IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 3 October 2023, accepted 23 October 2023, date of publication 25 October 2023, date of current version 14 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3327637

RESEARCH ARTICLE

Streaming Data Classification Based on Hierarchical Concept Drift and Online Ensemble

NING LIU[1,2] **AND JIANHUA ZHAO**[2,3]

[1]Faculty of Economics and Management, Shangluo University, Shangluo 726000, China
[2]Engineering Research Center of Qinling Health Welfare Big Data, Universities of Shaanxi Province, Shangluo 726000, China
[3]School of Mathematics and Computer Application, Shangluo University, Shangluo 726000, China

Corresponding author: Jianhua Zhao (zhaojh2009@aliyun.com)

This work was supported in part by the Natural Science Basic Research Program of Shaanxi under Grant 2022JM-339, in part by the Science and Technology Planning Project of Shangluo City under Grant 2022-J-0009, in part by the Natural Science Research Project of Shangluo University under Grant 22KYZX05, and in part by the Shangluo Universities Key Disciplines Project (Discipline Name: Mathematics).

ABSTRACT In order to improve the performance of online learning in the real-time distribution of streaming data, a streaming data classification algorithm based on hierarchical concept drift and online ensemble(SCHCDOE) is proposed in this paper. The concept drift index is calculated based on the newly arrived data instance, and the streaming data is divided into three states: stable state, concept drift warning state, and concept drift occurrence state. When the streaming data is in a stable state, the classifier is not updated. When the streaming data is in a concept drift warning state, online ensemble learning is achieved through random subspaces method to perform feature selection and efficiently update the classifier. When the streaming data is in a concept drift occurrence state, anomaly detection mechanism is used to eliminate abnormal data, and online ensemble learning method and incremental learning method are combined for learning. Local information and global distribution information of the streaming data are fully utilized to train the model, so that the learning model can respond quickly after concept drift occurs. Experiments are conducted on both synthetic and real datasets, and the experimental results show that the proposed algorithm performs well. Compared with other classic algorithms, classification accuracy and concept drift adaptability of the proposed algorithm are improved.

INDEX TERMS Conceptual drift, streaming data, online learning, incremental learning, ensemble learning, classification.

I. INTRODUCTION

In the era of big data, dynamic data is constantly emerging in various application fields, such as transportation data, web page clicks, and stock forecasting. Compared with traditional static data, these dynamic data have the characteristics of high speed, real-time, variability, and unpredictability, so they are referred to as streaming data [1], [2]. Unlike traditional static data, streaming data has characteristics such as dynamism, timing, infinity, and non reproducibility. Traditional machine learning algorithms and theories rely on the assumption of uniform distribution of data, making it difficult to apply to streaming data problems with constantly changing distributions. It brings serious challenges to the collection, storage, analysis, processing of data, as well as model construction and algorithm design for mining tasks [3], [4].

In streaming data, the instability and dynamic changes of data distribution can lead to changes in the data distribution and implicit target concepts with changes in environmental factors, namely concept drift [5], [6]. Concept drift is an important feature of streaming data mining in the real world, and it is also an inevitable difficult problem in streaming data analysis and mining. Concept drift breaks the assumption of fixed data distribution in traditional machine learning, and its typical feature is the inconsistency between real-time data

The associate editor coordinating the review of this manuscript and approving it for publication was Sotirios Goudos.

distribution and training data distribution, which has received increasing attention and research [7], [8].

For the common concept drift problem in streaming data mining, common processing strategies mainly include instance selection-based method and ensemble learning-based method. The instance selection-based method is usually implemented using sliding window technology, which stores data using one or more sliding windows and continuously sliding the windows forward to determine whether concept drift has occurred. The purpose of this method is to select the latest sample for model training and updating, to ensure that the data in the current window can follow the latest distribution [9], [11]. This method improves the real-time performance of online learning models to some extent by introducing a concept drift detection mechanism, however adjusting the sliding window size is a difficult problem, which to some extent affects the model performance.

The ensemble learning-based method can combine the temporal characteristics of streaming data to construct multiple differential base learners after concept drift occurs. This method integrates multiple weak classifiers into a strong ensemble model through a combination strategy, while flexibly updating the base classifier, effectively improving the generalization performance of the model. Therefore, using ensemble learning to handle concept drift is an effective approach for streaming data mining in unstable environments [12], [13], [14].

The ensemble-based learning method can be further divided into data block-based ensemble [15], [16] and online ensemble [17], [18], [19]. Online ensemble is an ensemble learning method that processes samples one by one. Compared with block-based ensemble method, online ensemble can effectively improve the real-time performance of the model [20], [21], [22].

However, common online ensemble learning methods have the following drawbacks: (1) The learning performance of the model is not high. Due to the limited amount of new distribution data obtained at the beginning of drift, the ensemble model retains a large number of base classifiers carrying old distribution data information. Most base classifiers have poor performance and cannot achieve a ‘‘good but different’’ ensemble effect, resulting in poor learning performance of the ensemble model. (2) The generalization performance of the model is poor. Most online ensemble learning methods cannot respond in a timely manner to new data distributions after drift occurs, resulting in online learning models not being able to quickly converge to new distributions after drift occurs, and model generalization performance is poor. (3) The efficiency of model learning is not high. Although online ensemble can effectively improve the real-time performance of the model, its learning efficiency is low due to the need to process samples one by one. At the same time, all features of the data in online ensemble are used for ensemble training, which also means high computational costs.

In order to improve the performance of online learning model, this paper proposes a streaming data classification algorithm based on hierarchical concept drift and online ensemble (SCHCDOE). The concept drift index is calculated based on the newly arrived data instance, and the streaming data is divided into three states: stable state, concept drift warning state and concept drift occurrence state. When the streaming data is in a stable state, the classifier is not updated. When the streaming data is in a concept drift warning state, online ensemble learning is achieved through random subspaces method to perform feature selection and efficiently update the classifier. When the streaming data is in a concept drift state, anomaly detection mechanism is used to eliminate abnormal data, and online ensemble method and incremental learning method are combined for learning. Local information and global distribution information of the streaming data are fully utilized to train the model, so that the learning model can respond quickly after concept drift occurs. Finally, the performance of the proposed algorithm is verified on both synthetic and real datasets.

The main contributions of this work are listed as follows:

(1)A conceptual drift detection method based on sliding window and drift level is proposed. Based on the relationship between the correct probability and minimum standard deviation predicted by the classifier on the sliding window of newly arrived data instances, the concept drift index is calculated, and the streaming data is divided into three states: stable state, concept drift warning state, and concept drift occurrence state. Each state is treated differently.

(2) A streaming data learning method based on concept drift classification and online ensemble is proposed. When the streaming data is in a stable state, the classifier is not updated. When the streaming data is in a concept drift warning state, an online ensemble method based on random subspace method is used for learning to efficiently achieve rapid classifier updating. When the streaming data is in a concept drift state, anomaly detection mechanism is used to eliminate abnormal data, and online ensemble method and incremental learning method are combined for learning.

The rest of this paper is arranged as follows. The section II reviews recent literature, the section III provides a detailed introduction to the methods used in this study, the section IV presents experimental results, discusses. In section V, the conclusion is summarized and the future work direction is introduced.

II. RELATED WORK

A. CONCEPT DRIFT

The concept drift in streaming data makes it difficult for learning models trained from historical data to adapt to new data after distribution changes. For example, in the field of meteorological prediction, weather conditions may be influenced by factors such as temperature, air humidity, and pressure. The changes in these factors may lead to different weather conditions. If real-time changes in meteorological factors cannot be detected, it is impossible to accurately predict changes in weather conditions. Therefore, in streaming data mining with concept drift, it is necessary to break the traditional assumption of fixed data distribution in machine learning, which is of great significance for improving the adaptive performance of online learning models [7], [8].

Different literature categorizes concept drift into different categories, while Chen Zhiqiang et al. [22] classified concept drift into four types: mutation, gradual change, increment, and repetition. De Barros and Santos [23] classified concept drift into three types: mutation type, gradient type, and repetition type. However, based on the time of its change, it mainly consists of two situations: sudden changes in data distribution and changes that persist for a period of time.

The research on concept drift is mainly divided into two key research directions: one is to actively detect whether there is concept drift in streaming data, and the other is how to make adaptive adjustments to the model after detecting concept drift. Based on this, algorithms related to concept drift can be divided into active detection method [24], [25] and passive adaptive method [26], [27].

The active detection method [24], [25] judges concept drift by detecting whether the data distribution has changed and whether the performance of the classifier has decreased. It mainly targets rapid distribution changes, with a focus on quickly detecting the occurrence and specific time points of concept drift. It is mainly divided into model performance-based, window mechanism-based, and hypothesis testing algorithm. Although the active detection method can avoid unnecessary detection in non-stationary streaming data and improve the efficiency of the algorithm, there may be errors, missed detection, and delayed detection of concept drift points during the learning process, which will lead to a decrease in the generalization performance of online learning models.

The passive adaptive method [26], [27] does not require the introduction of a drift detection mechanism to determine the occurrence of concept drift, but rather adapts to changes in data distribution by continuously adjusting the learner. In passive adaptive method, ensemble learning is a common processing method, which can be divided into data block-based online ensemble and single data sample-based online ensemble based on the size of the learning unit. The data block-based online ensemble performs batch processing on the data each time. Although the data block-based online ensemble method can greatly improve the predictive performance of the classifier, when concept drift occurs in the data blocks, the method cannot respond quickly, resulting in slower convergence speed of the model. Although single data sample-based online ensemble method has improved the model's response speed to concept drift to a certain extent, it is difficult to extract important historical information.

B. ENSEMBLE LEARNING

After concept drift occurs, the ensemble learning strategy can flexibly update the base classifier and effectively improve the generalization performance of the model. Therefore, ensemble learning is an effective way to solve concept drift. Ensemble learning methods can be divided into block-based ensemble and online ensemble.

The block-based ensemble divides the streaming data into fixed size data blocks for processing. The most common method is to construct a limited number of base classifiers, and replace the poorly performing classifiers in the ensemble classifier with the classifiers created on the latest data block according to certain rules. Typical methods include: 1)An ensemble classification method based on traditional streaming data [11]. It constructed the ensemble model by training base classifiers on the data block, and used the model built on the latest data block to replace the ensemble model with the worst performing base classifier based on certain heuristic rules. 2)A method based on dynamically adjusting the weight of the base classifier [12], [13], [14]. It adapted to concept drift by continuously adjusting the weight of the base classifier. 3)The online adaptive method based on selective ensemble [15] and the ensemble learning method based on transfer [16]. They improved the effectiveness of the base learner through selective ensemble and transfer learning techniques. Although the block-based ensemble method can greatly improve the overall prediction performance of the model, after concept drift occurs, more outdated base classifiers will reduce the effectiveness of the model.

Online ensemble is an ensemble method that processes samples one by one. Compared with block-based ensemble method, online ensemble can effectively improve the real-time performance of the model. Typical methods include: 1) Concept drift adaptation method based on dynamic weighted voting [17], [18]. It initialized weights based on the accuracy of prediction for new samples and updated weights based on global and local predictions to dynamically update the base classifier. 2) The single sample incremental model [19]. It first initialized a set of base classifiers, updated the ensemble model based on the individual samples arrived at each timestamp, and performed a weighted combination of classifiers. 3)The online learning method based on hybrid labeling strategy [20]. It integrated a fixed basis classifier and a dynamic basis classifier to adapt to concept drift. 4)The online learning method based on serial cross hybrid ensemble [21]. It achieved concept drift detection and convergence through a hybrid ensemble of serial based classifiers and cross based classifiers. 5)The Leveraging bagging method improved based on the online bagging method [22]. It combined the simplicity of bagging with adding more randomness to the input and output of the classifier, sequentially improving the performance of the classifier.

III. OUR METHOD

A. ALGORITHM FRAMEWORK

The flowchart of the algorithm SCHCDOE proposed in this paper is shown in Figure 1. Based on the relationship between the correct probability and the minimum standard deviation

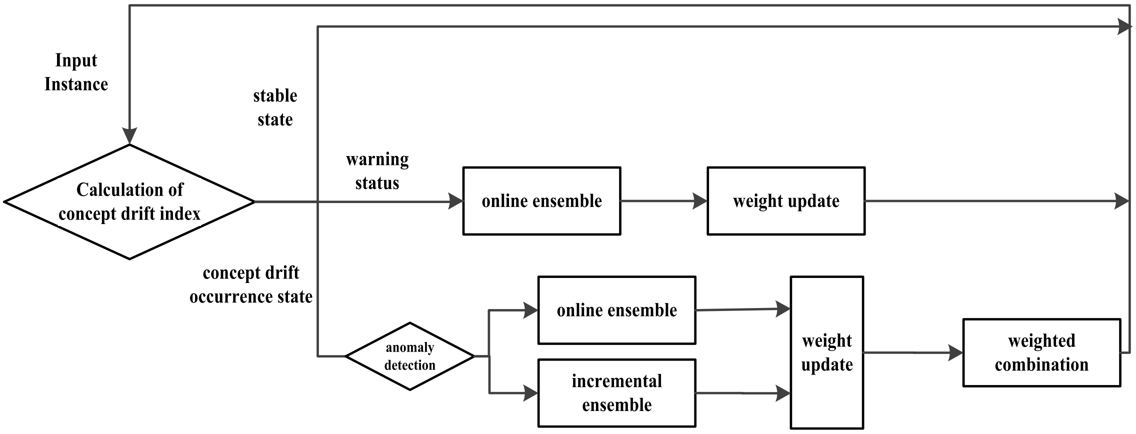


FIGURE 1. The flowchart of the proposed algorithm.

predicted by the classifier on the sliding window of the newly arrived data instance, the concept drift index is calculated, and the streaming data is divided into three states: stable state, concept drift warning state, and concept drift occurrence state. According to different states, the following three operations are performed separately: (1) When the data flow is in a stable state, the classifier is not updated; (2) When the streaming data is in a concept drift warning state, learning is carried out through online ensemble method to perform feature selection and efficiently achieve rapid classifier updating; (3) When the streaming data is in a concept drift state, anomaly detection mechanism is used to detect abnormal data. If it is determined to be an outlier, the instance will be skipped directly and the classifier will not be updated online. If it is determined that it is not an outlier, the online ensemble method and incremental learning method are combined to learn and update the classifier. The combination of online ensemble method and incremental learning method fully utilizes local information and global distribution information of streaming data to train the model.

B. THE IMPLEMENTATION PROCESS OF THE ALGORITHM

1) CONCEPT DRIFT DETECTION

Based on the relationship between the correct probability and minimum standard deviation predicted by the classifier on the sliding window of newly arrived data instances, the concept drift index is calculated, and the steaming data is divided into three states: stable state, concept drift warning state, and concept drift occurrence state.

DDM (Drift Detection Algorithm) is a widely used conceptual drift detection strategy. This algorithm was proposed by Gama et al [28]. When the streaming data is stable, the probability of the classifier predicting correctly for newly arrived data instances follows the Bernoulli distribution [28]. The maximum cumulative classification accuracy $P_{\max}$ denotes the probability of the classifier's prediction being correct. $S_{\min}$ denotes the minimum standard deviation of the classifier and $S_{\min} = \sqrt{P_{\max} \times (1 - P_{\max})}$.$P_w$ and S_w denotes the prediction accuracy and standard deviation of the currently saved sliding window respectively. Based on these values, the current streaming data is divided into three states, as shown in formula (1).

$$\begin{cases} P_w - S_s \geq P_{\max} - S_{\min}, & \text{stable state} \\ P_w - S_s < P_{\max} - S_{\min}, & \text{concept drift warning} \\ P_w - S_s < P_{\max} - 2S_{\min}, & \text{concept drift occurrence state.} \end{cases} \tag{1}$$

According to formula (1), the concept drift index I is used to describe the degree of concept drift in the current streaming data, and the calculation formula is shown in (2).

$$I = (P_w - S_w - P_{\max})/S_{\min}. \tag{2}$$

$$\begin{cases} I \geq 1, \text{Stable without classifier updates} \\ I \leftarrow 1, \text{Warning and online updates} \\ I \leftarrow 2, \text{Concepte drift occurs.} \end{cases} \tag{3}$$

According to the different values of the concept drift index, the steaming data is divided into three states. As shown in formula (3), when I>=1, it is judged that the streaming data is in a stable state and the model is not updated; When I←1, a concept drift warning is issued and the model is updated online. When I←2, it is determined that the steaming data has undergone concept drift, abnormal points are detected, and the model is updated online.

2) ABNORMAL POINT DETECTION

The streaming data often contain noise. If noise is not effectively distinguished and each newly arrived instance is processed using the same method, the classification decision boundary is easily damaged by noise (outliers). Here, an anomaly detection mechanism is used to detect data instances. If it is determined to be an outlier, the instance will be skipped directly and the classifier will not be updated online. If it is determined that it is not an outlier, the classifier is updated online. The anomaly detection mechanism can avoid excessive influence of decision boundaries by outliers, thereby improving the overall noise resistance of the model.

The anomaly detection method is to save N latest classification models. During anomaly detection, N classifiers are used to predict the input values, obtain the predicted values $\hat{y}_t$ corresponding to N models, and calculate the corresponding mean value and variance value.

$$0.5 \in [\hat{y}_{t\ mean} - 3\hat{y}_{t\ \mathrm{std}}, \hat{y}_{t\ mean} + 3\hat{y}_{t\ \mathrm{std}}], \tag{4}$$

where $\hat{y}_{t\ mean}$ denotes the mean of the predicted values of N classifiers, $\hat{y}_{t\ \mathrm{std}}$ denotes the variance value of N classifier predictions.

If formula (4) holds, it is considered that the current instance is not far from the classification decision boundary, and it is judged that it is not an outlier. Otherwise, it is considered that the current instance is far from the classification decision boundary, and it is judged as an outlier.

3) ONLINE ENSEMBLE METHOD

In the online ensemble process, on the one hand, a random subspace ensemble method is used to construct a base classifier by randomly selecting several subspaces from the original feature space. Multiple base learners are integrated to obtain the final result, increasing the diversity of classifiers and improving computational efficiency. On the one hand, local predictions are made through online ensemble strategies, and the weights of the base learners are updated to make the online ensemble model adapt to rapid changes in streaming data and improve learning efficiency.

Randomly sized feature subspaces can enhance the diversity of classifiers and improve ensemble learning performance. At the same time, it can reduce the impact of noise, making the model adapt faster to local concept drift subspaces that only affect certain features. The calculation of the number of features r is shown in formula (5).

$$r = \mu \times f + \frac{(1-\mu) \times f \times N(0,1)}{2}, \tag{5}$$

where $\mu \in [0, 1], f$ denotes the total number of features.

In streaming data learning problems, weight updating need to be completed online, meaning that the algorithm should adjust the model weight based on each round of data feedback. Here, the multiplicative weight update (MWU) algorithm is used for adaptive weight adjustment, as shown in formula (6). The weights of each model are adjusted in each round. If the model performs well on the current data (*i.e.* the predicted loss is small), the weight will be increased on the original basis. Otherwise, it will decrease. Ultimately, all weights will be normalized to the K models, making them legal probability distributions.

$$\omega_{t+1}^{k} = \frac{\omega_t^k \exp(-\eta \ell(T_k(x_t), y_t))}{\sum_{k=1}^{K} \omega_t^k \exp(-\eta \ell(T_k(x_t), y_t))}, \tag{6}$$

where x_t denotes the input value at time t, $T_k(x_t)$ denotes the predicted value of x_t, y_t denotes the actual value of x_t, k denotes the number of classifiers, $\ell(T_k(x_t), y_t)$ denotes the loss value, η denotes the step size of the adaptive weight adjustment algorithm.

4) LEARNING METHOD COMBINING ONLINE ENSEMBLE AND INCREMENTAL LEARNING

After concept drift occurs in streaming data, the predictive performance of most historical base learners in online ensemble learning will remain low in a relatively short period of time, and the corresponding weights of the base learners will experience an exponential decline. The online ensemble learning model cannot cover the distribution information of the entire streaming data, resulting in poor robustness of the base learners.

Therefore, after concept drift occurs, this paper adopts a learning method that combines online ensemble learning and incremental learning. By using online ensemble learning to locally predict new samples, the model responds promptly to concept drift. Meanwhile, incremental learning is used for global prediction. By utilizing an appropriate amount of key samples within the historical data block and newly arrived samples within the data block, key historical information and information on the latest data distribution are extracted separately to update the incremental learner and quickly adapt to concept drift. Combining the online ensemble learning and incremental learning mentioned above, a total testing model is formed, and the samples to be tested are weighted for voting, as shown in formula (7).

$$y = \sum_{i=1}^{k} w_i \cdot h_i(x) + \lambda \cdot h_s(x), \tag{7}$$

where $h_i(x)$ denotes an ensemble classifier, $h_s(x)$ denotes incremental classifier, λ denotes weight coefficient.

C. PSEUDO-CODE FOR THE PROPOSED ALGORITHM

The algorithm description of SCHCDOE is shown in Algorithm 1.

IV. EXPERIMENT

In order to verify the effectiveness of the proposed algorithm, two different types of concept drift datasets, namely synthetic and real datasets for experiments shown in Table 1, are used to conduct experiments. The experimental results will be analyzed and explained from different evaluation indicators.

A. EXPERIMENTAL DATA

1) SYNTHESIZED DATASET

The synthesized dataset mainly includes the following five types of datasets.

(1)SEAs, $\mathrm{SEA_G}$ and SEAm dataset. They are a set of streaming data generated using the SEA [11] generator. By changing the threshold, concept drift is simulated. The dataset contains three attributes, of which only two are related. Each dataset contains 20000 instances and adds 5% noise. In addition, all three datasets contain two concept drifts. Among them, the $\mathrm{SEA_s}$ dataset contains two mutation type concept drifts, the $\mathrm{SEA_G}$ dataset contains two gradual

Algorithm 1 SCHCDOE Algorithm

```
Input:
  SD: Streaming data D
Output:
  Accuracy, Precision, Recall, and F1-Score
Process:
  Step 1. The training set is learned to form multiple classifiers.
  Step 2. for $x$ in SD.
  Step 3. The concept drift index $I$ is calculated according to the formula (1)-(2).
  Step 4.  if I ≧ 1, the classifier does not update, goto Step2.
  Step 5.  end if
  Step 6.  if I←1, feature selection is performed according to formula (5), online learning and ensemble learning are performed according to formulas (6), and the classifier is updated.
  Step 7.  end if
  Step 8.  if I←2, determine if $x$ is an outlier according to formulas (4).
  Step 9.  if $x$ is an outlier, remove it and goto step 2.
  Step 10. end if
  Step 11. if $x$ is not an outlier, perform online ensemble and incremental learning according to formulas (7).
  Step 12.  Accuracy, Precision, Recall, and F1-Score are calculated according to formula (8)-(11).
  Step 13.  end if
  Step 14. end if
  Step 15. end for
```

TABLE 1. The experimental datasets.

Dataset	number	Attribute	Classes	noise	type
SEAs	20000	3	2	5%	Abrupt
SEA_G	20000	3	2	5%	gradual
SEA_m	20000	3	2	5%	Mixed
Mixed paper	20000	4	2	5%	Mixed
Hyperplane	100000	10	2	5%	Incremental
electricity	45312	6	2		Unknown
Weather	18159	8	2		Unknown

conceptual drifts, the SEA_m dataset contains one mutation type concept drift and one gradient type concept drift.

(2)Mixed paper dataset. Including mutation class and gradient class concept drift, four features, and 20000 instances and adds 5% noise.

(3)Rotating Hyperplane dataset [29]. Including the characteristic of conceptual gradient, the hyperplane in d-dimensional space continuously changes in position and direction, with $d = 10$, two categories, 100 000 instances and adds 5% noise.

2) REAL DATASET

The real dataset mainly includes the following two types of datasets.

(1)electricity dataset. Including six attributes and 45312 instances. It is a real-world dataset widely used in streaming data learning. This dataset is part of the data from the New South Wales electricity market in Australia from 1995 to 1998. As electricity prices in that area not fixed but vary based on supply and demand, the goal is to predict daily changes in electricity prices.

(2)Weather dataset. Including 18159 instances and 8 related attributes. The content of the data is weather information collected in Bellevue, Nebraska from 1949 to 1999, with the aim of predicting whether it will rain on a given date.

To solve overfitting problems, two methods are used in this paper. One is K-fold cross validation, the other one is adding noise to dataset.

(1)K-fold cross validation is used to solve overfitting problems.

The dataset is divided into k subsets, one subset is selected as the validation set each time, and the remaining k-1 subset is used as the training set. K cross validation tests are repeated, and the final performance indicator is obtained by averaging of the k validation results. Multiple cross validation can reduce the impact of randomness on model performance, making the evaluation results more reliable and stable. The use of k-fold cross validation can effectively reduce the risk of overfitting training data in the model, while also improving the model's generalization ability and robustness, thereby achieving better performance and predictive ability.

(2)Adding noise to dataset is used to solve overfitting problems.

Adding noise to the input dataset can stabilize the model without affecting data quality and privacy, while adding noise to the output can make the data more diverse. Noise addition should be carried out within a certain range to avoid incorrect data or significant differences.

B. EXPERIMENTAL SETUP

The experimental platform is Windows 10, with Intel i7-2.5GHz CPU and 32GB memory. All classification algorithms are implemented based on Python language.

The parameter values of the algorithm SCHCDOE proposed in the paper are shown in Table 2.

C. EVALUATING INDICATOR

In order to evaluate the performance of the proposed algorithm, Accuracy, Precision, Recall, F1-Score, AUC value and Robustness are selected as evaluation criteria for algorithm comparison. The higher the values, the better the performance of the algorithm.

The evaluation indicators of the data are based on the confusion matrix shown in Table 3. The descriptions of TP, TN, FN, and FP are as follows:

●True positive example (TP): Predicted as a positive sample, in fact, it is also a positive sample.

●True negative example (TN): Predicted as a negative example sample, but actually also a negative example sample.

TABLE 2. The parameter values of the proposed algorithm.

Serial Number	The name of parameters	Parameter value	Functional description of parameters
1	N	6	The num of latest classification models in formula (4).
2	μ	0.58	Parameters in randomly feature subspaces in formula (5).
3	k	10	The number of classifiers during adaptive weight adjustment in formula (6).
4	η	0.145	The step size of the adaptive weight adjustment algorithm in formula (6).
5	λ	0.725	The weight coefficient in formula (7).

TABLE 3. Confusion matrix.

category		Predicted value	
		Positive Example	Negative Example
True value	Positive example	TP	FN
	Negative example	FP	TN

•False Negative Example (FN): Predicted as a negative sample, but actually a positive sample.

•False positive sample (FP): Predicted as a positive sample, but actually a negative sample.

Accuracy, Precision, Recall, F1-Score, ROC curve, AUC value and Robustness are defined as follow:

(1)Accuracy: Reflect the classifier's ability to judge the entire sample, and determine positive as positive and negative as negative. The accuracy calculation formula is shown in (8):

$$Accuracy = \frac{TN + TP}{TP + TN + FN + FP} \tag{8}$$

(2)Precision: It refers to how many samples predicted to be true are indeed true. The calculation method is shown in Formula(9):

$$\Pr ecision = \frac{TP}{TP + FP} \tag{9}$$

(3)Recall: It represents the proportion of predicted positive samples in the actual positive samples. The calculation method is shown in formula(10):

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

(4)F1-Score: It represents the harmonic average of accuracy and recall, which is closer to the smaller value of accuracy and recall. The calculation method is shown in formula(11):

$$F1_{Score} = 2 * \frac{Recall * \Pr ecision}{Recall + \Pr ecision} \tag{11}$$

(5)TPR (**True Positive Rate):** It represents the proportion of actual positive instances to all positive instances in the positive instances predicted by the classifier. The calculation method is shown in formula(12):

$$TPR = \frac{TP}{TP + FN} \tag{12}$$

(6)FPR(False Positive Rate): It represents the proportion of actual negative instances to all negative instances in the positive class predicted by the classifier. The calculation method is shown in formula(13):

$$FPR = \frac{FP}{TN + FP} \tag{13}$$

(7)ROC curve: The ROC curve is plotted by TPR and FPR, where TPR is the y-axis and FPR is the x-axis.

(8)AUC value: The AUC value is the area covered by the ROC curve. Obviously, the larger the AUC, the better the classifier's classification performance.

(9)Robustness: Robustness is an effective evaluation indicator for the stability performance of a model, and it also reflects the generalization performance of the model. Here the robustness of different algorithms are analyzed in terms of average accuracy,and the robustness of algorithm A on different datasets is defined as:

$$R_A(D) = \frac{Accuracy_A(D)}{\min_a Accuracy_a(D)}, \tag{14}$$

where $Accuracy_A(D)$ denotes the average accuracy of algorithm A on dataset D, $Accuracy_a(D)$ denotes the minimum average accuracy value among all algorithms on dataset D. The overall robustness value of an algorithm is the sum of its robustness values on all datasets. Assuming there are n datasets, specifically defined as:

$$R_A = \sum_{i=1}^{n} R_A(D_i). \tag{15}$$

D. COMPARATIVE MODEL

There are three classic algorithms to compare with the algorithm proposed in this paper.

(1)SEA [11]: It is a traditional streaming ensemble classification algorithm. This algorithm classifies all data and establishes a single decision tree on all data, requiring approximately constant memory and quickly adjusting based on concept drift.

(2)AC_OE [30]: It is an adaptive classification method for concept drift based on online ensemble. Local prediction of streaming data is performed through online ensemble learning to dynamically adjust the weights of learners. The global distribution information of streaming data is performed through incremental learning strategy.

(3)KSHPR [31]: It is an online ensemble adaptive algorithm. Using a strategy combining non parametric testing and sliding window for concept drift detection, an ensemble learning model of four base learners is established, and weights are dynamically assigned based on the prediction accuracy of the base learners.

TABLE 4. Comparison of experimental results.

Dataset	Alogrithm	Types of Evaluation Indicators			
		Accuracy	Precision	Recall	F1-Score
SEAs	SCHCDOE	**0.9387**	**0.9274**	**0.9328**	**0.9332**
	SEA	0.8854	0.8698	0.8564	0.8617
	AC_OE	0.9175	0.9047	0.9064	0.9094
	KSHPR	0.9074	0.8974	0.9017	0.9042
SEA$_G$	SCHCDOE	**0.9287**	**0.9211**	**0.9208**	**0.9249**
	SEA	0.8814	0.8481	0.8757	0.8778
	AC_OE	0.9148	0.8947	0.8947	0.9012
	KSHPR	0.9002	0.9122	0.9141	0.9095
SEA$_m$	SCHCDOE	**0.9321**	**0.9282**	**0.9360**	**0.9348**
	SEA	0.8874	0.8717	0.8647	0.8712
	AC_OE	0.9119	0.9002	0.8998	0.9046
	KSHPR	0.8976	0.9046	0.9096	0.9015
Mixed paper	SCHCDOE	**0.9287**	**0.9274**	**0.9224**	**0.9266**
	SEA	0.8579	0.8478	0.8487	0.8501
	AC_OE	0.8825	0.8910	0.8941	0.8867
	KSHPR	0.9087	0.8975	0.9124	0.9105
Hyperplane	SCHCDOE	**0.9012**	**0.9059**	**0.9141**	**0.9084**
	SEA	0.8774	0.8756	0.8802	0.8787
	AC_OE	0.8892	0.8812	0.8754	0.8725
	KSHPR	0.8464	0.8558	0.8501	0.8468
electricity	SCHCDOE	**0.8836**	**0.8787**	**0.8798**	**0.8802**
	SEA	0.6455	0.6441	0.6420	0.6414
	AC_OE	0.8214	0.8202	0.8124	0.8137
	KSHPR	0.8821	0.8745	0.8838	0.8787
Weather	SCHCDOE	**0.8954**	**0.8974**	**0.8810**	**0.8852**
	SEA	0.8501	0.8574	0.8610	0.8542
	AC_OE	0.8720	0.8754	0.8734	0.8725
	KSHPR	0.8598	0.8541	0.8474	0.8499

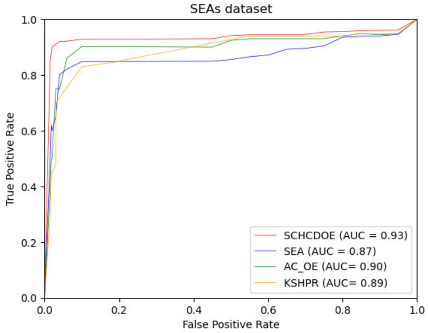


FIGURE 2. ROC curves and AUC values on SEAs dataset.

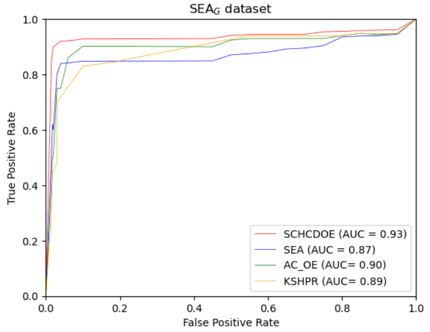


FIGURE 3. ROC curves and AUC values on SEA$_G$ dataset.

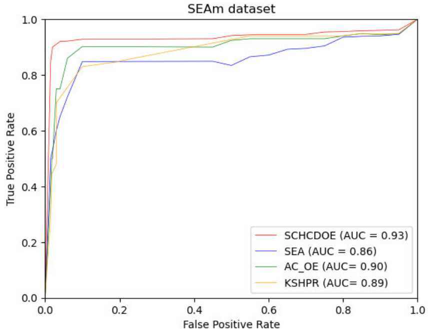


FIGURE 4. ROC curves and AUC values on SEA$_m$ dataset.

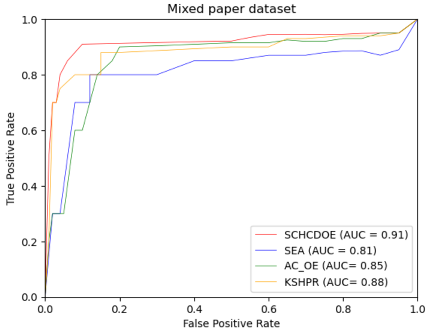


FIGURE 5. ROC curves and AUC values on Mixed paper dataset.

E. EXPERIMENTAL RESULTS AND ANALYSIS

1) COMPARATIVE EXPERIMENTS WITH OTHER ALGORITHMS

On synthesized datasets and real datasets, the algorithm SCHCDOE proposed in this paper is compared with SEA, AC_OE and KSHPR. The experimental results are shown in Table 4 and the results show that our algorithm SCHCDOE has overall improved in terms of Accuracy, Precision, Recall,

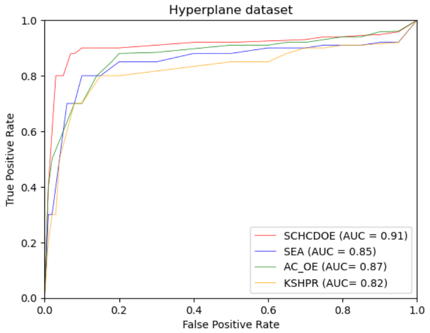


FIGURE 6. ROC curves and AUC values on Hyperplane dataset.

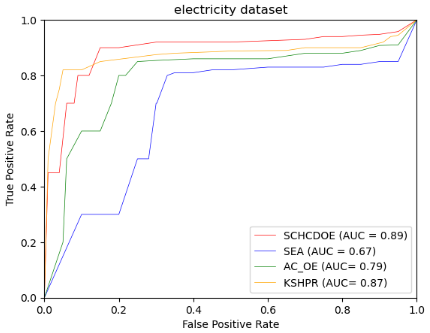


FIGURE 7. ROC curves and AUC values on electricity dataset.

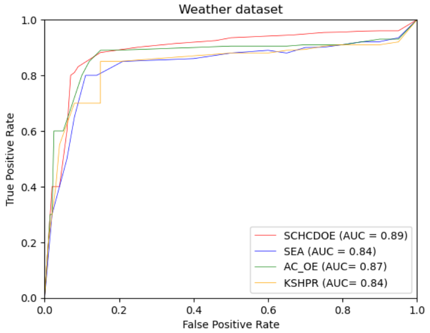


FIGURE 8. ROC curves and AUC values on Weather datase.

and F1 Score. It indicates that SCHCDOE can effectively handle various types of concept drift.

All ROC curves and AUC values are the result of randomized experiments, and the detailed ROC curves are shown from Figure 2 and Figure 8, the results show that the AUC value of our algorithm is higher than other algorithms.

This is because SCHCDOE can effectively detect concept drift. When concept drift occurs, it effectively extracts the latest data distribution information after concept drift, fully leveraging the role of a classifier, enabling the model to quickly converge, and effectively improving the overall performance of the model.

Compared with SEA, SCHCDOE reduces the impact of concept drift by detecting concept drift, using various mechanisms such as outlier detection, online ensemble learning, and incremental learning, effectively improving classification performance after concept drift occurs. Compared with AC_OE, SCHCDOE divides streaming data into three states: stable state, concept drift warning state, and concept drift occurrence state. The use of outlier detection mechanism can avoid excessive influence of decision boundaries by outliers, thereby improving the overall noise resistance of the model. Compared with KSHPR, SCHCDOE combines online ensemble learning and incremental learning methods for learning, fully utilizing local information and global distribution information of streaming data to train the model, enabling the learning model to respond quickly to concept drift while improving convergence and robustness.

2) ABLATION EXPERIMENT

Ablation experiments are conducted to verify the effectiveness of the concept drift detection (CDD), anomaly detection (AD), and random subspace method (RSM) introduced by SCHCDOE. In the ablation experiment, three algorithms SCHCDOE_del_CDD, SCHCDOE_del_AD and SCHCDOE_del_RSM are obtained by omitting the corresponding modules from SCHCDOE, and their performance are compared by experiment.

The specific functions of the three algorithms for ablation experiments are as follows:

SCHCDOE_del_CDD: All data is not subjected to concept drift detection, and is directly integrated through online learning to form a classifier.

SCHCDOE_del_AD: When the streaming data is in a concept drift state, anomaly detection is not performed on the data, and online ensemble learning and incremental learning methods are directly used for learning.

SCHCDOE_del_RSM: Anomaly detection is performed on streaming data, but in the online integration process, all features of the streaming data are integrated and random subspace integration methods are not used.

Table 5 shows the experimental results of the ablation experiment after removing the concept drift detection (CDD) module. After removing the CDD mechanism, the SCHCDOE_del_CDD algorithm performs poorly in online learning, resulting in a significant overall performance loss for the classifier. This is because SCHCDOE_del_CDD does not perform concept drift detection on all data and integrates it directly through online learning to establish a classifier. After concept drift occurs, the learning model trained from historical data is difficult to adapt to the new data after distribution changes. This further validates that concept drift can lead to the failure or decrease in accuracy of the prediction model, and concept drift detection is particularly important.

Table 6 shows the experimental results of the ablation experiment after removing the anomaly detection (AD) module. After removing the AD mechanism, the performance of the SCHCDOE_del_AD algorithm in online learning is not very good. This is because SCHCDOE_del_AD does not

TABLE 5. Comparison of experimental results.

Dataset	Alogrithm	Types of Evaluation Indicators			
		Accuracy	Precision	Recall	F1-Score
SEAs	SCHCDOE	**0.9387**	**0.9274**	**0.9328**	**0.9332**
	SCHCDOE_del_CDD	0.8251	0.8321	0.8254	0.8287
SEA$_G$	SCHCDOE	**0.9287**	**0.9211**	**0.9208**	**0.9249**
	SCHCDOE_del_CDD	0.8454	0.8302	0.8321	0.8451
SEA$_m$	SCHCDOE	**0.9321**	**0.9282**	**0.9360**	**0.9348**
	SCHCDOE_del_CDD	0.8147	0.8174	0.8245	0.8249
Mixed paper	SCHCDOE	**0.9287**	**0.9274**	**0.9224**	**0.9266**
	SCHCDOE_del_CDD	0.8541	0.8374	0.8241	0.8259
Hyperplane	SCHCDOE	**0.9012**	**0.9059**	**0.9141**	**0.9084**
	SCHCDOE_del_CDD	0.8478	0.8358	0.8356	0.8297
electricity	SCHCDOE	**0.8836**	**0.8787**	**0.8798**	**0.8802**
	SCHCDOE_del_CDD	0.6387	0.6454	0.6584	0.6478
Weather	SCHCDOE	**0.8954**	**0.8974**	**0.8810**	**0.8852**
	SCHCDOE_del_CDD	0.8248	0.8354	0.8458	0.8357

TABLE 6. Comparison of experimental results.

Dataset	Alogrithm	Types of Evaluation Indicators			
		Accuracy	Precision	Recall	F1-Score
SEAs	SCHCDOE	**0.9387**	**0.9274**	**0.9328**	**0.9332**
	SCHCDOE_del_AD	0.9120	0.9142	0.9121	0.9148
SEA$_G$	SCHCDOE	**0.9287**	**0.9211**	**0.9208**	**0.9249**
	SCHCDOE_del_AD	0.9101	0.9021	0.8985	0.8954
SEA$_m$	SCHCDOE	**0.9321**	**0.9282**	**0.9360**	**0.9348**
	SCHCDOE_del_AD	0.9101	0.9087	0.8952	0.9121
Mixed paper	SCHCDOE	**0.9287**	**0.9274**	**0.9224**	**0.9266**
	SCHCDOE_del_AD	0.8901	0.8921	0.8998	0.8922
Hyperplane	SCHCDOE	**0.9012**	**0.9059**	**0.9141**	**0.9084**
	SCHCDOE_del_AD	0.8814	0.8800	0.8891	0.8847
electricity	SCHCDOE	**0.8836**	**0.8787**	**0.8798**	**0.8802**
	SCHCDOE_del_AD	0.8320	0.8322	0.8328	0.8244
Weather	SCHCDOE	**0.8954**	**0.8974**	**0.8810**	**0.8852**
	SCHCDOE_del_AD	0.8778	0.8798	0.8782	0.8814

TABLE 7. Comparison of experimental results.

Dataset	Alogrithm	Types of Evaluation Indicators				
		Accuracy	Precision	Recall	F1-Score	Time(s)
SEAs	SCHCDOE	0.9387	0.9274	0.9328	0.9332	**9.8453**
	SCHCDOE_del_RSM	0.9390	0.9380	0.9381	0.9382	13.5414
SEA$_G$	SCHCDOE	0.9287	0.9211	0.9208	0.9249	**9.6344**
	SCHCDOE_del_RSM	0.9291	0.9220	0.9340	0.9300	14.4144
SEA$_m$	SCHCDOE	0.9321	0.9282	0.9360	0.9348	**9.5234**
	SCHCDOE_del_RSM	0.9354	0.9315	0.9351	0.9352	12.5447
Mixed paper	SCHCDOE	**0.9287**	**0.9274**	**0.9224**	**0.9266**	**9.9234**
	SCHCDOE_del_RSM	0.9214	0.9254	0.9212	0.9241	14.5444
Hyperplane	SCHCDOE	0.9012	0.9059	0.9141	0.9084	**11.2342**
	SCHCDOE_del_RSM	0.9123	0.9178	0.9185	0.9146	15.4114
electricity	SCHCDOE	0.8836	0.8787	0.8798	0.8802	**18.2352**
	SCHCDOE_del_RSM	0.8877	0.8854	0.8887	0.8880	28.8414
Weather	SCHCDOE	0.8954	0.8974	0.8810	0.8852	**12.2422**
	SCHCDOE_del_RSM	0.9014	0.8998	0.8941	0.8956	18.8441

perform anomaly detection on the streaming data when it is in a concept drift state, and directly uses online ensemble methods and incremental learning methods for learning. After concept drift occurs, there is often noise in the streaming data. If the noise is not effectively distinguished, but rather every newly arrived instance is treated the same and updated, the classification decision boundary is easily damaged by noise (outliers). This indicates that the anomaly detection mechanism can avoid the influence of too many outliers on the decision boundary, thereby improving the overall noise resistance of the model.

Table 7 shows the experimental results of the ablation experiment after deleting the random subspace method (RSM) module, where the time column represents the learning time. Figure 9 shows the comparison between the two algorithms in terms of learning time.

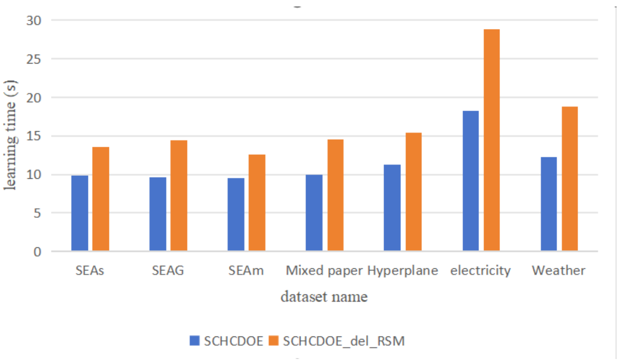


FIGURE 9. Comparison chart of learning time.

From Table 7 and Figure 9, it can be seen that after removing the RSM mechanism, although the online learning performance of the SCHCDOE_del_RSM algorithm has little improvement, the learning time has a significant increase. This is because SCHCDOE_del_RSM adopts online ensemble learning for all data when the streaming data is in a concept drift state, which increases training time.

However, in the process of online ensemble learning, the proposed SCHCDOE adopts a random subspace ensemble method, which randomly selects several subspaces from the original feature space to construct a base classifier, maintaining higher system diversity, increasing classifier diversity, and improving computational efficiency. Simultaneously it reduces the impact of noise to make the model adapt faster to local concept drift subspaces that only affect certain features.

On the other hand, SCHCDOE utilizes online ensemble learning strategies for local prediction and updates the weights of the base learners to adapt the online ensemble model to rapid changes in streaming data and improve learning efficiency. This mechanism improves learning efficiency while maintaining learning accuracy. From Table 7, it can be seen that on some datasets, such as Mixed paper, the accuracy of SCHCDOE is slightly improved compared to SCHCDOE_del_RSM, while greatly shortening the time. On other datasets, although the accuracy of SCHCDOE decreased slightly compared to SCHCDOE_del_RSM, it greatly shorten time and improves learning efficiency.

3) ROBUSTNESS ANALYSIS OF THE MODEL

Robustness is an important indicator of algorithm stability, and a larger value indicates that the model is more stable. Figure 10 shows the robustness of different algorithms on different datasets. The different heights of small rectangles represent the robustness values of the algorithm on different datasets. The values above each column represent the sum of the robustness values of the algorithm on all datasets, which is the overall robustness of the algorithm.

It can be seen that on most datasets, the robustness of the algorithm SCHCDOE proposed in this paper is superior to the other three algorithms, and the overall robustness has achieved the optimal value. There are two reasons: (1) SCHCDOE uses an ensemble learning framework to combine multiple weak classifiers, improving the overall generalization performance of the model. (2) SCHCDOE uses feature selection algorithms to achieve dimensionality reduction, effectively reducing the risk of overfitting the training data, improving the generalization ability and robustness of the model.

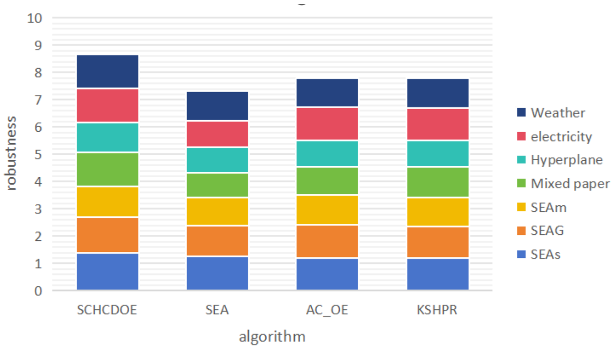


FIGURE 10. Robustness comparison of different algorithms.

V. CONCLUSION

To address the issue of online learning models not being able to respond in a timely manner to changes in distribution after concept drift occurs in streaming data, this paper proposes a streaming data reliability learning model SCHCDOE. The concept drift index is calculated based on the newly arrived data instance, and the streaming data is divided into three states: stable state, concept drift warning state and concept drift occurrence state. When the streaming data is in a concept drift state, anomaly detection mechanism is used to eliminate abnormal data, online ensemble method and incremental learning method are combined for learning. Local information and global distribution information of the streaming data are fully utilized to train the model, so that the learning model can respond quickly after concept drift occurs while improving the robustness.

In the future, further optimization of the algorithm will be considered to expand its applicability while further reducing time costs.

REFERENCES

[1] G. Krempl, I. Žliobaite, D. Brzeziński, E. Hüllermeier, M. Last, V. Lemaire, T. Noack, A. Shaker, S. Sievi, M. Spiliopoulou, and J. Stefanowski, "Open challenges for data stream mining research," *ACM SIGKDD Explor. Newslett.*, vol. 16, no. 1, pp. 1–10, Sep. 2014.

[2] E. Lughofer and M. Pratama, "Online active learning in data stream regression using uncertainty sampling based on evolving generalized fuzzy models," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 1, pp. 292–309, Feb. 2018.

[3] T. Zhai, Y. Gao, and J. W. Zhu, "Survey of online learning algorithms for streaming data classification," (in Chinese), *J. Softw.*, vol. 31, no. 4, pp. 912–931, 2020.

[4] P. Zhao and Z.-H. Zhou, "Learning from distribution-changing data streams via decision tree model reuse," *Scientia Sinica Informationis*, vol. 51, no. 1, pp. 1–12, 2021.

[5] J. Lu, A. Liu, F. Dong, F. Gu, J. Gama, and G. Zhang, ''Learning under concept drift: A review,'' *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 12, pp. 2346–2363, Dec. 2019.

[6] M. Tennant, F. Stahl, O. Rana, and J. B. Gomes, ''Scalable real-time classification of data streams with concept drift,'' *Future Gener. Comput. Syst.*, vol. 75, pp. 187–199, Oct. 2017.

[7] Y. Wen, S. Liu, Y. Miao, X. Yi, and C. Liu, ''Survey on semi-supervised classification of data streams with concept drifts,'' (in Chinese), *Ruan Jian Xue Bao/J. Softw.*, vol. 33, no. 4, pp. 1287–1314, 2022.

[8] J. Tanha, N. Samadi, Y. Abdi, and N. Razzaghi-Asl, ''CPSSDS: Conformal prediction for semi-supervised classification on data streams,'' *Inf. Sci.*, vol. 584, pp. 212–234, Jan. 2022.

[9] L. Du, Q. Song, and X. Jia, ''Detecting concept drift: An information entropy based method using an adaptive sliding window,'' *Intell. Data Anal.*, vol. 18, no. 3, pp. 337–364, Apr. 2014.

[10] A. Pesaranghader and H. L. Viktor, ''Fast Hoeffding drift detection method for evolving data streams,'' in *Machine Learning and Knowledge Discovery in Databases*, vol. 9852, 2016, pp. 96–111.

[11] W. Street and Y. Kim, ''A streaming ensemble algorithm (SEA) for large-scale classification,'' in *Proc. 7th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*. San Francisco, NY, USA: ACM, 2001, pp. 377–382.

[12] Y. Lu, Y.-M. Cheung, and Y. Y. Tang, ''Adaptive chunk-based dynamic weighted majority for imbalanced data streams with concept drift,'' *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 8, pp. 2764–2778, Aug. 2020.

[13] D. Brzezinski and J. Stefanowski, ''Reacting to different types of concept drift: The accuracy updated ensemble algorithm,'' *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 1, pp. 81–94, Jan. 2014.

[14] R. Elwell and R. Polikar, ''Incremental learning of concept drift in nonstationary environments,'' *IEEE Trans. Neural Netw.*, vol. 22, no. 10, pp. 1517–1531, Oct. 2011.

[15] H. Guo, S. Zhang, and W. Wang, ''Selective ensemble-based online adaptive deep neural networks for streaming data with concept drift,'' *Neural Netw.*, vol. 142, pp. 437–456, Oct. 2021.

[16] Y. Sun, K. Tang, Z. Zhu, and X. Yao, ''Concept drift adaptation by exploiting historical knowledge,'' *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 10, pp. 4822–4832, Oct. 2018.

[17] J. Z. Kolter and M. A. Maloof, ''Dynamic weighted majority: A new ensemble method for tracking concept drift,'' in *Proc. 3rd IEEE Int. Conf. Data Mining*, Melbourne, FL, USA, Nov. 2003, pp. 123–130.

[18] P. Sidhu, M. Bhatia, and A. Bindal, ''A novel online ensemble approach for concept drift in data streams,'' in *Proc. IEEE 2nd Int. Conf. Image Inf. Process. (ICIIP)*, Dec. 2013, pp. 550–555.

[19] A. Bifet, G. Holmes, B. Pfahringer, R. Kirkby, and R. Gavaldà, ''New ensemble methods for evolving data streams,'' in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jun. 2009, pp. 139–148.

[20] J. Shan, H. Zhang, W. Li, and Q. Liu, ''Online active learning ensemble framework for drifted data streams,'' *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 2, pp. 486–498, Feb. 2019.

[21] H. Guo, S. Gao, and W. Wang, ''Concept drift detection and convergence based on hybrid ensemble of serial and cross,'' *J. Data Acquisition Process.*, vol. 37, no. 5, pp. 997–1011, 2022.

[22] Z. Chen, M. Han, M. Li, H. Wu, and X. Zhang, ''Survey of concept drift handling methods in data streams,'' *Comput. Sci.*, vol. 49, no. 9, pp. 14–32, 2022.

[23] R. S. M. de Barros and S. G. T. de Carvalho Santos, ''An overview and comprehensive comparison of ensembles for concept drift,'' *Inf. Fusion*, vol. 52, pp. 213–244, Dec. 2019.

[24] Z. Yang, S. Al-Dahidi, P. Baraldi, E. Zio, and L. Montelatici, ''A novel concept drift detection method for incremental learning in nonstationary environments,'' *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 1, pp. 309–320, Jan. 2020.

[25] G. Cheng, D. Qian, J. Guo, and Y. Zhao, ''A classification approach based on divergence for network traffic in presence of concept drift,'' *J. Comput. Res. Develop.*, vol. 57, no. 12, pp. 2673–2682, 2020.

[26] H. Yu and G. I. Webb, ''Adaptive online extreme learning machine by regulating forgetting factor by concept drift map,'' *Neurocomputing*, vol. 343, pp. 141–153, May 2019.

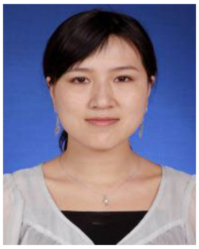
[27] A. Cano and B. Krawczyk, ''Kappa updated ensemble for drifting data stream mining,'' *Mach. Learn.*, vol. 109, no. 1, pp. 175–218, Jan. 2020.

[28] J. Gama, P. Medas, G. Castillo, and P Rodrigues, ''Learning with drift detection,'' in *Proc. Brazilian Symp. Artif. Intell.* Berlin, Germany: Springer, 2004, pp. 286–295.

[29] H. Wang, W. Fan, P. Yu, and J. Han, ''Mining concept-drifting data streams using ensemble classifiers,'' in *Proc. 9th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*. Hawthorne, NY, USA: IBM Thomas J. Watson Research, 2003, pp. 226–235.

[30] H. Guo, L. Cong, S. Gao, and W. Wang, ''Adaptive classification method for concept drift based on online ensemble,'' *J. Comput. Res. Develop.*, vol. 60, no. 7, pp. 1592–1602, 2023.

[31] R. Cui, X. Qi, Y. Liu, and L. Lin, ''Online ensemble adaptive algorithm for concept drift of streaming data,'' *J. Nanjing Univ., Natural Sci.*, vol. 59, no. 1, pp. 134–144, 2023.



NING LIU received the master's degree in computer software and theory from the School of Information Science and Technology, Northwestern University, China, in 2007. She is currently an Assistant Professor with the Faculty of Economics and Management, Shangluo University, China. Her current research interests include natural language processing, recommendation systems, sentiment analysis, and machine learning.

JIANHUA ZHAO received the Ph.D. degree in computer science and technology from the Computer School, Northwestern Polytechnical University, China, in 2014. He is currently a Professor with the School of Mathematics and Computer Application, Shangluo University, China. His current research interests include natural language processing, recommendation systems, sentiment analysis, and machine learning.