

International Journal of Reasoning-based Intelligent Systems

ISSN online: 1755-0564 - ISSN print: 1755-0556
<https://www.inderscience.com/ijris>

Fast classification algorithm of uncertain big data stream based on distributed limit learning machine

Zhiling Yang

DOI: [10.1504/IJRIIS.2023.10057747](https://doi.org/10.1504/IJRIIS.2023.10057747)

Article History:

Received:	22 February 2023
Last revised:	05 May 2023
Accepted:	05 May 2023
Published online:	18 March 2025

Fast classification algorithm of uncertain big data stream based on distributed limit learning machine

Zhiling Yang

Department of Information Engineering,
Zhujiang College of South China Agricultural University,
Guangzhou, 510900, China
Email: yzlzhhr@126.com

Abstract: Aiming at the problems of low classification accuracy and long classification time in traditional uncertain big data stream classification algorithm, a fast classification algorithm for uncertain big data stream based on distributed limit learning machine is proposed. First, construct the uncertain big data model, and measure the adaptive mixed distance of the data flow. Then, according to the measurement results, use the DSUFIM – mine mining algorithm to mine the uncertain big data flow. Then combine the K-nearest neighbour and label correlation to conduct online feature screening for the mined uncertain big data flow, and use the cosine similarity method to detect whether the feature has conceptual drift. Finally, according to the detection results, the distributed limit learning machine is used to quickly classify uncertain big data streams. The experimental results show that the proposed algorithm has higher classification accuracy and faster classification speed for uncertain large data streams.

Keyword: distributed limit learning machine; uncertain big data flow; quick classification; K nearest neighbour; cosine similarity.

Reference to this paper should be made as follows: Yang, Z. (2025) 'Fast classification algorithm of uncertain big data stream based on distributed limit learning machine', *Int. J. Reasoning-based Intelligent Systems*, Vol. 17, No. 1, pp.73–81.

Biographical notes: Zhiling Yang received her Master's degree in Computer Technology from School of Computer Science and Engineering, South China University of Technology in 2015. She is currently a Lecturer in Zhujiang College of South China Agricultural University. Her research interests include big data technology, data mining, and machine learning.

1 Introduction

With the continuous development of information technology, data exists in the form of data flow and grows in gradient. In the big data environment, due to external influences, privacy protection, data loss, and other reasons, it is not possible to determine whether data items in the big data stream exist, resulting in the formation of uncertain data streams. Therefore, big data streams are classified into deterministic data streams and uncertain data streams (Lin et al., 2021a; Yang et al., 2020). For the classification of big data streams, it is necessary to deeply analyse the characteristics of data streams and the uncertainty of data to finally solve the problem efficiently. Because the amount of data on the uncertain data stream is generally massive, and the data on the data stream is uncertain, the most important challenge for classification processing of uncertain data stream is how to complete the concept drift and uncertainty processing of data stream, and how to efficiently process the increasingly large data information, and fully consider the rapidity of data arrival (Wang et al., 2022).

Shi and Huang (2020) proposes an uncertain big data stream classification algorithm based on the density peak to obtain the uncertain big data stream. Under the online and offline two-level processing framework, the uncertain big

data stream is characterised by the density peak to obtain the data stream feature vector, and the uncertain big data stream is classified by combining the attenuation function and the micro-cluster deletion mechanism. Tu and Chen (2021) proposes an uncertain data stream classification algorithm based on the attenuation window, collects the uncertain data stream of the online part through the sub-window sampling mechanism, stores the data stream using the two-level summary statistical structure linked list, pre-processes the stored uncertain data stream through the attenuation window mechanism, and clusters the uncertain data stream using the dynamic exception grid deletion mechanism according to the pre-processing results, Get classification results. However, the above two algorithms do not use mixed distance measures for data streams, resulting in low accuracy and poor classification results for uncertain large data streams. Zhou and Liu (2020) proposes an uncertain data flow classification algorithm based on multi-source migration learning method. First, the uncertain data flow is obtained, and then the multi-source migration learning method is used to build a multi-source classifier, calculate the class posterior probability of each source classifier for each sample in the data block, output the sample deterministic value based on the calculation results,

and use this value as a threshold to classify the uncertain data flow. Hu et al. (2020) proposes a distributed fast short text data stream classification algorithm based on the FTRL method, establishes a distributed fast short text corpus, obtains the short text data stream, uses the FTRL method to update the short text data stream in real-time, uses the time factor weighting mechanism to reduce the dimension of the updated data stream, and uses the LR classifier to classify the short text data stream according to the processing results to obtain the classification results. However, in the process of classifying uncertain data streams using the above two algorithms, there is no detection of conceptual drift in the characteristics of uncertain data streams, resulting in longer classification time, slower classification speed, and lower classification efficiency.

The distributed limit learning machine is a simple and effective SLFNs learning algorithm for single hidden layer feedforward neural networks. Traditional neural network learning algorithms (such as BP algorithm) require a large number of network training parameters to be manually set, and are prone to generate local optimal solutions. The limit learning machine only needs to set the number of hidden layer nodes of the network, and does not need to adjust the input weights and hidden element bias of the network during the algorithm execution process, and produces a unique optimal solution. Therefore, it has the advantages of fast learning speed and good generalisation performance. Aiming at the problems of the above algorithms and aiming at achieving fast classification of uncertain big data streams, this paper proposes a fast classification algorithm of uncertain big data streams based on distributed limit learning machine. The specific research route is as follows:

- 1 Under the analysis of the causes of uncertain data, the uncertain big data is modelled by the possible world semantic model, and in the established uncertain big data model, the data flow is measured by adaptive hybrid distance through three steps: kernel distance, Jensen-Shannon divergence and adaptive weight
- 2 Use the DSUFIM – mine mining algorithm to mine the data stream in the uncertain big data model, and store the mined uncertain big data stream into the UIT-lists structure
- 3 Because the multi-label data in the mined uncertain big data stream shows a large number of features and improves the data dimension, in order to solve the problem of dimension disaster, this paper uses online feature selection method to screen the uncertain big data stream online features based on K-nearest neighbour and label correlation, so as to reduce the dimension effect
- 4 According to the online feature screening results of uncertain big data flow, the cosine similarity method is used to detect whether the concept drift of uncertain big data flow features occurs
- 5 The distributed limit learning machine is used to quickly classify the uncertain big data stream. The

selected online features of the uncertain big data stream are used as the input dataset, and the input weight value, hidden layer node offset value, and output weight value are calculated to obtain the fast classification output result of the uncertain big data stream.

2 Rapid classification of uncertain big data streams

2.1 Adaptive hybrid distance measure for uncertain big data flow

First, analyse the reasons for the generation of uncertain data. Then, in order to obtain uncertain big data streams, it is necessary to model uncertain big data using a possible world semantic model. Finally, in the established uncertain big data model, adaptive hybrid distance measures are applied to the data streams through three steps: kernel distance, Jensen Shannon divergence, and adaptive weight.

2.1.1 Uncertain big data modelling

Because big data contains uncertain data, in order to obtain uncertain big data streams, it is necessary to model uncertain big data through a possible world semantic model. Before establishing the uncertain big data model, first analyse the causes of the uncertain big data. In many application fields, data acquisition and processing are affected by various aspects, and the existence of data is usually inaccurate. Therefore, with the deepening of research, people increasingly realise the importance of studying uncertain data, and the first step of studying uncertain data is to understand its formation (Zhang et al., 2020). The reasons for the formation of uncertain data are mainly divided into the following aspects:

- 1 The original data is inaccurate. In the process of data acquisition, due to the accuracy limitation of the instrument itself, network delay and other factors, the specific differences in the process of data acquisition result in a large number of uncertain data, which is the most direct reason for the generation of uncertain data.
- 2 Meet the requirements of special applications. Some private data can not give the original data directly, and will interfere with the original data by adding noise data, such as bank deposits and loans, telecommunications, etc.
- 3 Processing of missing values. Due to equipment failure or data integration process, some missing values are generated due to various reasons. A common method to deal with missing values is to manually supplement the value based on experience, which also creates uncertainty in the original data.

The uncertain data is divided into existence level and attribute level. The existence level uncertainty data is that the value of each object is known and certain, but the

existence of the object is uncertain. The probability is used to represent the possibility of the existence of the data object. Attribute-level uncertain data refers to that each data object has multiple possible values, which does not involve the uncertainty of the entire object, that is, the object is certain to exist, but the value of each feature of the object is uncertain. The traditional data model is not applicable to this kind of data, so a new uncertain big data model should be established to obtain the uncertain big data flow when analysing (Mousavi et al., 2022).

At present, there are many models for uncertain data, including possible world semantic model, data flow data model, semi-structured data model, relational data model, and multi-dimensional data model. The modelling of uncertain big data models mainly includes probabilistic or-set table and Probabilistic table. Among them, the most widely used is the possible world semantic model, and other models can be converted into the possible world model (Zhu et al., 2020; Zebhi et al., 2020). For this reason, this article constructs an uncertain big data model through the possible world semantic model.

2.1.2 Adaptive hybrid distance measure

In the established uncertain big data model, adaptive mixed distance measurement is applied to the data flow. The adaptive hybrid distance measure consists of three parts: kernel distance, Jensen-Shannon divergence and adaptive weight. The details are as follows:

2.1.2.1 Nuclear distance

Kernel distance has been widely used in measuring the geometric distance between determined data. In this section, a kernel distance based on Gaussian kernel is used to measure the geometric distance between uncertain data streams (Liu et al., 2020).

Given two uncertain data points P and Q , $C(P)$ and $C(Q)$ represent the centroids of P and Q respectively, the nuclear distance $KD(P||Q)$ between P and Q is defined as follows:

$$KD(P||Q) = 1 - \exp\left(-\frac{\|C(P) - C(Q)\|_2}{2\sigma^2}\right) \quad (1)$$

where, σ represents the parameter, and σ is set to be equal to the median Euclidean distance between the centres of gravity of all uncertain data points in the dataset.

According to the above formula, the nuclear distance satisfies the symmetry and boundedness $KD(P||Q) \in (0, 1]$. In addition, due to the introduction of kernel function, the kernel distance is more suitable for processing linearly indivisible data (Lin et al., 2021b).

2.1.2.2 Jensen-Shannon divergence

Jensen-Shannon divergence is a widely used probability distribution distance measure. In the discrete case, the Jensen-Shannon divergence $JSD(f||g)$ between $f(x)$ and $g(x)$ is defined as follows:

$$JSD(f||g) = \frac{1}{2} \left(\sum_{x \in R} \frac{2f(x)}{f(x) + g(x)} \right) + \left(\sum_{x \in R} \log \frac{2g(x)}{f(x) + g(x)} \right) \quad (2)$$

where, $f(x)$ and $g(x)$ represent probability mass functions in different spaces.

In the continuous case, the Jensen-Shannon divergence $JSD(f||g)$ between $f(x)$ and $g(x)$ is defined as follows:

$$JSD(f||g) = \frac{1}{2} \left(\sum_{x \in R} f(x) \log \frac{2f(x)}{f(x) + g(x)} \right) + \left(\sum_{x \in R} g(x) \log \frac{2g(x)}{f(x) + g(x)} \right) \quad (3)$$

From the above formula, Jensen-Shannon divergence is symmetric, and when the base of log is 2, its value range is $[0, 1]$.

Given two uncertain data points P and Q according to equations (2) and (3), we can calculate the Jensen-Shannon divergence between them according to the following formula:

$$JSD(P||Q) = \frac{1}{2} \left(E \left(\log \frac{2P}{P+Q} \right) + E \left(\log \frac{2Q}{P+Q} \right) \right) \quad (4)$$

where, E represents expectation, P and Q represent probability distribution corresponding to uncertain data points.

Assuming that the sample set of uncertain data point P is $\{p_1, p_2, \dots, p_r\}$ and the sample set of Q is $\{q_1, q_2, \dots, q_r\}$, calculate the Jensen-Shannon divergence between uncertain data points P and Q :

$$JSD(P||Q) = \frac{1}{2} \left(\frac{1}{r} \sum_{i=1}^r \log \frac{2P(p_i)}{P(p_i) + Q(p_i)} + \frac{1}{t} \sum_{i=1}^t \log \frac{2Q(q_i)}{P(q_i) + Q(q_i)} \right) \quad (5)$$

where, p_i represents the sample of uncertain data point P , and q_i represents the sample of uncertain data point Q .

According to the above formula, the calculated Jensen-Shannon divergence satisfies the symmetry and boundedness.

2.1.2.3 Adaptive weight

Given two uncertain data points o_1 and o_2 , $R(o_1)$ and $R(o_2)$ represent the uncertain areas of o_1 and o_2 respectively, and $v(o_1)$ and $v(o_2)$ represent the super-volume of o_1 and o_2 respectively, then the position overlap $\rho(o_1, o_2)$ between o_1 and o_2 is:

$$\rho(o_1, o_2) = \frac{v(R(o_1) \cap R(o_2))}{\min\{v(o_1), v(o_2)\}} \quad (6)$$

where, $v(R(o_1) \cap R(o_2))$ represents the hypervolume of uncertain regions $R(o_1)$ and $R(o_2)$. For any two uncertain data points, if their position overlap is greater than 0.5, then

we think that these two data points overlap seriously (Rambabu and Venkatram, 2021).

Given the uncertain dataset $UD = \{o_1, o_2, \dots, o_n\}$, n represents the number of data points in the dataset, then the position serious overlap degree W_{UD} of UD is:

$$W_{UD} = \frac{\sum_{i,j=1}^n \delta[\rho(o_i, o_j) \geq 0.5]}{n(n-1)} \quad (7)$$

where, $\delta[\cdot]$ is the logical judgment function. According to the above formula, W_{UD} is in direct proportion to the logarithm of the uncertain data points with serious overlapping positions in the dataset.

The position overlap degree of uncertain datasets can be determined adaptively according to the datasets. According to the observation of the tendency of the dataset to measure the distance, we will use the position serious overlap of the dataset as the weight parameter to adjust the importance of the kernel distance and the probability distribution distance (Chen, 2021).

According to the above steps, the adaptive mixing distance measure of uncertain big data flow is:

$$SMDM(P \parallel Q) = (1 - W_{UD})KD(P \parallel Q) + W_{UD}JSD(P \parallel Q) \quad (8)$$

2.2 Uncertain big data stream mining

According to the results of adaptive mixed distance measurement of uncertain big data streams calculated above, the DSUFIM – mine mining algorithm is used to mine uncertain big data streams.

Set $UDS = \{T_1, T_2, \dots, T_n\}$ to represent an indeterminate large data flow with infinite arrival. The collection of data items is $I = \{I_1, I_2, \dots, I_n\}$, UDS is composed of several different tuples, which are also called transactions, while tuples are composed of several different items, and transactions are identified by transaction number TID , where $I_1, I_2, \dots, I_n$ is different transactions that arrive in chronological order.

The uncertain data flow UDS can also represent several basic windows that arrive continuously, and the basic window is composed of fixed-length transactions. Assuming that the size of each basic window is L , the sliding window can be represented as $SW = (SW_1, SW_2, \dots, SW_n)$, and the maximum number of windows that can be accommodated is represented by SW , that is, the maximum value of SW . SW moves in the basic window unit with the arrival of the following time. When the window reaches the maximum value, the oldest window is deleted and a new window is added.

- 1 $|UDS|$ refers to the total length of the transaction, and the expected support $E[s(X)]$ of project X is:

$$E[s(X)] = \sum_{j=1}^{|UDS|} p(X, t_j) \quad (9)$$

where, t_i represents tuple.

The probability $p(X, t_i)$ of X in tuple t_i is the product of the probabilities of X in t_i tuple, and its expression is:

$$p(X, t_i) = \prod_{x \in X} p(x, t_i) \quad (10)$$

Assume that the true expected support of item set X is $E[s_{true}(X)]$, the estimated expected support of item set X is $E[s_{est}(X)]$, the estimated expected support of attenuation is $E[s_{DEST}(X)]$, the expected support of current item set X is $E[s_{curr}(X)]$, the error factor in the summary structure is Δ , N is the total length of the uncertain data stream, and N_i is the length of the event when the batch i data stream arrives.

- 2 In UDS, set the error parameter as ε and the support threshold as σ . If the itemset X is a frequent itemset, it must comply with $E[s_{true}(X)] \geq \sigma N$.
- 3 If the data flow number of the current batch is i and the attenuation factor is γ , then the expected support of itemset X in the i batch of data can be defined as:

$$E_i[s(X)] = E_{i-1}[s(X)] * \gamma + E_i[s_{curr}(X)] \quad (11)$$

The attenuation factor is defined as D_i , and the attenuation length of the data stream can be defined as:

$$N_i = wD_i \quad (12)$$

- 4 Given an UDS, the attenuation length is wD_i , the support threshold is σ , and the attenuation factor is γ . If the itemset X is a frequent itemset, then $E[s_{true}(X)] \geq \sigma wD_i$. Assuming that the current batch i data stream is being processed, the expected attenuation support of itemset X in batch i data can be redefined as:

$$E_i[s_{DEST}(X)] = E_i[s_{DEST}(X)] * \gamma + E_i[s_{curr}(X)] \quad (13)$$

If the user sends a mining request in the last j batch data stream, and the mining is uncertain about the big data stream:

$$E_j[s_{DEST}(X)] = E_i[s_{DEST}(X)] * \gamma \quad (14)$$

The mined uncertain big data stream is stored in the UIT-lists structure. The construction process of the global list UIT-lists can be divided into the following steps:

- 1 Scan all data in the current execution window, obtain the item name, existing transaction number and probability of data items in each existing transaction, calculate the expected value of data items, and obtain frequent data items.
- 2 The data items are stored in the UIT-lists structure according to the dictionary order. The expected value and information list of the items are also stored in each storage space in the UIT-lists. The information list contains a tuple composed of the existence transaction number of a series of data items and the probability information in the corresponding transaction.
- 3 When the window slides, the oldest batch of transactions are eliminated, and the tuple of the

eliminated transaction number is deleted from the information list of all data items in the UIT-lists structure. Obtain the data information of the newly added transaction, update the information list of the original data item, store the new data item and its information in UIT-lists, and delete the item with empty information list after the update.

- 4 Repeat step (3) to continuously update the global storage list UIT-lists.

2.3 Online feature selection of uncertain big data flow

Because the multi-label data in the mined uncertain big data stream often shows a large number of characteristics, this brings some problems and challenges to the research of fast classification of uncertain big data streams. The high dimension of data means that the information of a single data block is not fully expressed, and the classification accuracy of the trained classifier for other data blocks is reduced. Therefore, it is particularly important to establish an efficient classification model in the high-dimensional uncertain big data environment. At this stage, the work of processing high-dimensional data classification mostly uses feature selection or feature extraction to transform high-dimensional data into separable low-dimensional datasets. Feature extraction is to transform the original feature space into a low-dimensional feature space to achieve the purpose of dimensionality reduction, but the new feature space does not retain the physical meaning of these original features, so it is difficult to explain. On the contrary, feature selection obtains the optimal subset from the original data, which not only realises the dimensionality reduction, but also preserves the intuitive meaning and physical interpretation, so as to achieve better feature space and achieve better classification performance. Therefore, feature selection has become a hot topic in many fields. In order to solve the problem of dimension disaster, this paper uses online feature selection method to filter the uncertain big data flow based on K-nearest neighbour and label correlation, and achieve the effect of reducing dimension by eliminating redundant or irrelevant features. The specific steps are as follows:

- 1 In the simulated uncertain big data flow environment, vectorise each incoming data block D_k to get $D'_k = \{e_1^k, e_2^k, \dots, e_N^k\}$. First, use the cosine function to calculate the label correlation of D'_k 's class label vector $Y(D'_k)$, and get the label correlation matrix L_k of D'_k :

$$L_k = \{lc_{s,t}^k \mid 1 \leq s, t \leq L\} \quad (15)$$
 where, $lc_{s,t}^k$ represents the correlation between dimension s label l_s and dimension t label l_t in L_k .
- 2 Calculate the average label correlation ave_lc^k of L_k , and use the value of ave_lc^k as the filtering threshold of the label pair set to obtain the label relationship pair LP^k of the data block:

$$LP^k = \{ \langle l_s, l_t \rangle \mid lc_{s,t}^k > ave_lc^k, l_s, l_t \in Y(D'_k) \} \quad (16)$$

- 3 Find the K nearest neighbour sets $kNN_{f_r^k}(e_i^k)$ of e_i^k under the r feature f_r^k , and judge the dependency degree of f_r^k on the nearest neighbour set by combining the label correlation matrix L_k and the label space information:

$$dep(kNN_{f_r^k}(e_i^k)) = \frac{\sum LP^k(e_i^k, e_j^k) / |LP^k|}{K} \quad (17)$$

where, $|LP^k|$ represents the total number of label pairs in label pair set LP^k , and $LP^k(e_i^k, e_j^k)$ represents the number of co-occurrences of the J class label and the K class label of the j data e_j^k in the t nearest neighbour $kNN_{f_r^k}(e_i^k)$.

Calculate the dependency of all data under feature f_r^k , in turn, and finally calculate the average dependency $dep_{f_r^k}^{D'_k}$ of the r feature f_r^k , of D'_k feature set X^k . Only when the threshold α condition $dep_{f_r^k}^{D'_k} \geq \alpha$ is met, the feature f_r^k is retained in the flow feature selection process.

Use the above method to perform the same feature extraction operation for each new data block, and obtain the reconstructed feature set M_k after feature selection of data block D'_k from the original feature set X^k . The online feature selection algorithm based on K-nearest neighbour and label correlation not only considers the correlation between feature data, but also introduces label correlation as an important basis for feature selection as far as possible.

2.4 Detection of feature concept drift in uncertain big data flow

In the process of classifying uncertain big data streams, the infinity of data also implies the differences of data content and structure in different time periods, that is, the generation of concept drift. It is defined that the concept drift caused by the change of feature data distribution in the feature space is considered as virtual drift, and the concept drift caused by the change of label data distribution in the label space is considered as actual drift. The generation of these two drifts will cause the loss of accuracy in the classification process of the established integration model. In order to solve this problem, this paper uses cosine similarity method to detect whether the features of uncertain big data flow have conceptual drift.

The occurrence of concept drift is determined by detecting the joint distribution of features or label data in the data block. Cosine Similarity has its own advantages, which can distinguish the difference performance from the perspective of data direction and is not sensitive to absolute values, and can effectively distinguish the correlation between different features and label data. The concept drift detection is performed between adjacent new and old data

blocks and from the perspective of feature distribution and label distribution.

It can be seen from the online feature selection of uncertain big data streams that each new data block is vectorised, label space correlation calculation and stream feature dimensionality reduction. Then for the quantised data block D'_k at time k and the quantised data block D'_{k-1} at time $k-1$, the cosine similarity first determines the distribution difference between L_k and L_{k-1} , and then determines whether the feature drift occurs according to the feature distribution difference between the feature sets M_k and M_{k-1} after dimensionality reduction of D'_k and D'_{k-1} . The specific process is as follows:

When $k = 1$, train data block D'_k at this time as test data TE_k . If $1 < k < n$, first judge the possibility of change of data block D'_k on the label distribution, and then judge the change of feature distribution if there is no drift. That is, first calculate the cosine similarity $\cos(L_{k-1}, L_k)$ between L_k and L_{k-1} , and compare it with the threshold β , when $\cos(L_{k-1}, L_k) \geq \beta$ indicates that there is no conceptual drift in the label distribution of data block D'_k at time k ; Otherwise, it indicates conceptual drift. If the label distribution concept drift does not occur, compare the feature sets M_k and M_{k-1} of data blocks D'_k and D'_{k-1} to see if they are completely consistent. If it is completely consistent, it means that there is no feature drift on the feature distribution in D'_k at this time; Otherwise, it indicates that feature drift has occurred.

2.5 Fast classification of distributed limit learning machines

Distributed limit learning machine is a new type of fast learning algorithm, which is considered as a parametric linear model focusing on the solution of linear system. Compared with traditional fuzzy neural network learning algorithm, distributed limit learning machine has the characteristics of significant efficiency and easy to obtain the global optimal solution, and the operation process of distributed limit learning machine is relatively simple, which can improve the classification speed. The specific process of distributed limit learning machine algorithm can be described as follows:

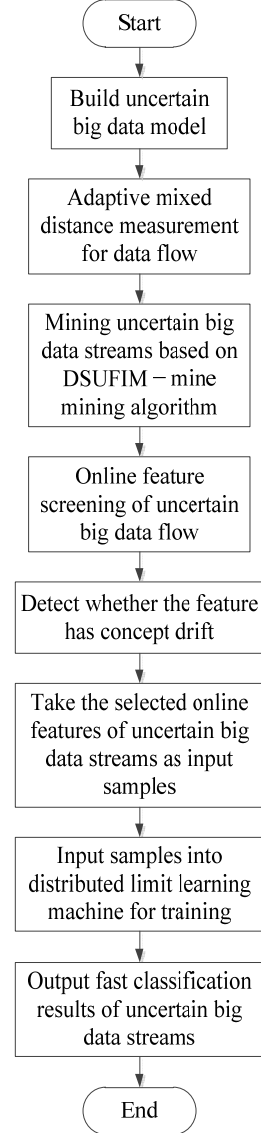
- 1 Random generation of weight w
- 2 Calculate the output matrix H of hidden layer nodes according to the formula
- 3 Calculate the output weight

$$\beta' = H'T \quad (18)$$

- 4 Find the optimal decision function.

According to the above algorithm flow of distributed limit learning machine, this paper uses distributed limit learning machine to quickly classify uncertain big data streams based on the detection results of concept drift of uncertain big data streams.

Figure 1 Specific flow chart of this method



- *Input*: take the selected online feature x_i of uncertain big data flow as the input sample dataset, t_i is the classification mark of this sample i in class m , w_{ij} is the input weight value connecting the sample input x_i and the j neuron, p is the output weight value connecting the i neuron and the output classification mark t_i , L is the number of hidden layer nodes, and b_i is the offset of the i node. The expected output is:

$$T = H\beta \quad (19)$$

- *Training*: when the experimental input data is known, take L hidden layer nodes, and the classification training steps of distributed limit learning machine are as follows:

- 1 The input weight value w_i and the hidden layer node offset value b_i are randomly generated according to the corresponding needs

- 2 From the above formula, the hidden layer input matrix H can be obtained
- 3 Calculate the output weight β according to the known quantity

- *Output*: the final fast classification output of uncertain big data stream based on distributed limit learning machine is:

$$f(x) = h(x)\beta' = h(x)(H'T) \quad (20)$$

where, H' is the generalised inverse of matrix H and $h(x)$ is the activation function.

Based on the above analysis, the specific flow of the fast classification algorithm of uncertain big data flow based on distributed limit learning machine proposed in this paper is shown in Figure 1.

3 Simulation experiment analysis

3.1 Experimental scheme

In order to verify the effectiveness of the fast classification algorithm of uncertain big data streams based on distributed limit learning machine proposed in this paper in practical applications, the classification accuracy and classification time of uncertain big data streams are selected as the experimental indicators, and the classification algorithm based on multi-source migration learning method proposed in Zhou and Liu (2020) and the classification algorithm based on FTRL method proposed in Hu et al. (2020) are used as the comparison method to compare and test with this method.

Table 1 Experimental dataset

Dataset	Number of data	Number of dimensions	Number of clusters
Synthetic 1	150	2	4
Synthetic 2	150	2	4
Synthetic 3	160	2	5
Synthetic 4	180	2	7

3.2 Experimental data

The experiment will be conducted in four artificial datasets, Synthetic 1 to Synthetic 4, including standard datasets and real uncertain datasets, with a total of 640 data. The experimental sample data will be divided into two parts, namely, 440 test sample data and 200 experimental sample data. The test sample data will be input to the simulation software for pre experiment testing, and the optimal operating parameters of the simulation software will be obtained through continuous adjustment, This parameter is used as the initial simulation parameter during the experimental process to ensure the authenticity and reliability of the experimental results. The detailed information of the dataset is shown in Table 1.

3.3 Experimental indicators

- 1 Classification accuracy Acc: the measure of the accuracy rate of all classification, indicating the ratio of the two types of samples correctly detected to all samples. According to the previous analysis, Acc is not a good indicator to measure the performance of classifiers in the classification of uncertain big data flow problems, because it tends to be a large number of categories. The higher the classification accuracy, the better the classification effect. The expression is:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \quad (21)$$

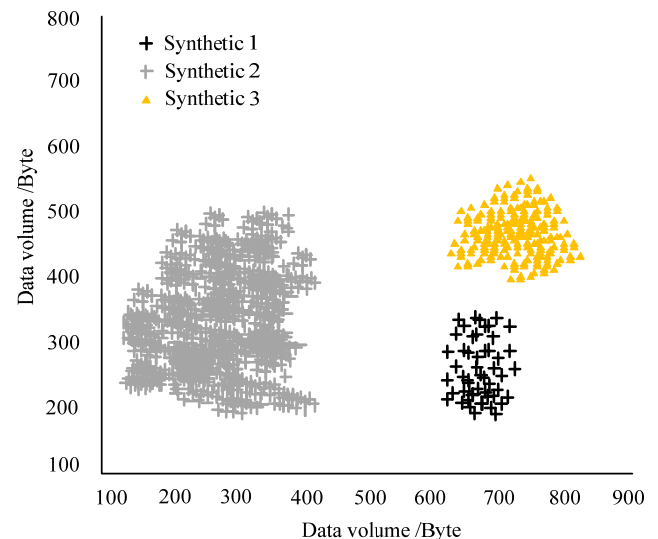
In the formula, TP and TN respectively represent the number of positive and negative samples to be detected correctly, FP represents the number of samples whose real class is marked as negative, but which are detected as positive, and FN is the opposite.

- 2 Classification time: the shorter the classification time, the faster the classification speed and the higher the classification efficiency of the uncertain big data stream.

3.4 Experimental result

The fast classification algorithm of uncertain big data stream based on distributed limit learning machine, algorithm of Zhou and Liu (2020) and Hu et al. (2020) are used to classify the uncertain big data stream and test the classification effect of the three algorithms. The test results are shown in Figures 2 to 4.

Figure 2 Classification results of uncertain big data flow of this algorithm (see online version for colours)



According to the above figure, the fast classification algorithm of uncertain big data streams based on distributed limit learning machine proposed in this paper can accurately classify uncertain big data streams, while the algorithm in Zhou and Liu (2020) and Hu et al. (2020) have poor effect in classifying uncertain big data streams, which shows that

the method in this paper can accurately classify uncertain big data streams.

Figure 3 Uncertain big data stream classification results of reference Zhou and Liu (2020) algorithm (see online version for colours)

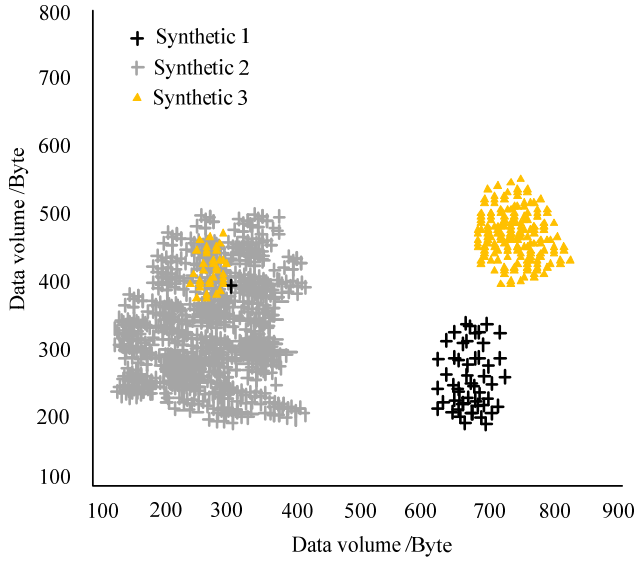
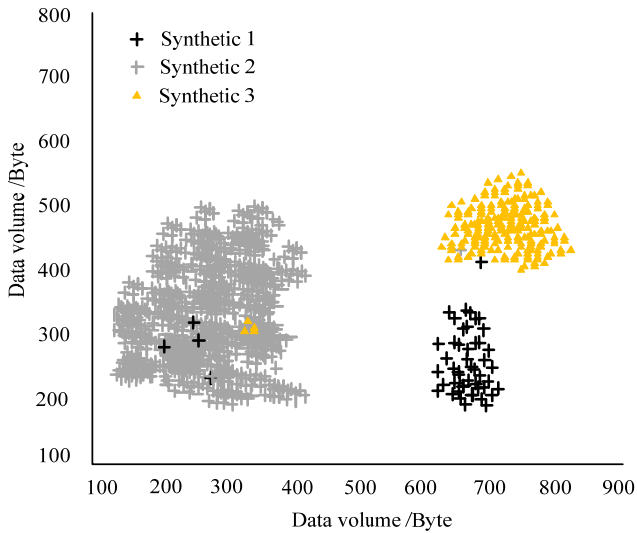


Figure 4 Uncertain big data stream classification results of reference Hu et al. (2020) algorithm (see online version for colours)

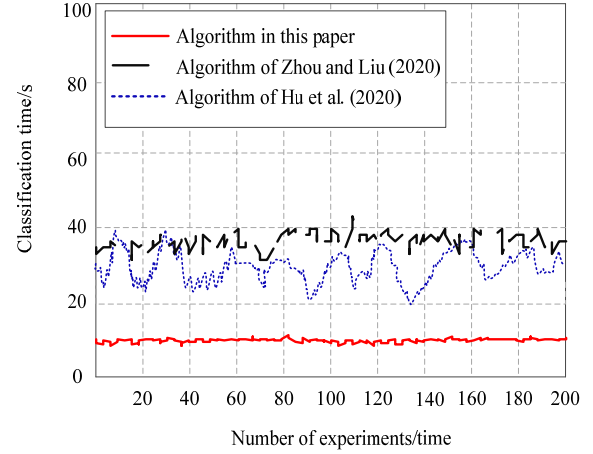


The fast classification algorithm of uncertain big data streams based on distributed limit learning machine, algorithm of Zhou and Liu (2020) and Hu et al. (2020) are used to compare the time used for classification of uncertain big data streams. The comparison results are shown in Figure 5.

According to Figure 5, the proposed fast classification algorithm of uncertain big data streams based on distributed limit learning machine in this paper takes less than 15s to classify uncertain big data streams, and the algorithm in Zhou and Liu (2020) and Hu et al. (2020) take less than 40s to classify uncertain big data streams. The algorithm in this

paper takes the shortest time to classify uncertain big data streams, and the classification speed is the fastest.

Figure 5 Comparison results of uncertain big data stream classification time of three methods (see online version for colours)



4 Conclusions

Due to the slow speed and poor classification effect of traditional methods for uncertain big data streams, this paper proposes a new fast classification algorithm for uncertain big data streams based on distributed limit learning machine. Analyse the causes of uncertain big data, establish an uncertain big data model, measure the adaptive mixing distance of uncertain big data flow in the model through the three steps of kernel distance, Jensen-Shannon divergence and adaptive weight, and mine the uncertain big data flow using the DSUFIM – mine mining algorithm based on the data flow distance measurement results, and adopt the online feature selection method and cosine similarity method, Conduct online feature selection and concept drift detection for uncertain big data streams, and rapidly classify uncertain big data streams through distributed limit learning machines. The experiment shows that the fast classification algorithm of uncertain big data streams based on distributed limit learning machine proposed in this paper can accurately classify uncertain big data streams, and the time for classification of uncertain big data streams is within 15s. The algorithm in this paper has good classification effect and high classification efficiency, and can quickly complete the classification of uncertain big data streams.

Acknowledgements

This work was supported by the programs of Guangdong Provincial Education Science Planning under grant no. 2022GXJK404, Beifang International Education Group under grant no. 20210608004, and Teaching Quality and Reform Project of Guangdong Provincial Undergraduate Colleges under grant no. YJGH [2023]4.

References

- Chen J. (2021) 'Optimisation of massive MIMO data classification algorithm based on fuzzy C-means and differential evolution method', *International Journal of Information and Communication Technology*, Vol. 2021, No. 2, pp.156–167.
- Hu, Y., Hu, X.G. and Li, P.P. (2020) 'Fast short text data stream classification method based on spark', *Computer Engineering and Applications*, Vol. 56, No. 14, pp.138–147.
- Lin, K., Xu, X. and Gao, H. (2021a) 'TSCRNN: a novel classification scheme of encrypted traffic based on flow spatiotemporal features for efficient management of IIoT', *Computer Networks*, Vol. 190, No. 5, pp.107974.1–107974.11.
- Lin, Q.Z., Zhao, Y. and Sun, Z.G. (2021b) 'Automatic classification of large-scale optical fiber network data flow', *Laser Journal*, Vol. 42, No. 9, pp.80–84.
- Liu, W.X., Cai, J. and Wang, Y. (2020) 'Fine-grained flow classification using deep learning for software defined data center networks', *Journal of Network and Computer Applications*, Vol. 41, No. 12, pp.102766–102766.
- Mousavi, S.N., Chen, F. and Abbasi, M. (2022) 'Efficient pipelined flow classification for intelligent data processing in IoT', *Digital Communication and Network: English Version*, Vol. 8, No. 4, pp.561–575.
- Rambabu, K. and Venkatram, N. (2021) 'Ensemble classification using traffic flow metrics to predict distributed denial of service scope in the internet of things (IoT) networks', *Computers and Electrical Engineering*, Vol. 96, No. 3, pp.107444–1–107444-13.
- Shi, L.J. and Huang, D.C. (2020) 'Clustering algorithm for position uncertain data expressed by connection number', *Journal of Chinese Computer Systems*, Vol. 41, No. 2, pp.361–368.
- Tu, L. and Chen, L. (2021) 'Clustering algorithm for uncertain data stream based on damped sliding window', *Application Research of Computers*, Vol. 38, No. 9, pp.2673–2677+2682.
- Wang, J., Jin, C.J. and Yan, D. (2022) 'A frequent pattern mining algorithm based on uncertain data stream', *Journal of Armored Forces*, Vol. 2022, No. 2, pp.78–86.
- Yang, P., Zhu, Y. and Zhang, L. (2020) 'Malware family classification based on attributed dataflow graph', *Journal of Information Security Research*, Vol. 6, No. 3, pp.228–234.
- Zebhi, S., Almodarresi, S. and Abootalebi, V. (2020) 'Converting video classification problem to image classification with global descriptors and pre-trained network', *IET Computer Vision*, Vol. 14, No. 8, pp.614–624.
- Zhang, Y.F., Yu, H. and Wang, G.Y. (2020) 'An active learning method for data stream adaptive classification', *Journal of Nanjing University: Natural Science Edition*, Vol. 56, No. 1, pp.67–73.
- Zhou, S. and Liu, S.M. (2020) 'Research on multi-source transfer learning data streams classification based on dynamic strategy', *Computer Engineering*, Vol. 46, No. 5, pp.139–143,149.
- Zhu, B., Zhong, Y.L. and Wang, X.T. (2020) 'Outlier detection on uncertain data streams', *Journal of Hunan University: Natural Science Edition*, Vol. 47, No. 2, pp.134–140.