Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 21, No. 1, 2026

A concept drift-oriented active learning method

Jing Zhang, Siqi Zhang

School of Computer Science and Technology, Taiyuan Normal University, Jinzhong 030619, China

Abstract: The continuous growth and evolving internal distribution of streaming data were long challenged by the phenomenon of concept drift in classification tasks. In practical applications, labeling costs were prohibitively high, and active learning was frequently employed to mitigate the scarcity of labels. However, in drifting environments, a single sampling strategy was prone to induce selection bias and information redundancy. In this paper, a concept drift-oriented active learning method, termed CDAL, was proposed, which integrated drift detection with a clustering-based representative sampling mechanism. Upon detecting a distribution change, a candidate set was constructed from neighboring samples and structurally partitioned using online clustering. A global average uncertainty was then calculated as a dynamic reference, and the overall uncertainty level within each cluster was compared. Based on this comparison, either a centroid sample or a random sample was adaptively selected from the cluster for labeling. This design eliminated the need for preset fixed thresholds and balanced the information content of samples with their distribution coverage while controlling annotation costs. Experiments conducted on multiple real-world and synthetic datasets demonstrated that CDAL achieved superior cumulative accuracy and related evaluation metrics compared to baseline methods and was capable of rapidly restoring classification performance for new concepts after drift occurred, thereby validating the effectiveness of the proposed strategy.

Keywords: Data stream; Active learning; Concept drift.

1. Introduction

Under the wave of the big data era, the speed of data generation has accelerated sharply. These data are often continuously and dynamically changing and continue to grow over time, with the data volume being nearly infinite. Data with these characteristics are called Streaming Data or Data Stream[1]. One of the remarkable characteristics of streaming data is that as time progresses, the data distribution changes, and the relationship between the input features of samples and the output labels also changes. Generally, this phenomenon where data distribution changes unpredictably over time in a certain way is called concept drift[2].

Currently, the handling of concept drift includes two methods: passive adaptation[3] and active detection[4]. Passive adaptation methods do not actively detect concept drift in data streams; instead, they respond to concept drift by continuously adjusting the classification model. Such methods often have insufficient sensitivity in capturing drifts. Active detection methods involve adding a concept drift detection mechanism to the model to actively detect whether drift occurs. Once drift occurs, the current model is adjusted to timely adapt to new concepts. Moreover, in real-world environments, characteristics such as concept drift in data streams and high labeling costs often simultaneously exist and influence each other. Labeling all samples one by one is impractical, making it difficult to obtain completely true labeling information[5]. Traditional detection methods typically assume unrestricted access to the true labels of all samples in the data stream, which is not suitable for real-world scenarios. Therefore, active learning methods have significant advantages in this context. For active learning, its performance largely depends on the effectiveness of the sampling strategy. However, when concept drift occurs, the mapping relationship between the features of historical samples and their class labels changes over time.

To address the aforementioned issues, a concept drift-oriented active learning method, termed CDAL, was proposed in this paper. An online deep neural network model was designed and combined with a drift detection mechanism to identify changes in data distribution. Furthermore, a clustering-based representative active sampling strategy was introduced, in which incremental clustering techniques were employed to capture latent patterns within the newly emerging concept distribution. By incorporating a dynamic uncertainty threshold, samples with low redundancy and high representativeness were selected, thereby significantly improving labeling efficiency and the model's adaptability to evolving data streams.

2. Related Work

2.1. Concept Drift Detection Methods.

In the two major approaches to handling concept drift, the passive adaptation strategy typically collects samples from data streams at preset fixed time intervals and then periodically adjusts the classification model, without the need for a concept drift detection process[6]. The Very Fast Decision Tree (VFDT) algorithm proposed by Domingos[7] uses Hoeffding bound theory to determine how many samples are needed at a node to perform a split on a specific attribute. However, passive adaptation methods update models at a constant rate, lacking specificity to the actual dynamics of data changes. This often leads to unnecessary consumption of computational and time resources, significantly increasing time overhead.

Unlike passive adaptation, active detection methods typically rely on declines in classification model performance or changes in data distribution to actively detect concept drift, and update the classification model only when concept drift is detected. In distribution-based detection methods, such as the Region Drift Detection (RDD)[8], data blocks are divided into multiple regions through diverse partitioning patterns to measure local drift levels, and only the regions with the greatest divergence are selected for drift adaptation.

Most of the above detection methods highly depend on full

ground-truth labels, but acquiring labels is extremely costly in the real world, making drift detection techniques under weakly supervised conditions more practically significant.

2.2. Data Stream Active Learning Methods

In the research on active learning sample selection strategies, there are mainly three sampling criteria: uncertainty, representativeness, and diversity[9]. The uncertainty criterion determines the uncertainty of samples based on the reliability of the classifier's classification of samples. This strategy often only focuses on the category to which a sample is most likely to belong, while overlooking the possibility that samples may belong to other categories. The representativeness criterion aims to select a set of samples that can maximally represent all categories in the entire data distribution, while avoiding selecting outliers or noisy points during sampling. Zgraja et al.[10] developed a clustering-based approach that employs incremental clustering and a weighting mechanism for each cluster. According to a predefined labeling budget, multiple data instances are randomly selected from the clusters with the highest weight rankings for labeling, and the classifier is adjusted and optimized using the labels of these samples.

In this context, the above-mentioned active learning methods all have certain limitations.

3. Concept Drift-Oriented Active Learning

Currently, most data stream classification algorithms rely on supervised learning, but acquiring real-world labels often incurs substantial human and material costs in practice. Active learning methods enable classification tasks to be effectively completed with limited labeling costs. However, the occurrence of concept drift and the use of single sampling strategies can cause the samples queried by active learning to fail to accurately represent the true distribution of current data, contain excessive redundant information, and overlook other samples of critical value. This situation thus leads to the problem of sampling bias, limiting the information learned by the model. To address this issue, this paper proposes a concept drift-oriented active learning method. Figure 1 illustrates the overall framework of the method.

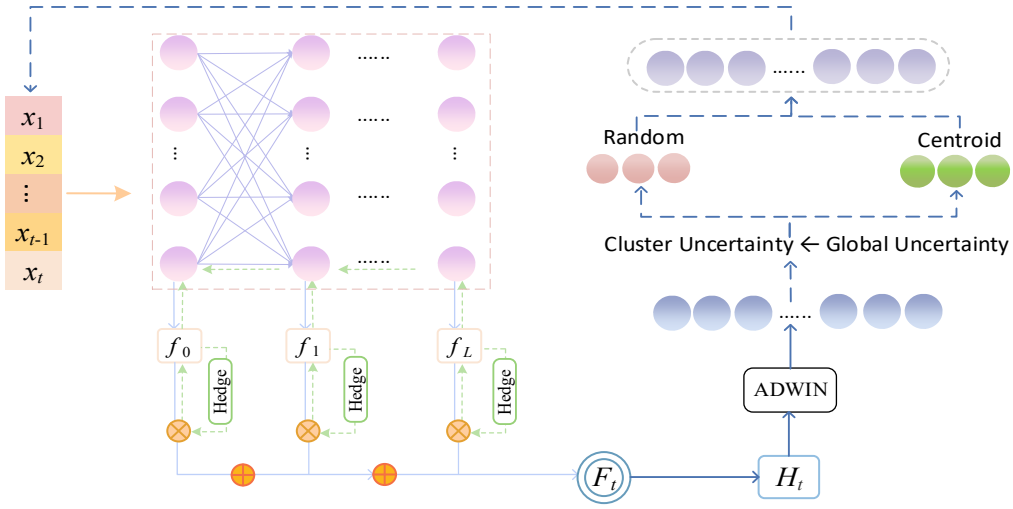

Fig. 1 Overall framework of CDAL

3.1. Problem Definition.

Formally, the data distribution refers to the joint probability distribution of the data, i.e.,$P(x, y)$. Concept drift is defined as the change in the joint probability distribution of data between two different time instances t and $t+1$, which can be expressed as:

$$Pt(x,y) \neq Pt+1(x,y) \quad (1)$$

Where x represents the feature attributes of instances in the data stream, and y denotes the instance labels.

By decomposing the joint probability into $P(x,y) = P(x)P(y|x)$concept drift can be attributed to changes in the prior probability $P(x)$ of samples and the posterior probability $P(y|x)$.

This paper adopts different sensitivity $\varepsilon \in (0,1)$ setting approaches for different data streams:

$$\varepsilon = \sqrt{\frac{1}{2m} \cdot ln\frac{4}{\delta'}} \quad (2)$$

m denotes the mixed mean of two subwindow lengths. δ is used to correct the hypothesis testing error, with the expectation that the global error remains below δ The optimization for m and δ is as follows:

$$\delta' = \frac{\delta}{n} \quad (3)$$

$$m = \frac{1}{\frac{1}{n_0}+\frac{1}{n_1}} \quad (4)$$

Here, n_0 and n_1 denote the lengths of two adjacent subwindows, so $n = n_0 + n_1 n = n_0 + n_1$ is the total observation window length. Concept drift is detected when the absolute value of the mean difference between the two subwindows exceeds ε.

3.2. Clustering-Based Representative Active Learning Sampling Strategy.

The performance of active learning methods heavily depends on the quality of the label query strategy, which directly affects the model's generalization ability under limited annotation budgets. When concept drift occurs, the new data distribution often exhibits several previously unknown category patterns and density structures. If sampling is conducted solely based on a single uncertainty metric, the selected samples are prone to noise interference

and may overly concentrate on low-density regions near the decision boundary, thereby failing to comprehensively capture the overall distribution of the new concept. To address this issue, this paper proposes a clustering-based representative active sampling strategy. Once drift detection is triggered, incremental clustering techniques are employed to structurally partition the candidate samples. Within each resulting cluster, representative samples are selected in a differentiated manner according to the cluster's uncertainty level, thereby maximizing distribution coverage and noise robustness while ensuring sufficient information content.

For each cluster obtained from the clustering process, the average entropy of all samples within that cluster is first computed as the overall uncertainty level of the cluster. This cluster-level average uncertainty is then compared with the global average:

(1) If the average uncertainty of a cluster is higher than the global average, it indicates that the cluster as a whole lies near the decision boundary of the current model, where old and new concepts heavily overlap and information value is high. To avoid redundancy caused by repeatedly selecting similar samples within the same highly uncertain region, one sample is randomly chosen from the cluster and submitted for expert labeling.

(2) If the average uncertainty of a cluster is not higher than the global average, it indicates that the samples within the cluster can be classified with relative certainty by the model. However, this precisely suggests that the cluster represents a new distribution pattern that the model has preliminarily adapted to but has not yet firmly mastered. To accurately capture the central characteristics of this pattern, the sample closest to the cluster centroid is selected for labeling.

3.3. Datasets and Comparative Methods.

Detailed information on the datasets is presented in Table 1.In this paper, the proposed method is compared with six baseline methods: No Det[11], Ran Retr[11], Eql Retr[11], UDD[11], KSWIN[12]. Since the KSWIN method is prone to generating false positive drift signals, two groups of comparisons are configured in the experiments.

Table 1. Datasets Information

Datasets	Instances/103	Attributes	Classes	Drift Type	Drift Position/103
Kddcup99	494	41	23	unknown	-
Electricity	45	6	2	unknown	-
Weather	95	9	3	unknown	-

("-" denotes an uncertain drift position)

3.4. Evaluation Metrics.

In this paper, the generalization performance of the model is evaluated and analyzed using the following metrics.

1)Cumulative Accuracy (Cumacc): This metric refers to the ratio of the cumulative number of correctly predicted samples from the beginning up to the current moment to the total number of samples, expressed as:

$$Cumacc = \frac{1}{T*n}\sum_{t=1}^{T} n_t \quad (5)$$

where n denotes the total number of samples, and n_t represents the number of correctly predicted samples at time t.

2)Matthews Correlation Coefficient (MCC): This metric takes into account all four categories in the confusion matrix (true positives, true negatives, false positives, and false negatives). It is suitable for imbalanced datasets and avoids the bias that may arise from using a single metric under class imbalance conditions.

3)Area Under the ROC Curve (AUC): The ROC curve describes the relationship between the true positive rate and the false positive rate of a classifier under different thresholds. The AUC value represents the area under the ROC curve and is used to measure classifier performance. An AUC value closer to 1 indicates better classifier performance.

3.5. Experimental Results and Analysis.

3.5.1. Analysis of Cumulative Accuracy Results

Table 2 presents the cumulative accuracy results of different methods across various datasets. As can be observed from Table 2, the proposed CDAL method achieves the highest cumulative accuracy across all three datasets, with an average rank of 1.0, significantly outperforming the other compared methods. On the Electricity dataset, CDAL attains a cumulative accuracy of 0.7325, representing an improvement of approximately 3%. On the Kddcup99 dataset, CDAL maintains the top position despite a narrower margin of improvement. On the Weather dataset, CDAL achieves an accuracy of 0.9857, substantially higher than the second-best Kswin. The introduction of a clustering-based representative sampling strategy in CDAL effectively reduces redundant labeling in high-confidence regions and concentrates the limited labeling budget on samples with the highest information value, thereby yielding further performance gains.

Table 2. Cumulative Accuracy Comparison of Different Methods

Datasets	Average Ranking of Each Method in Cumulative Accuracy						
	UDD	Eql Retr	Kswin	Kswin_unl	No det	Ran Retr	CDAL
Electricity	0.7003(4)	0.7020(3)	0.6923(5)	0.6893(6)	0.6719(7)	0.7032(2)	0.7325(1)
Kddcup99	0.9891(5)	0.9907(4)	0.9913(2)	0.9913(2)	0.9820(7)	0.9889(6)	0.9952(1)
Weather	0.9729(7)	0.9810(3)	0.9837(2)	0.9792(5)	0.9783(6)	0.9801(4)	0.9857(1)
Average Rank	5.3	3.3	3	4.3	6.7	4.0	1.0

3.5.2. Analysis of Matthews Correlation Coefficient (MCC) Results

Table 3 presents the experimental results of CDAL and six comparative methods based on the Matthews correlation coefficient. As can be seen from Table 3, CDAL achieves significantly higher MCC values than all compared methods

across all test datasets, further demonstrating the considerable advantage of the proposed method in terms of overall performance evaluation. This notable improvement is attributed to the fact that, upon detecting concept drift, CDAL is capable of identifying newly emerging clusters through clustering and prioritizing the labeling of high-uncertainty samples within these clusters. This approach effectively reduces misclassification on minority class samples, thereby substantially enhancing the MCC values.

Table 3. Matthews correlation coefficient (MCC)

Datasets	UDD	Eql Retr	Kswin	Kswin_unl	No det	Ran Retr	CDAL
Electricity	0.4162	0.4302	0.4023	0.4019	0.4002	0.4328	0.5102
Kddcup99	0.9732	0.9779	0.9836	0.9836	0.9624	0.9689	0.9908
Weather	0.8901	0.9005	0.9025	0.8932	0.8925	0.8996	0.9285

3.5.3. Analysis of Area Under the ROC Curve (AUC) Results

Table 4 presents the experimental results of CDAL and six comparative methods in terms of the area under the ROC curve (AUC).

As can be observed, the proposed CDAL method achieves AUC values of 0.8179, 0.9701, and 0.7426 on the Electricity, Kddcup99, and Weather datasets. This outcome is primarily attributed to the fact that, upon detecting concept drift, CDAL performs structural partitioning of candidate samples via online clustering and adaptively selects either centroid samples or random samples for labeling by incorporating a dynamic global uncertainty threshold. Such a mechanism effectively captures the critical support vectors of the new concept distribution while controlling labeling redundancy, enabling the deep neural network model to rapidly adjust the classification boundary and thereby enhancing overall AUC performance.

Table 4. Area under the ROC curve (AUC)

Datasets	UDD	Eql Retr	Kswin	Kswin_unl	No det	Ran Retr	CDAL
Electricity	0.7749	0.7756	0.7674	0.7662	0.7602	0.7783	0.8179
Kddcup99	0.9627	0.9665	0.9683	0.9683	0.9597	0.9610	0.9701
Weather	0.7229	0.7286	0.7310	0.7251	0.7243	0.7256	0.7426

4. Conclusion

Concept drift in data streams can impair the sampling effectiveness of active learning, as a single strategy often struggles to ensure both the information value of selected samples and adequate coverage of the overall data distribution. To address this issue, this paper combines drift detection with clustering-based representative sampling to construct the CDAL framework. Upon detecting a drift, the method performs online clustering on candidate samples surrounding the detection point and adaptively decides whether to select a centroid sample or a random sample from each cluster by comparing the cluster's average uncertainty with the global average. This design eliminates the need for manually preset fixed thresholds and achieves a reasonable balance between sampling efficiency and representativeness. Future research will focus on sample selection in imbalanced data streams to further enhance the applicability of the proposed method.

References

[1] GAMA J, GANGULY A, OMITAOMU O, et al. Knowledge discovery from data streams[J]. Intelligent Data Analysis, 2009, 13(3): 403-404.

[2] WEBB G L, HYDE R, CAO H, et al. Characterizing concept drift[J]. Data Mining & Knowledge Discovery, 2016, 30(4): 964-994.

[3] SUÁREZ-CETRULO A L, QUINTANA D, CERVANTES A. A survey on machine learning for recurring concept drifting data streams[J]. Expert Systems with Applications, 2023, 213: 118934.1-118934.17.

[4] YU H, LIU W, LU J. et al. Detecting group concept drift from multiple data streams[J]. Pattern Recognition, 2023, 134: Article No. 109113.

[5] SHAHRAKI, A, ABBASI, M, TAHERKORDI, A, et al. Active learning for network traffic classification: A Technical Study[J]. IEEE Transactions on Cognitive Communications and Networking, 2022, 8(1): 422-439.

[6] KARIMIAN M, BEIGY H. Concept drift handling: a domain adaptation perspective[J]. Expert Systems with Applications, 2023, 224: Article No. 119946.

[7] DOMINGOS P, HULTEN G. Mining high-speed data streams [C]// Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining. New York: ACM, 2000: 71-80.

[8] LIU A, LU J, ZHANG G. Diverse instance-weighting ensemble based on region drift disagreement for concept drift adaptation[J]. IEEE transactions on neural networks and learning systems, 2020, 32(1): 293-307.

[9] DE ROSA R, CESA-BIANCHI N. Confidence decision trees via online and active learning for streaming data[J]. Journal of Artificial Intelligence Research, 2017, 60: 1031-1055.

[10] ZGRAJA J, GAMA J, AND WO´ZNIAK M. Active learning by clustering for drifted data stream classification [C]// Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases Workshops. Dublin: Springer, 2019: 80-90.

[11] DOMINIK K, NIKLAS K, SEBASTIAN H. Machine learning operations (MLOps): Overview, definition, and architecture[J]. IEEE Access, 2023, 11: 31866-31879.

[12] MAURRAS U T, YOUSRA C, ALIOU B, et al. Anomalies detection using isolation in concept-drifting data streams[J]. Computers, 2021, 10(1): 1-21.