Journal of Network Intelligence ©2025 ISSN 2414-8105
Taiwan Ubiquitous Information Volume **10**, Number **4**, November **2025**

Hierarchical Clustering and Key Rule Based Fuzzy Random Mining Algorithm for Big Data

Min-Juan Liu*, Xi-Zhi Zhang

College of Information Engineering
Zhengzhou Shengda University
Zhengzhou 451191, P. R. China
25741190@qq.com, 1908879334@qq.com

He-Ling Cao

College of Information Science and Engineering
Henan University of Technology
Zhengzhou 450001, P. R. China
caohl@haut.edu.cn

Kuang Jing

Bulacan State University
Malolos 3000, Philippines
704161029@qq.com

*Corresponding author: Min-Juan Liu

Received December 11, 2024, revised May 18, 2025, accepted August 22, 2025.

Abstract. *Big data mining is the process of extracting valuable information from a large amount of random and fuzzy data, and due to the characteristics of massive big data with multi-dimensionality, sparsity and dynamics, it is difficult to accurately obtain its distributional characteristics, and it is difficult to realize random mining directly. For this purpose, a fuzzy stochastic mining algorithm for big data based on hierarchical clustering and key rules is proposed. Firstly, for the issue that the cohesive hierarchical clustering algorithm (HAC) is easily affected by boundary points, the HAC is optimized by using the idea of redistribution of boundary points (EHAC), and hierarchical clustering is carried out within and between hierarchical levels, respectively, to identify the boundary points and redistribute the boundary points, so as to improve the accuracy of the clustering. Then, based on the association rules between the data, the big data are fused, and the EHAC algorithm is used to obtain the linguistic association characteristics of the fused big data. Finally, combining the association rule and feature weighting optimization Naive Bayes, adaptive uniform traversal of the associated features, to find the optimal solution of the mining objective function, to achieve the optimization of big data mining. The experimental outcome indicates that the clustering effect of the proposed algorithm is better, and the mining accuracy can reach more than 90%, which can effectively realize the fuzzy random mining of big data.*

Keywords: big data mining; hierarchical clustering; key rule; feature weighting; Naive Bayes

1. **Introduction.** Data mining is an important method to analyze and process data, which can mine valuable information from data and provide decision support for decision makers [1]. Recently, as the information technology growing, the amount of data generated by people every day is growing and accumulating at an unprecedented rate. The huge

amount of data and complex data types pose new challenges to the existing data mining algorithms [2]. Along with the concept of big data, traditional data mining algorithms are limited by factors such as memory, efficiency, scalability, and so on, and cannot effectively deal with big data, and it is difficult to meet the demand of extracting valuable information efficiently and accurately [3, 4]. Especially when it comes to the processing of fuzzy and random data, the existing algorithms often appear to be incompetent. Therefore, how to process big data more efficiently and mine the required information from it has become a hot topic in current research on big data.

1.1. **Related work.** Traditional algorithms mainly use time-frequency analysis [5], cluster analysis [6], beam analysis [7], etc. Bocca and Rodrigues [8] offered a data mining algorithm based on autocorrelation feature matching to form autocorrelation beams for data in massive cascading databases, but the disadvantage of this algorithm is that the computational overhead is too large, and real-time performance of massive data mining is not good. Riaz et al. [9] used k-means clustering and text detection for nonlinear feature space reconstruction of the data, on the basis of which semantic association feature retrieval and feature filtering and matching are performed, but this algorithm reduces the accuracy of data mining in the case of high interference. Subramaniam et al. [10] used fuzzy C-mean clustering method to achieve data mining, when the data spatial features of the similarity difference is small, the accuracy of data mining is not high. Ianni et al. [11] first to the minimum quadratic distance as the criterion of data clustering processing, through parallelization clubs to complete the fuzzy stochastic mining of big data, but the computational consumption is large. Methods such as k-means and fuzzy C-mean clustering require determining the optimal cluster centers as well as defining the number of classes, resulting in high computational overhead. Hierarchical Clustering Algorithm (HCA) based on unsupervised learning has attracted the attention of scholars due to its simple logical principles as well as accurate clustering results. Dogan and Birant [12] proposed a big data digging approach on the ground of single link cohesive HCA but the effectiveness metrics measurements are insufficient. Pandey and Shukla [13] proposed a hierarchical clustering approach relied on the Minimum Spanning Tree (MST) and used it for e-commerce big data mining, which reduces the complexity of the merging process while ensuring the effectiveness of the clustering. The introduction of Hierarchical Clustering Algorithm (HCA) improves the clustering effect of big data mining to a certain extent, but it does not take into account the association rules between the data, which leads to low mining accuracy. The association rules are mainly structured mathematical models to objectively reflect the intrinsic correlation between a large amount of data. Hao et al. [14] analyzed association rules for consumers and then used K-means clustering algorithm for educational big data mining, but the running time is long. Li et al. [15] suggested an enhanced fetching approach for text big data on the ground of related semantic integration density gathering, which is with high calculation intricacy and poor real-time digging performance. Fernandez-Basso et al. [16] designed a data digging algorithm relied on fuzzy semantic relation principle characteristic capture and adaptive context-discriminative mapping, but it is susceptible to fuzzy semantic perturbation, which diminishes the anti-perturbation capability of the digging process. Alarifi et al. [17] processed attributes through support metric feature analysis, used similarity algorithm to filter association rules, and combined with bidirectional LSTM to output fuzzy stochastic mining results for big data. Yang et al. [18] proposed log features using association rules for associative hierarchical clustering and output optimized results for data mining through a Navie Bayesian (NB) classifier, which enhances the accuracy of big data digging, but still suffers from high execution intricacy.

1.2. **Contribution.** Although existing research has improved the efficiency of data mining to a certain extent, the ability to deal with the ambiguity and randomness of data is also limited, and it is often difficult to accurately capture the intrinsic laws and patterns of the data, which are affected by the semantic ambiguity factor. For the goal of addressing the above issues, this paper designs a fuzzy stochastic mining algorithm for big data based on hierarchical clustering and key rules. The innovative work of this algorithm is reflected in the following four aspects. (1) Focusing on the issue that the cohesive hierarchical clustering algorithm (HAC) is easily affected by boundary points, the data points are divided into high-density layer, intermediate layer and low-density layer, and hierarchical clustering is carried out within and between the layers respectively to obtain the initial clustering results, and the concept of certainty is introduced to redistribute the boundary points, which improves the accuracy of the clustering of HAC. (2) Based on the association rules between the data, the big data with nonlinear features are fused, and the improved HAC algorithm is adopted to decide the semantic similarity and association of the big data, and the words corresponding to the first several weight coefficients are selected as the semantic association characteristics of the big data through the weighting of the association features by TF-IDF. (3) Combining association rules and feature-weighted optimized Naive Bayes (ONB), adaptive consistent traversal studying methods for information fusion processing of related characteristic quantities in big data digging, to find the optimal solution of the digging objective function, to achieve the big data on-the-fly fuzzy mining. (4) Experimental outcome on the public dataset Forest show that the proposed algorithm's Accuracy and Normalized Mutual Information (NMI) are improved by at least 4.95% and 7.05%, respectively, compared with other algorithms, and it can effectively realize the clustering of big data with high accuracy of data mining.

2. Theoretical analysis.

2.1. Hierarchical clustering algorithm.

HCA is a cluster analysis method that aims to partition the data set through different levels, resulting in a tree-like cluster structure. Compared to clustering algorithms such as k-means [19], k-modes [20], and DBSCAN [21], the advantage of HCA is that it does not require to specify the amount of clusters in advance, and it can provide more detailed information about the clustering structure, which is categorized into Cohesive Hierarchical Clustering (HAC) and Split Hierarchical Clustering (SHC), as shown in Figure 1.

HAC(cohesion)

SHC(disrupt)

Due to the high computational complexity of SHC, HAC is commonly used in practical applications, and its main idea is to start clustering with individual points as clusters, and merge the two clusters with the closest related attributes successively until only one cluster is left at the end or the termination condition is reached. The key of HAC is how to calculate the proximity between two classes. The distance between any two points in various clusters is defined as the proximity between the two points, and the definition of proximity varies in different HAC algorithms, and the most common definitions of proximity are as follows. (1) Single-chain based method [22]. This method defines the shortest distance between any two points between two clusters as the proximity of the two clusters, the single chain algorithm is suitable for dealing with non-ellipsoidal clusters, but it is sensitive to the noise points and outliers, and its proximity distance is calculated as follows.

$$d_{\min}(C_i, C_j) = \min_{x \in C_i,\, y \in C_j} \|x - y\| \tag{1}$$

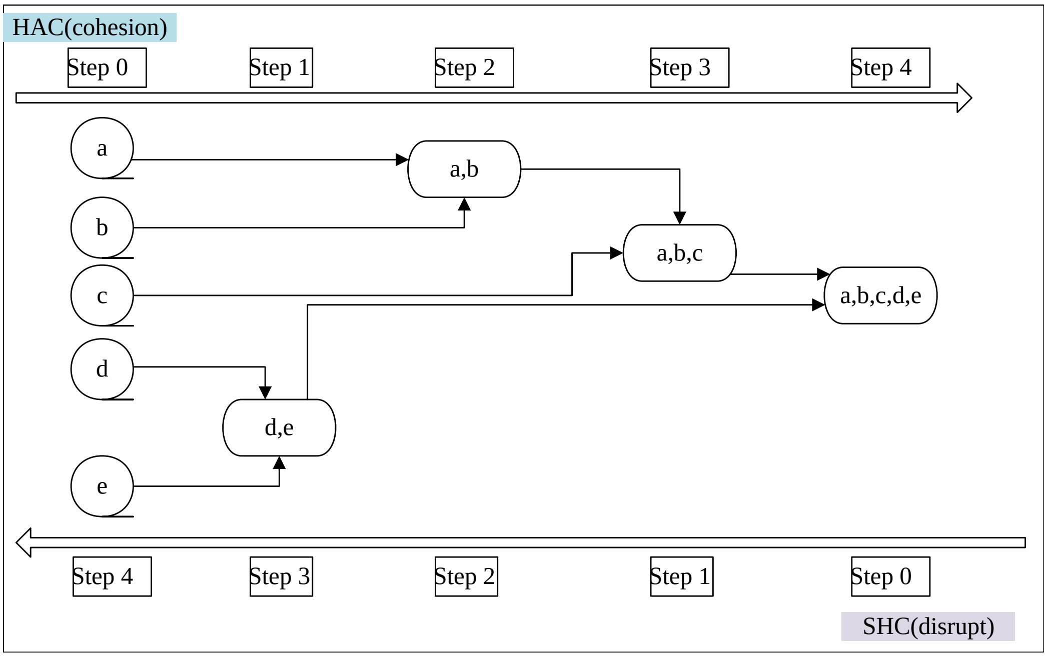

Figure 1. Hierarchical clustering

(2) Full chain based method [23]. In this method the proximity is defined as the longest distance between any two points in C_i, C_j. The algorithm is not sensitive to noise and outliers, but it splits large clusters and applies to spherical clusters. This algorithm is not sensitive to noise and outliers, but it splits large clusters and is applicable to spherical clusters, and the proximity is calculated as follows.

$$d_{\text{avg}}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{x \in C_i} \sum_{y \in C_j} \|x - y\| \tag{2}$$

2.2. **Linked rules grounded theory.** An association is a pattern that exists between two or more variables, and association rule algorithms are responsible for identifying hidden associations between data items in a given dataset, describing the closeness of the data is the main purpose of association analysis. Association rules are used to describe patterns of knowledge that occur regularly and simultaneously between related things [24].

Given a dataset D, $I = \{i_1, i_2, i_3, \ldots, i_m\}$ is a collection of m distinct items from which each transaction T is composed, i.e., $T \subseteq I$. Assume that each transaction T has a transaction identifier TID and that X is a collection of items, when $X \subseteq T$ means that T contains X. The mined association rules are shaped as $A \Rightarrow B$, where $A \subset I$, $B \subset I$, and $A \cap B = \varnothing$, the rule $A \Rightarrow B$ holds only if it satisfies the set minimum support threshold minsup and minimum confidence threshold minconf, and their probabilities are $P(A \cup B)$ and $P(B \mid A)$, respectively. The association rule $A \Rightarrow B$ has support S, denoted S is the percentage of D that contains both A and B, and it is probability $P(A \cup B)$.

$$S(A \to B) = P(A \cup B) = \frac{|A \cup B|}{|D|} \tag{3}$$

where $|D|$ is the total number of transactions in D. Rule $A \Rightarrow B$ has credibility C, which means that C is the probability that the set of items containing A also contains the set of items B. This is the conditional probability $P(B \mid A)$.

$$C(A \to B) = P(B \mid A) = \frac{|A \cup B|}{|A|} \tag{4}$$

where $|A|$ is the number of transactions in D that contain the item set A.

3. **Hierarchical clustering algorithm based on boundary point redistribution optimization.** The HAC algorithm can obtain the clustering tree at one time, and can find the hierarchical relationship between clusters, but it is very easy to be affected by the boundary points, resulting in poor clustering effect. To deal with the above issues, the data are first divided into high-density layer, intermediate layer and low-density layer, and the initial clustering results are obtained by hierarchical clustering within and between the layers, respectively. Then the results are optimized by the boundary point redistribution idea to identify the boundary points and redistribute the boundary points, which improves the accuracy of the HAC algorithm. The optimization flow of the HAC algorithm is shown in Figure 2.

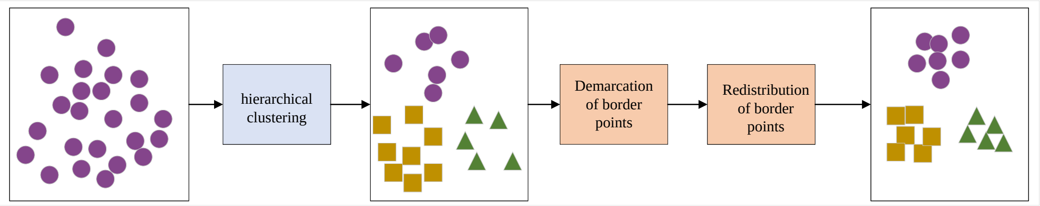

Figure 2. Optimized HAC algorithm flow

Suppose the data set $D = \{x_1, x_2, \ldots, x_n\}$, the number of clusters is k, and the stratification ratio is rate. the distance of the data points is calculated according to Equation (2), and the probability density ρ_i of each data point is computed using Gaussian kernel function.

$$\rho_i = \sum_i \chi(d_{ij} - d_{\mathrm{avg}}) \tag{5}$$

where d_{ij} is the distance to the data points and d_{avg} is the proximity distance. When $x < 0$, $\chi(x) = 1$, $x \geq 0$, $\chi(x) = 0$.

D is then categorized into high-density, intermediate, and low-density layers according to rate pairs, and at each layer the clusters in which the two points that are closest and do not belong to the same cluster are merged. The initial clustering results obtained after hierarchical clustering are not satisfactory, mainly due to the influence of boundary points, resulting in poor clustering accuracy. The traditional HAC algorithm does not involve the redistribution of boundary points, so there is a phenomenon that two similar clusters are wrongly classified into one cluster, therefore, choosing a suitable principle of boundary point allocation will greatly improve the accuracy of clustering. After the above hierarchical clustering preclassification, the dataset has been formed into K clusters and the data points with the highest fitness values in the clusters have been labeled. Based on this, the boundary points of each cluster are defined as the data points belonging to the cluster but whose distance from the data point with the highest fitness value in the cluster is greater than a set radius threshold.

$$\mathrm{BP}_a = \{\, i \mid i \in a,\ d_{ic} > \mathrm{rad},\ c \neq i \,\} \tag{6}$$

where a is the cluster that has been divided, BP_a is the set of boundary points of cluster a, and c is the data point with the largest fitness value in the cluster.

After obtaining the set of boundary points for each cluster, it is necessary to redistribute them. Assuming any boundary point i, the contribution of the data points in the K-nearest-neighbors to i is calculated based on the set of K-nearest-neighbors to boundary point i. Assuming that $i \in \mathrm{BP}_a$, $j \in N_k(i)$ and j have been assigned to cluster a, the contribution of data point j to boundary point i is defined as follows.

$$\mathrm{CON}_a(i) = \mathrm{SUM}\big(i \in N_k(j),\, j \in a\big) \times e^{-d_{ij}} \tag{7}$$

From Equation (7), it can be concluded that when the distance between data points i and j is closer and the more data points in $N_k(j)$ belong to the cluster, it is considered that the contribution of data point j to i belonging to the cluster is greater. Based on the contribution counting data point i, the confidence level for the cluster is defined as follows.

$$\mathrm{CF}(i) = \sum_{j \in N_k(i)} \mathrm{CON}_a \tag{8}$$

After calculating the confidence level of the boundary points belonging to each cluster, the boundary points are divided into the clusters where the maximum value of the confidence level is located, so as to realize the redistribution of the boundary points, thus improving the clustering accuracy of the HAC algorithm.

4. Hierarchical clustering and key rule based fuzzy random mining algorithm for big data.

4.1. Association rule-based fusion of nonlinearly distributed big data. Intending to the existing big data mining algorithms with fuzzy semantic associations, resulting in low mining accuracy, this paper proposes a fuzzy stochastic mining algorithm for big data based on hierarchical clustering and key rules, as indicated in Figure 3. The algorithm is relied on the association rules between the data, fusion processing of big data with nonlinear characteristics, using the EHAC algorithm to gain the big data's semantic association features, clustering analysis of the extracted features, the clustered features as an input to the improved Naive Bayesian classifier, adjustive uniform traversal studying approach to find the optimal solution of the mining objective function, and to achieve the optimization of big data digging.

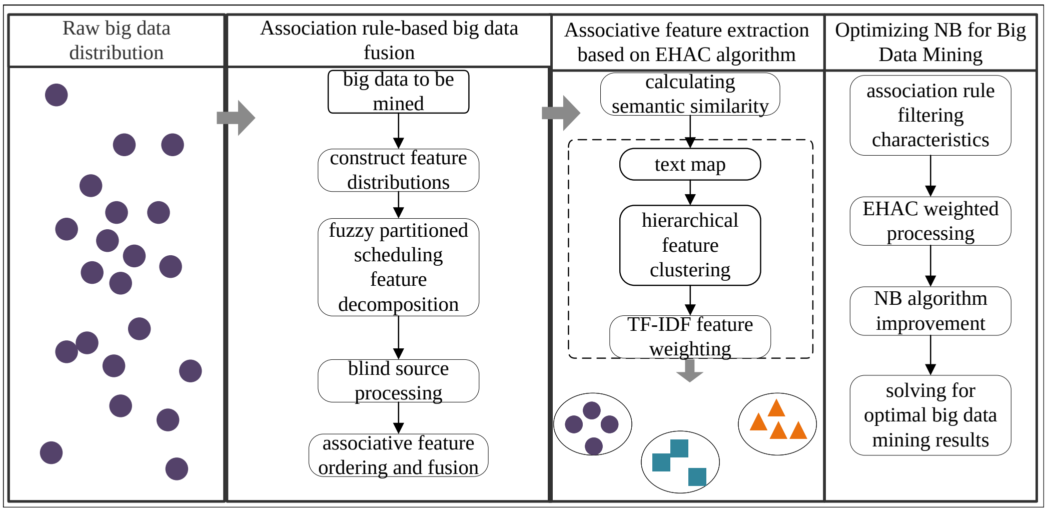
Raw big data distribution
Association rule-based big data fusion
Associative feature extraction based on EHAC algorithm
Optimizing NB for Big Data Mining
big data to be mined
construct feature distributions
fuzzy partitioned scheduling feature decomposition
blind source processing
associative feature ordering and fusion
calculating semantic similarity
text map
hierarchical feature clustering
TF-IDF feature weighting
association rule filtering characteristics
EHAC weighted processing
NB algorithm improvement
solving for optimal big data mining results

Figure 3. Hierarchical clustering and key rule based fuzzy random mining algorithm for big data

The distribution of big data has nonlinear characteristics [25], and the useful data mining needs to be fused according to the correlation between the data to lay a good foundation for the subsequent data feature extraction. Let Y_n be the nonlinear time series of the big data to be fuzzy randomly mined, d be the spatial dimension of the data, then the data mining semantic ontology is as follows.

$$\mathrm{TL}_X(x,y) = \text{Text } f\big(\mathrm{GDX}(x,y) > \theta_X\big) \tag{9}$$

where $\mathrm{TL}_X(x,y)$ is the semantic ontology, $\mathrm{GDX}(x,y)$ is the big data semantic covariance value, and θ_X is the threshold value.

Let l be the length of the big data samples, the similarity attribute ranking, its ranking structure is $Y_n = \{Y_n, Y_{n-1}, Y_{n-2l}, \ldots, Y_{n-(d-1)l}\}$, and the matrix of big data similarity attribute ranking $d * l$ dimensions is R_{d*l}, and the spectrum separation method is used to construct the feature distribution $R = \{Y_1, Y_2, \ldots, Y_d\}^T$ of big data fuzzy stochastic mining. The correlation fusion process is performed by using the association semantic rule for the results of expression formula is as follows.

$$R^T R = \{Y_1, Y_2, \ldots, Y_d\}\{Y_1, Y_2, \ldots, Y_d\}^T \tag{10}$$

In fusion of big data, feature decomposition of the above equation using fuzzy partition scheduling method is shown below.

$$R_1^T R_1' = Q * \Sigma Q^T \tag{11}$$

where Q is the exponential separation matrix. When the feature distribution is transformed from $l+1$-dimensional space to $2l$-dimensional space, the phase space reconstruction algorithm is used to reorganize the associated features to obtain the new features of the associated big data, as shown below.

$$R_2^T R_2' = Q * \Sigma Q^T \tag{12}$$

Based on the result of Equation (12), the information index separation matrix of big data association rules is established, and its expression is as follows.

$$Q = [Q_1, Q_2, \ldots, Q_m] \in \mathbb{R}^{m*m} \tag{13}$$

where m is the number of elements within the matrix. Since the big data association rule information exponential separation matrix is orthogonal matrix [26, 27]], here the exponential universal separation algorithm is used to decompose its features $QQ^T = Q'$, and blind source processing is performed on Q' to obtain Equation (14).

$$O' = \Sigma \operatorname{diag}(O_1', O_2', \ldots, O_m') \in \mathbb{R}^{m*m} \tag{14}$$

The result of the above equation is sorted, so that after the blind source separation of big data associated features are arranged according to the size, the nonlinear distribution of big data fusion is realized.

4.2. **Fuzzy semantic association feature extraction based on improved HAC algorithm.** After the fusion of non-linearly distributed big data, the EHAC algorithm and big data feature correlation are used for feature extraction. Firstly, the big data's semantic similarity is calculated, and the big data semantic association mapping differential is constructed as follows.

$$\tilde{\kappa}_i = f_i(\kappa_i, u_j, u_j)_{i,j} \tag{15}$$

where i and j are the number of feature clustering dimensions and the number of sampling points, κ_i is the big data sample, $\tilde{\kappa}_i$ is the semantic association mapping differential value, u_i and u_j are the feature independent variables.

Contextual ontology mapping process is applied to Equation (15), and the discrete degree generalized relations between ontology classes of big data after approximation [28] are obtained as follows.

$$\tilde{\kappa} = f(\kappa, u) \tag{16}$$

In solving Equation (16), the target vocabulary filtering approach is used to calculate the big data feature independent variable $u \in \mathbb{R}^{m*n}$, where m and n are the spatial and feature dimensions, respectively. Let d_i and d_j be the big data mapping distribution trajectories, and calculate the distance $D(d_i, d_j)$ between the two distribution trajectories corresponding to the discrete classes of semantically related feature distributions, as shown below.

$$D(d_i, d_j) = \frac{1}{d_i * d_j} * d_i d_j \tag{17}$$

Based on the above equation, we obtain $D(d_i, d_j)$, and use the mutual information form to describe its relationship between contexts in the distribution space, and calculate the mutual information $I(w_i, k)$ of big data text words w_i and feature words k.

$$I(w_i, k) = \log_2 \frac{P(w_i \mid k)}{P(k)} \tag{18}$$

where $P(w_i \mid k)$ is the number of occurrences of w_i in the interval of action of k, and $P(k)$ is the statistical probability of k, i.e., $P(k) = \mathrm{fre}(k)/n$. Subsequently, use the words of $P(k) > 0$ to construct a set of key feature words of big data text K. Combine the contextual information of the word to select the feature word K', so that the feature vector of the feature word is $\mathrm{Term} = \{t_1, t_2, \ldots, t_n\}$, where the word feature $t_i \in (K \cup K')$. The TF-IDF function was used to calculate the weight values of t_i as follows.

$$G = \max(\mathrm{TFIDF}(t_i, e)) \tag{19}$$

where G is the maximum weight value of t_i, d is the frequency of occurrence of t_i, and $\mathrm{TFIDF}(t_i, e)$ is the weight of t_i. The expression is as follows.

$$TFIDF(\ t_i, e\ ) == \frac{t_i f(t_i, e) * (\log_2 \frac{N}{n_{t_i}} + \log_2 \zeta)}{\sqrt{\sum\limits_{\in e} [t_i f(t_i, e)]^2 * (\log_2 \frac{N}{n_{t_i}} + \log_2 \zeta)}} \tag{20}$$

where ζ is the importance factor. After obtaining the weight coefficients of the textual features of big data by the above formula, the words corresponding to the first several weight coefficients are selected as the semantically related features of big data after descending the order.

4.3. **Association rule based optimized Naive Bayesian classifier for big data mining.** After obtaining the semantic association features of big data, they are used as inputs to NB, which is utilized to classify them, thus obtaining the big data fuzzy random mining results. However, when there are highly correlated attributes in the dataset, the mining accuracy of the traditional NB algorithm cannot reach the maximum. To enhance the issue of different influence degree of different features, we introduce association rules to filter high correlation features and weight different features with EHAC. For example, the flow of the feature f and the overall mining U is shown as follows. (1) Calculate and find the maximum values of the conditional probabilities of , and for each category in f, denoted by , and , where , if it fails to satisfy , then f has a weak effect on the excavation effect and is deleted. (2) Calculate the confidence means , , and of f that are positively correlated with f and the correlation of the results, such that , yields the weighted result
.

(21)

where is the value of , M is the total number of features, is the corresponding , and , is the probability of , and denotes the probability that belongs to . Through the above algorithm, NB is improved (ONB), and ONB is used to complete fuzzy stochastic mining of big data. Since ONB assumes that the relationship between the feature attributes of big data features is independent of each other, the joint probability of big data text features can be calculated by the product of the distribution probability of each feature, as shown below.

(22)

Bringing Equation (22) into Equation (21), there is obtained Equation (23).

(23)

where D is the dataset with mining, through the recursive algorithm can be solved to get the maximum value of the above equation can get the optimal big data mining results.

5. Performance testing and analysis.

5.1. Model predictive performance analysis. For the goal of comprehensively verifying the practicality and effectiveness of the proposed big data mining algorithm in real scenarios, an experimental framework based on MATLAB software and VC++ compilation environment was built. The simulation hardware environment is as follows: Intel Core 3-530 CPU, 1 Gigabyte RAM, and Windows 10. The Forest dataset [29] in UCI is selected as the core dataset for the experiments, which covers up to 50 000 large data samples with rich semantic information and practical application value. To ensure the accuracy and reliability of the experiments, the training set is divided into 500 sets according to their semantic association features, the sampling frequency of the big data is 10kHz, and the range of fuzzy differentiated association semantic sets is 36. The SVTC algorithm [15], HCDC algorithm [18] and HCKR algorithm proposed in this paper are compared experimentally on Forest dataset, where the performance of data mining algorithms is evaluated using Accuracy, F-measure, NMI and ARI metrics. Detailed information of these four performance evaluation metrics is available in the literature [30], and all of them are positively correlated with data mining effectiveness. The experimental results are shown in Table 1. The Accuracy and F-measure of HCKR are 90.95% and 92.57%, respectively, which are at least 4.95% higher compared to SVTC and HCDC. HCKR optimizes the hierarchical clustering algorithm by redistributing the boundary points, extracts the semantically related features of the big data by using the EHAC, and mines the feature data accurately by the optimized NB, which greatly improves the accuracy of data mining. The NMI and ARI of HCKR were improved by 7.05%–15.29% compared to SVTC and HCDC, and SVTC was poorly clustered by density clustering when dealing with a large amount of data. Although HCDC uses HAC to cluster big data text log features, it is not optimized for HAC and NB, so the mining performance is not as good as HCKR.

Table 1. Comparison of data mining effectiveness of different algorithms (%).

Algorithm	Accuracy	F-measure	NMI	ARI
SVTC	74.68	79.05	74.85	69.36
HCDC	84.18	87.62	80.19	75.05
HCKR	90.95	92.57	87.24	84.65

In the process of data mining, if underfitting occurs, it means that the data mining results are inaccurate, combined with MATLAB will be three algorithms of the fitting

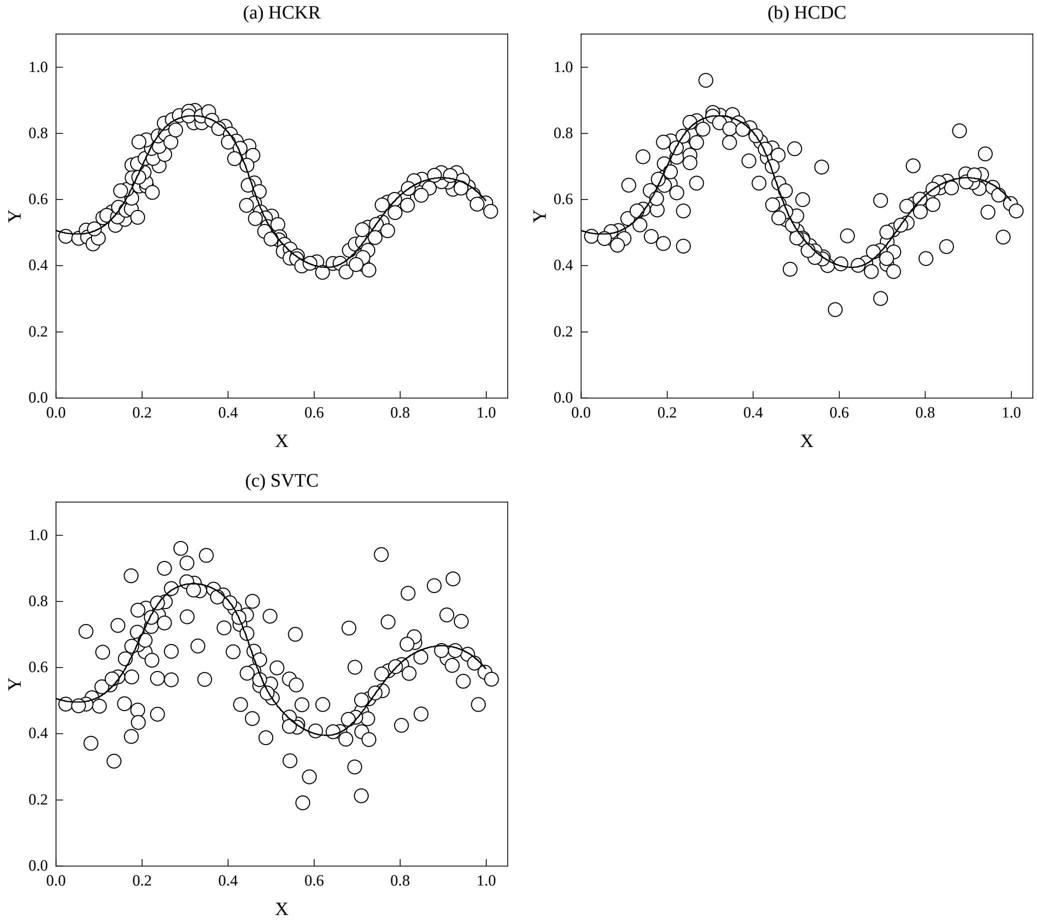

Figure 4. Comparison of reliability experiment results for big data mining

curve and the distribution of data points are plotted, as shown in Figure 4. First, observing the fitting curve of HCKR, almost every data point fits closely on the curve, reflecting its strong generalization ability and stability. The fitted curves of SVTC, while fitting the data well in some regions, show significant deviations in other regions, indicating that the control SVTC does not fully and accurately reflect the true state of the data. A large number of data points of HCDC are far away from the fitted curve, implying that it may not be able to handle the ambiguity and randomness of the data effectively. Taken together, HCKR performs best in terms of consistency between the fitted curve and the data distribution, and its mining results are the most reliable.

5.2. **Analysis of the results of the ablation experiments of the model.** To compare the clustering ability of the three mining algorithms, 20,000 data in the above big data sample collection are used as training samples, and the three mining methods are used to mine the big data, and the distribution of the real and imaginary parts of the big data after mining are given, and the results of the mining are plotted as shown in Figure 5, and the difference in the clustering ability of the different algorithms in dealing with the big data samples can be clearly observed.

The data in HCKR exhibit a highly regular distribution of states. Data with different features are tightly clustered, each cluster is compact and clear, and the distance between data points is moderate, without appearing too dense or too sparse. In contrast, SVTC and HCDC are haphazard. the data distribution of SVTC is not regular, and the data boundaries between different features are blurred, making it difficult to accurately

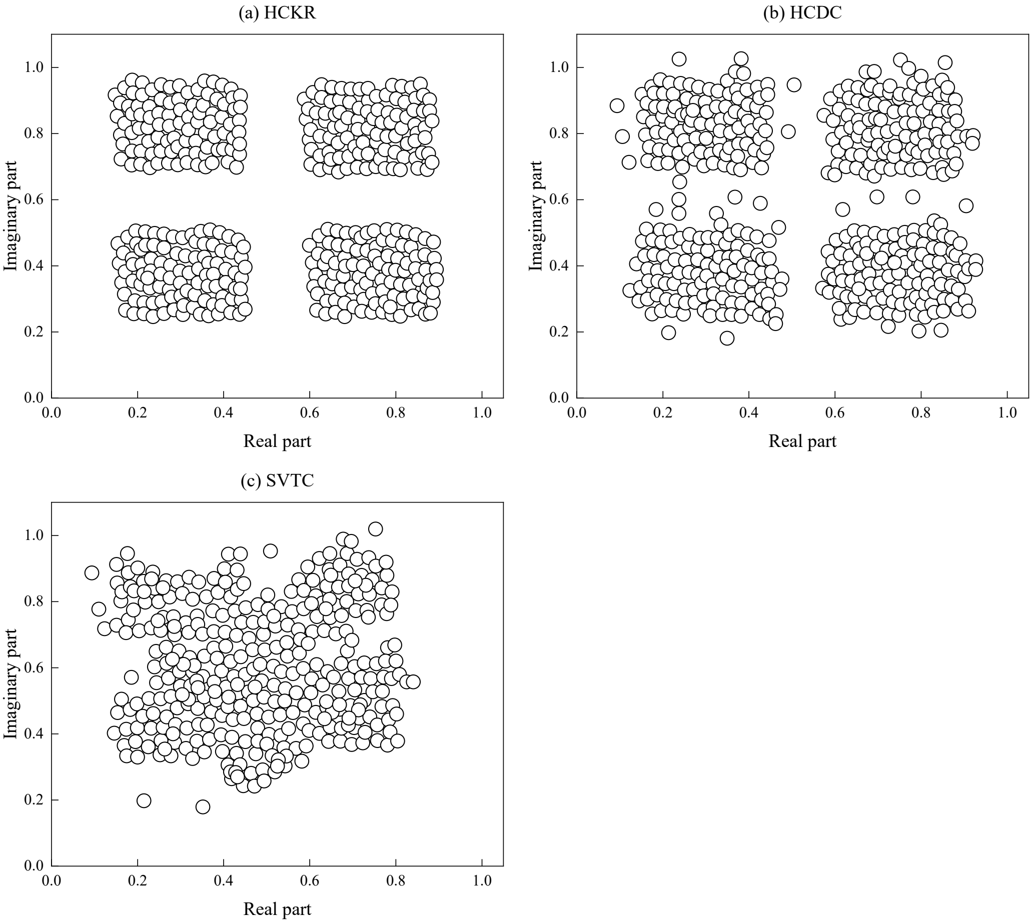

Figure 5. Comparison of big data mining clustering effects

delineate them. And the data distribution of HCDC is not regular at all, the data of different features are mixed together, and effective clustering cannot be realized. Therefore, HCKR has the best performance in terms of clustering effect. The algorithm is not only able to accurately capture the intrinsic correlation between the data, but also able to realize effective data clustering, providing strong support for subsequent data analysis and decision-making.

6. **Conclusion.**

The process of big data mining is affected by semantic ambiguity factors, which leads to unsatisfactory mining accuracy, so a fuzzy stochastic mining algorithm for big data based on hierarchical clustering and key rules is proposed. Aiming at the problem that the HAC algorithm is easily affected by the boundary points, which leads to poor clustering effect, the HAC is optimized by dividing the data points into high-density layer, intermediate layer and low-density layer, respectively, and carrying out hierarchical clustering within the layer and between the layers to get the initial clustering results, and introducing the concept of certainty to redistribute the boundary points, so as to improve the clustering effect. On this basis, based on the association rules between the data, the EHAC algorithm is adopted to decide the semantic similarity and association of the big data, and the associated features are weighted by the TF-IDF, and the words corresponding to the first several weighting coefficients are selected as the semantic associated characteristics of the big data. Finally, the proposed algorithm combines association rules and feature weighted optimization NBC, adaptive uniform traversal learning method for information fusion processing of associated features in big data mining, to find the optimal solution of the mining objective function, and to realize the random fuzzy mining of big data. Comparison experiments prove that the proposed algorithm has higher mining reliability, stronger mining clustering ability, and more ideal mining accuracy. The research in this paper has made some performance breakthroughs in big data mining, but due to my limited energy and time, there are still shortcomings in the improvement and application of the algorithm. In future research, it is hoped that the operational efficiency of the EHAC algorithm can be further improved, the algorithm can be tried to be improved by using the sampling strategy, and the principle of the method can be utilized to expand the scope of problem solving.

REFERENCES

[1] F. Coenen, "Data mining: past, present and future," The Knowledge Engineering Review, vol. 26, no. 1, pp. 25-29, 2011.

[2] P. Sunhare, R. R. Chowdhary, and M. K. Chattopadhyay, "Internet of things and data mining: An application oriented survey," Journal of King Saud University-Computer and Information Sciences, vol. 34, no. 6, pp. 3569-3590, 2022.

[3] J. Yang, Y. Li, Q. Liu, L. Li, A. Feng, T. Wang, S. Zheng, A. Xu, and J. Lyu, "Brief introduction of medical database and data mining technology in big data era," Journal of Evidence-Based Medicine, vol. 13, no. 1, pp. 57-69, 2020.

[4] D. K. Jadhav, "Big data: the new challenges in data mining," International Journal of Innovative Research in Computer Science & Technology, vol. 1, no. 2, pp. 39-42, 2013.

[5] L. Huang, X. Ni, W. L. Ditto, M. Spano, P. R. Carney, and Y.-C. Lai, "Detecting and characterizing high-frequency oscillations in epilepsy: a case study of big data analysis," Royal Society Open Science, vol. 4, no. 1. 160741, 2017.

[6] H. Jung, and K. Chung, "Social mining-based clustering process for big-data integration," Journal of Ambient Intelligence and Humanized Computing, vol. 12, no. 1, pp. 589-600, 2021.

[7] W. Li, J. Zhu, Y. Zhang, and S. Zhang, "Design and implementation of intelligent traffic and big data mining system based on internet of things," Journal of Intelligent & Fuzzy Systems, vol. 38, no. 2, pp. 1967-1975, 2020.

[8] F. F. Bocca, and L. H. A. Rodrigues, "The effect of tuning, feature engineering, and feature selection in data mining applied to rainfed sugarcane yield modelling," Computers and Electronics in Agriculture, vol. 128, pp. 67-76, 2016.

[9] S. Riaz, M. Fatima, M. Kamran, and M. W. Nisar, "Opinion mining on large scale data using sentiment analysis and k-means clustering," Cluster Computing, vol. 22, pp. 7149-7164, 2019.

[10] M. Subramaniam, A. Kathirvel, E. Sabitha, and H. A. Basha, "Modified firefly algorithm and fuzzy c-mean clustering based semantic information retrieval," Journal of Web Engineering, vol. 20, no. 1, pp. 33-52, 2021.

[11] M. Ianni, E. Masciari, G. M. Mazzeo, M. Mezzanzanica, and C. Zaniolo, "Fast and effective Big Data exploration by clustering," Future Generation Computer Systems, vol. 102, pp. 84-94, 2020.

[12] A. Dogan, and D. Birant, "K-centroid link: a novel hierarchical clustering linkage method," Applied Intelligence, vol. 10, pp. 1-24, 2022.

[13] K. K. Pandey, and D. Shukla, "NDPD: an improved initial centroid method of partitional clustering for big data mining," Journal of Advances in Management Research, vol. 20, no. 1, pp. 1-34, 2023.

[14] L. Hao, T. Wang, and C. Guo, "Research on parallel association rule mining of big data based on an improved K-means clustering algorithm," International Journal of Autonomous and Adaptive Communications Systems, vol. 16, no. 3, pp. 233-247, 2023.

[15] H. Li, J. Liu, K. Wu, Z. Yang, R. W. Liu, and N. Xiong, "Spatio-temporal vessel trajectory clustering based on data mapping and density," IEEE Access, vol. 6, pp. 58939-58954, 2018.

[16] C. Fernandez-Basso, M. D. Ruiz, and M. J. Martin-Bautista, "Spark solutions for discovering fuzzy association rules in Big Data," International Journal of Approximate Reasoning, vol. 137, pp. 94-112, 2021.

[17] A. Alarifi, A. Tolba, Z. Al-Makhadmeh, and W. Said, "A big data approach to sentiment analysis using greedy feature selection with cat swarm optimization-based long short-term memory neural networks," The Journal of Supercomputing, vol. 76, pp. 4414-4429, 2020.

[18] Q.-F. Yang, W.-Y. Gao, G. Han, Z.-Y. Li, M. Tian, S.-H. Zhu, and Y.-h. Deng, "HCDC: A novel hierarchical clustering algorithm based on density-distance cores for data sets with varying density," Information Systems, vol. 114. 102159, 2023.

[19] A. Likas, N. Vlassis, and J. J. Verbeek, "The global k-means clustering algorithm," Pattern Recognition, vol. 36, no. 2, pp. 451-461, 2003.

[20] F. Cao, J. Liang, D. Li, L. Bai, and C. Dang, "A dissimilarity measure for the k-modes clustering algorithm," Knowledge-Based Systems, vol. 26, pp. 120-127, 2012.

[21] A. Latifi-Pakdehi, and N. Daneshpour, "DBHC: A DBSCAN-based hierarchical clustering algorithm," Data & Knowledge Engineering, vol. 135. 101922, 2021.

[22] P. Daie, and S. Li, "Hierarchical clustering for structuring supply chain network in case of product variety," Journal of Manufacturing Systems, vol. 38, pp. 77-86, 2016.

[23] X. Ran, Y. Xi, Y. Lu, X. Wang, and Z. Lu, "Comprehensive survey on hierarchical clustering algorithms and the recent developments," Artificial Intelligence Review, vol. 56, no. 8, pp. 8219-8264, 2023.

[24] S. Kotsiantis, and D. Kanellopoulos, "Association rules mining: A recent overview," GESTS International Transactions on Computer Science and Engineering, vol. 32, no. 1, pp. 71-82, 2006.

[25] Q. Zhang, L. T. Yang, and Z. Chen, "Deep computation model for unsupervised feature learning on big data," IEEE Transactions on Services Computing, vol. 9, no. 1, pp. 161-171, 2015.

[26] A. Fahad, N. Alshatri, Z. Tari, A. Alamri, I. Khalil, A. Y. Zomaya, S. Foufou, and A. Bouras, "A survey of clustering algorithms for big data: Taxonomy and empirical analysis," IEEE Transactions on Emerging Topics in Computing, vol. 2, no. 3, pp. 267-279, 2014.

[27] A. K. Sangaiah, S. Rezaei, A. Javadpour, and W. Zhang, "Explainable AI in big data intelligence of community detection for digitalization e-healthcare services," Applied Soft Computing, vol. 136. 110119, 2023.

[28] P. Pieta, and T. Szmuc, "Applications of rough sets in big data analysis: an overview," International Journal of Applied Mathematics and Computer Science, vol. 31, no. 4, pp. 659-683, 2021.

[29] S. Lakshmanaprabu, K. Shankar, M. Ilayaraja, A. W. Nasir, V. Vijayakumar, and N. Chilamkurti, "Random forest for big data classification in the internet of things using optimal features," International Journal of Machine Learning and Cybernetics, vol. 10, no. 10, pp. 2609-2618, 2019.

[30] N. Bharill, A. Tiwari, and A. Malviya, "Fuzzy based scalable clustering algorithms for handling big data using apache spark," IEEE Transactions on Big Data, vol. 2, no. 4, pp. 339-352, 2016.