distance	106.3412	76.3452	68.3386
Average number of iterations of convergence	-	81	57

Table 9. Experimental results for the Glass dataset

	K-means	**Differential evolution based K-means**	**Improving K-means**
Minimum intra-class distance	5498.9892	5287.3343	4982.3461
Maximum intra-class distance	9316.3411	6577.3429	5734.4522
Average intra-class distance	7283.1187	5967.2645	5225.6436
Average number of iterations of convergence	-	1402	1164

It can be seen that the numerical results of the improved K-means clustering algorithm are the smallest compared to the other two algorithms. In addition, the improved K-means clustering algorithm has the smallest average intra-class distance and therefore, the clusters have a smaller fluctuation range and are more stable. In terms of average convergence generations, the improved K-means clustering algorithm has an improved convergence rate compared to the other two algorithms due to the use of a two-subpopulation mixing strategy. The convergence characteristic curves also indicate the convergence advantage of the improved K-means clustering algorithm with good global optimisation seeking ability.

The above experimental results validate the feasibility and efficiency of the improved K-means clustering algorithm. Compared with the other two algorithms, the improved K-means clustering algorithm is able to obtain the global optimum at a faster rate and has better robustness.

Table 10. Experimental results for the Vowel dataset

	K-means	**Differential evolution based K-means**	**Improving K-means**
Minimum intra-class distance	5498.9892	5287.3343	4982.3461
Maximum intra-class distance	9316.3411	6577.3429	5734.4522
Average intra-class distance	7283.1187	5967.2645	5225.6436
Average number of iterations of convergence	-	1402	1164

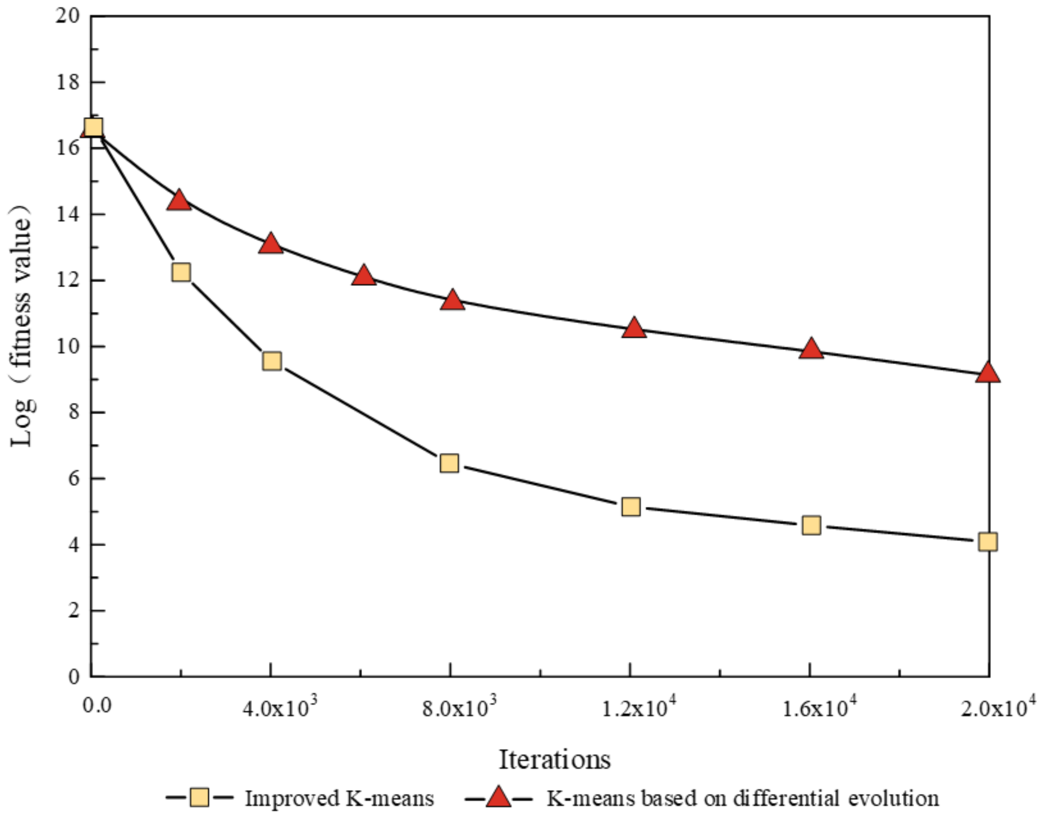

Figure 6. Comparison of convergence characteristics

5.2. **Clustering results of employee behavioural characteristics data.** Then, data such as consumption information and employee performance appraisal records were selected as the experimental data set to test the designed system.

The amount spent is divided into four categories of attributes. ”$Consumptionamount < \$10$” corresponds to ”1”, ”$\$10 \leq consumptionamount < \$20$” corresponds to”2”, ”$\$20 \leq \$20 < \$30$” corresponds to ”3” and ”$\$30 \geq$ ” corresponds to ”4”. Similarly, the employee performance appraisal is divided into four categories of attributes. ”Excellent”, ”Good”, ”Pass” and ”Fail” correspond to ”1”, ”2”, ”3” and ”4” respectively. The attribute weights for consumption are shown in Figure 7. The attribute weights for employee performance appraisal records are shown in Figure 8.

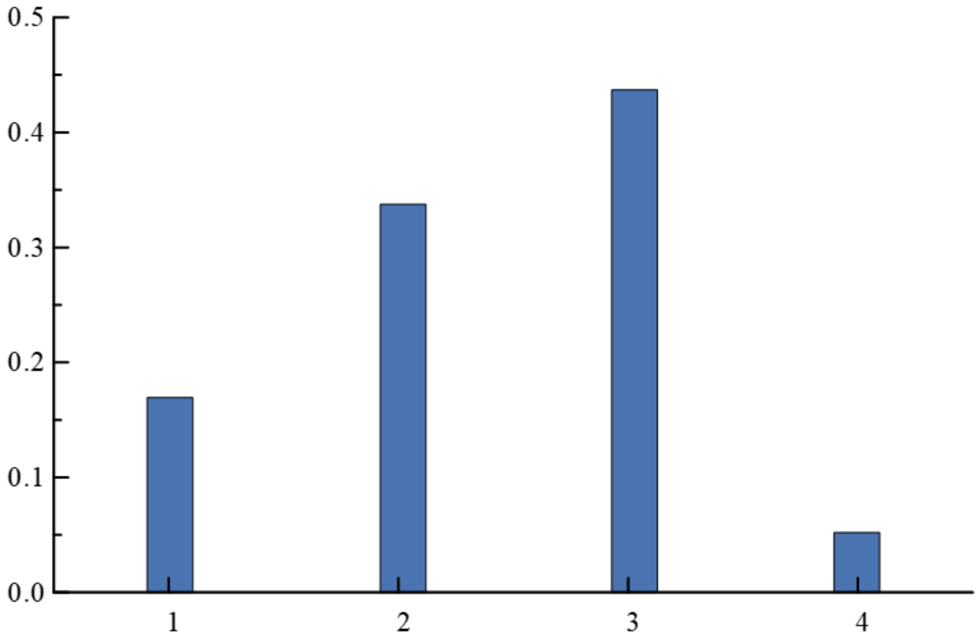

Figure 7. Attribute weights for consumption

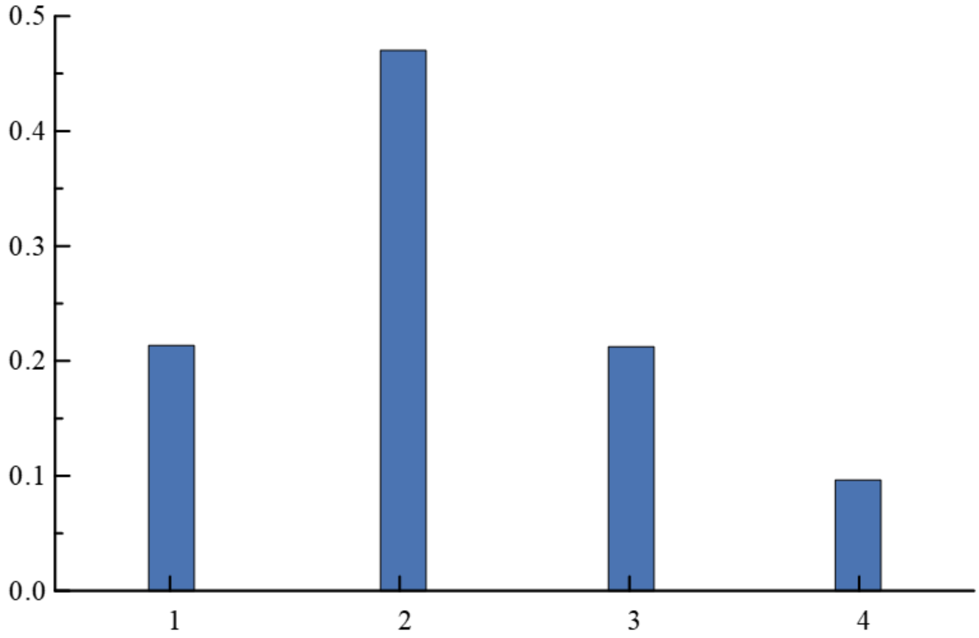

Figure 8. Attribute weights for employee performance appraisals

The card consumption dataset and the employee performance appraisal dataset were clustered using modified K-means and original K-means respectively, yielding the experimental results shown in Tables 11 and 12. The experimental results were evaluated using clustering accuracy.

It can be seen that for the consumption dataset, the improved K-means increases the clustering accuracy from 88.75% to 97.25%. For the performance appraisal dataset, the improved K-means increased the clustering accuracy from 89.00% to 97.50%. The clustering results in Experiment 2 were almost identical to those in Experiment 1. The test

Table 11. Clustering results for the consumption dataset

Category	Number of instances	Original K-means	Improving K-means
1	100	87	99
2	100	90	95
3	100	88	97
4	100	90	98
Accuracy		88.75%	97.25%

Table 12. Clustering results for the employee performance appraisal dataset

Category	Number of instances	Original K-means	Improving K-means
1	50	45	50
2	50	48	49
3	50	39	50
4	50	46	46
Accuracy		89.00%	97.50%

results show that the proposed method is able to assess the behavioural characteristics of employees more accurately than the traditional algorithm.

5.3. **Sentiment classification.** Finally, the Python language was used to crawl employee social data on corporate forums and microblogs and extract the textual content from the web pages.

The time frame for the employee social data was from January 1, 2021 to January 1, 2022. TF-IDF was used to process all the collected texts to determine the weight of each feature word. The improved K-means clustering algorithm was used to cluster the word bank after the word separation, thus classifying the text information into three sentiment categories: positive, negative and neutral. The distribution of opinion sentiment is shown in Figure 9. It can be seen that there are 13 positive messages, 29 negative messages and

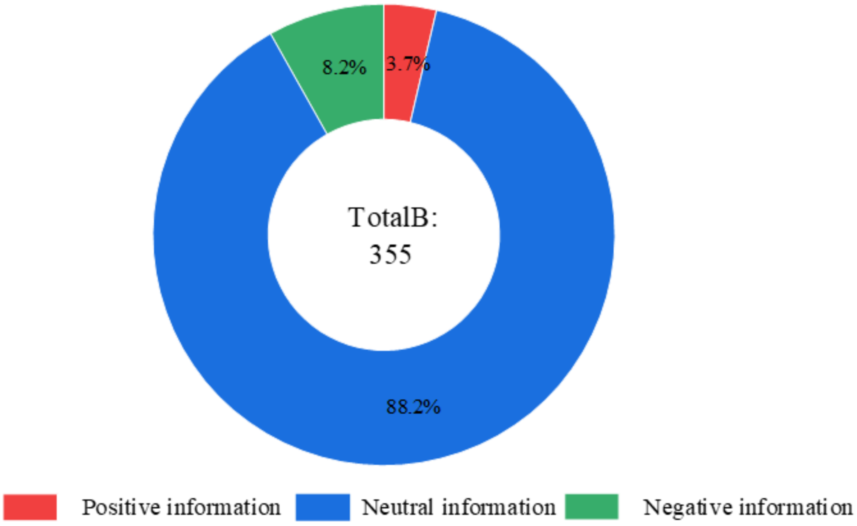

Figure 9. Distribution of public sentiment

313 neutral messages. By analysing the employee social network data, it is possible to keep abreast of the employees' thought dynamics. The clustering results for topic discovery are shown in Table 13. It can be seen that by analysing the emotions in employee commentary information, managers can provide early warning of abnormal employee behaviour or speech, thus enabling the monitoring of online public opinion.

Table 13. Clustering results for topic discovery

Subject	**Quantity**
Unemployment problem	45
Brokenhearted, emotional problems	34
Job promotion	23
Diet in the canteen	34
Interpersonal relationship	46
Other	272

6. **Conclusion.** In this paper, a text mining and sentiment classification based on an improved K-means clustering algorithm was constructed so as to cluster and analyse the behavioural characteristics data of enterprise employees. Enterprises can use this system to classify and evaluate their employees, thus enabling early warning of abnormal employee behaviour. The K-means clustering algorithm is used to analyse the massive amount of historical data on employees in order to evaluate them. In order to overcome the shortcomings of the traditional K-means algorithm, an adaptive differential evolution algorithm based on dynamic subpopulation is proposed and used as an improvement to the K-means algorithm. The improved K-means algorithm was implemented in the Spark platform and the employee behavioural trait assessment metrics were constructed. The experimental results show that the improved K-means clustering algorithm is able to obtain the global optimum at a faster rate and has better robustness than the other two algorithms. Compared with the traditional K-means algorithm, the improved K-means clustering algorithm can more accurately assess the behavioural characteristics of employees, thus enabling early warning of abnormal employee behaviour.

Acknowledgement. This work was supported by the Guangdong Education Science Planning Project (No.2021GXJK684).

REFERENCES

[1] J. H. Marler, X. Liang, and J. H. Dulebohn, "Training and Effective Employee Information Technology Use," *Journal of Management*, vol. 32, no. 5, pp. 721-743, 2006.

[2] M. Siponen and A. Vance, "Neutralization: New Insights into the Problem of Employee Information Systems Security Policy Violations," *MIS Quarterly*, vol. 34, no. 3, pp. 487-502, 2010.

[3] P. Kent and T. Zunker, "Attaining legitimacy by employee information in annual reports," *Accounting, Auditing & Accountability Journal*, vol. 26, no. 7, pp. 1072-1106, 2013.

[4] N. Shah, "A study of the relationship between organisational justice and employee readiness for change," *Journal of Enterprise Information Management*, vol. 24, no. 3, pp. 224-236, 2011.

[5] G. Solomon and I. Brown, "The influence of organisational culture and information security culture on employee compliance behaviour," *Journal of Enterprise Information Management*, vol. 14, no. 4, pp. 89-102, 2020.

[6] W. Yaokumah, D. O. Walker, and P. Kumah, "SETA and Security Behavior," *Journal of Global Information Management*, vol. 27, no. 2, pp. 102-121, 2019.

[7] S. Chaudhry, "Managing Employee Attitude for a Successful Information System Implementation: a Change Management Perspective," *Journal of International Technology and Information Management*, vol. 27, no. 1, pp. 57-90, 2018.

[8] J. D'Arcy and P.-L. Teh, "Predicting employee information security policy compliance on a daily basis: The interplay of security-related stress, emotions, and neutralization," *Information & Management*, vol. 56, no. 7, 103151, 2019.

[9] T. Papadopoulos, T. Stamati, and P. Nopparuch, "Exploring the determinants of knowledge sharing via employee weblogs," *International Journal of Information Management*, vol. 33, no. 1, pp. 133-146, 2013.

[10] C.-M. Chen, Z. Z. Zhang, J. M.-T. Wu, and K. Lakshmanna, "High Utility Periodic Frequent Pattern Mining in Multiple Sequences", *Computer Modeling in Engineering & Sciences*, vol. 137, no. 1, pp. 733-759, 2023.

[11] B. L. Chen, W.-S. Gan, Q. Lin, S.Q. Huang, and C.-M. Chen, "OHUQI: Mining on-shelf high-utility quantitative itemsets", *The Journal of Supercomputing*, vol. 78, pp. 8321-8345, 2022.

[12] H. Hong, P. Tsangaratos, I. Ilia, J. Liu, A-Xing. Zhu, and W. Chen, "Application of fuzzy weight of evidence and data mining techniques in construction of flood susceptibility map of Poyang County, China," *Science of The Total Environment*, vol. 625, pp. 575-588, 2018.

[13] C.-M. Chen, L. Chen, W.S. Gan, L. Qiu, and W.-P. Ding, "Discovering high utility-occupancy patterns from uncertain data", *Information Sciences*, vol. 546, pp. 1208-1229, 2021.

[14] J. Lee, N. Ohba, and R. Asahi, "Discovery of zirconium dioxides for the design of better oxygen-ion conductors using efficient algorithms beyond data mining," *RSC Advances*, vol. 8, no. 45, pp. 25534-25545, 2018.

[15] A. Mohamed, J. Molendijk, and M. M. Hill, "Lipidr: A Software Tool for Data Mining and Analysis of Lipidomics Datasets," *Journal of Proteome Research*, vol. 19, no. 7, pp. 2890-2897, 2020.

[16] J. A. M. Demattê, C. T. Fongaro, R. Rizzo, and J. L. Safanelli, "Geospatial Soil Sensing System (GEOS3): A powerful data mining procedure to retrieve soil spectral reflectance from satellite images," *Remote Sensing of Environment*, vol. 212, pp. 161-175, 2018.

[17] Y. Djenouri, A. Belhadi, and R. Belkebir, "Bees swarm optimization guided by data mining techniques for document information retrieval," *Expert Systems with Applications*, vol. 94, pp. 126-136, 2018.

[18] A. Agrawal and A. Choudhary, "An online tool for predicting fatigue strength of steel alloys based on ensemble data mining," *International Journal of Fatigue*, vol. 113, pp. 389-400, 2018.

[19] C.-H. Lin, C.-C. Chen, H.-L. Lee, and J.-R. Liao, "Fast K-means algorithm based on a level histogram for image retrieval," *Expert Systems with Applications*, vol. 41, no. 7, pp. 3276-3283, 2014.

[20] G. Douzas, F. Bacao, and F. Last, "Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE," *Information Sciences*, vol. 465, pp. 1-20, 2018.

[21] C. Lohrmann and P. Luukka, "A novel similarity classifier with multiple ideal vectors based on k-means clustering," *Decision Support Systems*, vol. 111, pp. 27-37, 2018.

[22] J. Avanija and D. K. Ramar, "A Hybrid Approach Using Pso and K-Means for Semantic Clustering of Web Documents," *Journal of Web Engineering*, vol. 12, no. 3-4, pp. 249-264, 2013.

[23] S. Cuomo, V. De Angelis, G. Farina, L. Marcellino, and G. Toraldo, "A GPU-accelerated parallel K-means algorithm," *Computers & Electrical Engineering*, vol. 75, pp. 262-274, 2019.

[24] G. Chen, Y. Li, K. Zhang, X. Xue, and J. Wang, "Efficient hierarchical surrogate-assisted differential evolution for high-dimensional expensive optimization," *Information Sciences*, vol. 542, pp. 228-246, 2021.

[25] X. Yu, C. Li, and J. Zhou, "A constrained differential evolution algorithm to solve UAV path planning in disaster scenarios," *Knowledge-Based Systems*, vol. 204, p. 106209, 2020.

[26] T. Wang, H. Ke, X. Zheng, and K. Wang, "Big Data Cleaning Based on Mobile Edge Computing in Industrial Sensor-Cloud," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 2, pp. 1321-1329, 2020.

[27] M. Ahmed, R. Seraj, and S. M. S. Islam, "The k-means Algorithm: a Comprehensive Survey and Performance Evaluation," *Electronics*, vol. 9, no. 8, 1295, 2020.

[28] A. M. Ikotun, M. S. Almutari, and A. E. Ezugwu, "K-Means-Based Nature-Inspired Metaheuristic Algorithms for Automatic Data Clustering Problems: Recent Advances and Future Directions," *Applied Sciences*, vol. 11, no. 23, 11246, 2021.

[29] P. Mansueto and F. Schoen, "Memetic differential evolution methods for clustering problems," *Pattern Recognition*, vol. 114, 107849, 2021.