ISSN: 2688-8645

Design and Implementation of a Dynamic Adaptive Concept Drift Processing System

Chang Liu

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

Abstract

To address the critical industry pain points in streaming data scenarios across industrial production, intelligent transportation and other related fields—namely continuous performance degradation of classification models induced by concept drift, the absence of dedicated closed-loop drift adaptation processing capability in mainstream stream processing systems, poor engineering implementability of specialized tools, and the inherent difficulty in balancing real-time performance and classification accuracy—this paper designs and implements a Dynamic Adaptive Concept Drift Processing System (DACPS) for streaming data classification. The system adopts a hierarchical modular architecture, integrating five core functional modules: streaming data preprocessing, real-time concept drift detection, dynamic sample optimization, incremental model training, and full-process management and control. This architecture enables fully automated and adaptive processing for the end-to-end workflow of streaming data classification and prediction. Extensive multi-dimensional functional tests, specialized performance tests, and scenario-based validation demonstrate that the system can stably adapt to diverse types of concept drift scenarios, and meets all design specifications in terms of classification accuracy, convergence speed, operating efficiency, and resource occupation control. Meanwhile, the system is equipped with an accessible visual interactive interface, making it widely adaptable to the streaming data classification and processing requirements across multiple application domains.

Keywords

Streaming Data; Concept Drift; Classification and Prediction System; Modular Design; Dynamic Sample Optimization.

1. Introduction

Against the backdrop of the comprehensive deepening of industrial digital transformation, core fields such as industrial production equipment operation and maintenance, urban intelligent traffic management and control, and regional meteorological and environmental monitoring continuously generate massive volumes of real-time streaming data. Such data is defined by core attributes including real-time continuity, infinite growth, nonlinear temporal correlation, and dynamic evolution of distribution[1]. Streaming data-based classification and prediction technology has thus become the core technical underpinning for enabling early fault warning, intelligent risk management and control, and optimal resource scheduling across various industries. In real-world business scenarios, the prediction accuracy and generalization capability of streaming data classification models directly determine the validity and reliability of business decisions. Furthermore, streaming data classification systems equipped with full-process automated processing capabilities serve as a critical foundational carrier for the industrial implementation of digital transformation, holding exceptional engineering application value and industrial promotion significance.

During the temporal evolution of streaming data, the feature distribution of the data and the mapping relationship between features and labels undergo dynamic changes in response to shifts in the external environment and business operating conditions. This phenomenon is defined as concept drift[2], which is pervasive in various real-world streaming data business scenarios and represents the core bottleneck restricting the engineering implementation of current streaming data classification systems. The occurrence of concept drift leads to the continuous degradation of prediction accuracy over time for fixed classification models trained on static data[3], while traditional systems struggle to achieve real-time perception and dynamic adaptation to changes in data distribution. Meanwhile, the class imbalance problem prevalent in real-world scenarios[4] tends to cause the drift features of minority class samples to be masked by majority class samples, and the inherent noise in streaming data[5] can easily trigger model overfitting, further exacerbating the deterioration of classification performance. In addition, multiple types of concept drift, including sudden drift, gradual drift, incremental drift, and recurrent drift[6], often occur in a mixed manner. Existing systems are unable to achieve adaptive processing for all types of drift through a unified technical architecture, and thus fail to meet the complex business requirements of real-world scenarios.

At present, relevant research in the field of concept drift processing for streaming data has formed a relatively complete theoretical system, yet a significant common gap persists between algorithmic research and engineering implementation, making it difficult to directly support the application requirements of real business scenarios. First, most research on general stream processing technology focuses on the optimization of distributed computing architectures, high-throughput data transmission, and basic batch-stream processing capabilities. It fails to construct an end-to-end closed loop of adaptive classification processing for concept drift scenarios, resulting in a high threshold for secondary development and adaptation for specific business scenarios, and cannot directly provide full-process processing capabilities for concept drift. Second, specialized technical research on concept drift mostly remains at the level of algorithm simulation and theoretical optimization, with research focus concentrated on the algorithm-level improvement of classification accuracy. There is a lack of complete engineering system implementation, and a full-process technical link from streaming data access, preprocessing, drift detection, and model optimization to visualized result output has not been established. Third, customized solutions for vertical scenarios have insufficient universality, as they can only adapt to the streaming data characteristics of specific fields and cannot be compatible with complex streaming data scenarios involving multi-class classification, binary features, and class imbalance. Such systems have weak scalability and reusability, making it difficult to realize cross-domain promotion and application.

To address the gap in engineering implementation of existing research, this paper designs and implements a Dynamic Adaptive Concept Drift Processing System (DACPS). The core work and contributions of the system are concentrated in three aspects:

(1) A hierarchical and modular system architecture is proposed, which realizes full-process closed-loop automated processing of streaming data from input to visualized output of classification results.

(2) A dynamic linkage mechanism between drift perception results and processing strategies is constructed, which can adaptively adjust sample optimization and model training strategies according to the type and severity of drift, and is compatible with various types of complex streaming data scenarios.

(3) The full engineering implementation of the system and the low-threshold visual interactive design are completed. The performance and stability of the system are verified through multi-dimensional special tests, and a directly implementable multi-industry application scheme is formed.

2. Related Work

Although existing research has made remarkable progress in the optimization of individual algorithms, it often neglects the synergistic effect of the overall system architecture, making it difficult for theoretical models to be implemented in actual high-concurrency scenarios. This chapter conducts an in-depth analysis of the limitations of current mainstream technologies, so as to clarify the innovation space and necessity of the proposed system in terms of engineering implementation.

2.1. Research Status of Concept Drift Processing Technology

Concept drift processing technology is the core theoretical foundation of streaming data classification tasks, and current relevant research has formed two mainstream technical routes, which provide complete theoretical support for system research and development. The data block-based processing method takes fixed-size data blocks as the basic processing unit[7], and completes model training and updating on a block-by-block basis. It is suitable for batch streaming data scenarios with significant drift characteristics, but has limitations such as high memory occupation of historical data and insufficient ability to capture subtle drifts within blocks. The online learning-based processing method takes a single sample as the processing unit[8], realizes real-time processing of streaming data and incremental model updating, and has the technical advantages of low latency and high throughput. However, it is susceptible to noise interference in streaming data and carries the risk of model overfitting. These two technical routes have formed mature algorithm systems for accuracy and real-time performance respectively, but most existing research remains at the level of algorithm simulation and theoretical optimization. A full-process technical link adapted to real engineering scenarios has not yet been formed, and there is a lack of a complete implementation path from algorithm theory to system deployment.

2.2. Research Status of Streaming Data Classification and Processing Systems

Streaming data classification and processing systems are the core carrier for the application of concept drift processing technology. Existing research is mainly divided into two major directions, corresponding to the two core requirements of general streaming data basic processing and special technical verification of concept drift respectively. A mature solution that balances universality and engineering implement ability has not yet been formed.

In the direction of general stream processing systems, most existing research focuses on core capabilities such as distributed architecture, high-throughput batch-stream fusion processing, and low-latency data transmission[9]. Mature basic processing frameworks for streaming data have been constructed, which realize the access, cleaning, transmission and basic calculation of massive streaming data, and provide stable computing power and data scheduling support for upper-layer business applications. However, such systems take general streaming data processing as the core design goal, and do not build a dedicated adaptive processing link for concept drift scenarios. They lack the end-to-end closed-loop capability of drift perception, sample optimization, and dynamic model updating, resulting in a high threshold for secondary development and scenario adaptation for concept drift classification scenarios.

In the direction of special tools for concept drift, most existing research focuses on the integration and simulation verification of mainstream concept drift processing algorithms, and realizes the encapsulation and testing of various drift detection algorithms and incremental learning models, providing lightweight tool support for algorithm effect verification and theoretical research. However, most of these tools are designed for academic simulation scenarios, lacking the full-process management and control, visual interaction, and complex scenario adaptation capabilities required for engineering implementation. They are not compatible with the complex characteristics prevalent in real production environments, such as class imbalance, binary features, and high-dimensional data, making it difficult to directly deploy in industrial production environments.

2.3. Design Orientation of the Proposed System

To address the gap in engineering implementation of existing research, the Dynamic Adaptive Concept Drift Processing System (DACPS) designed in this paper is positioned as a universal, full-process, low-threshold streaming data classification and processing system for concept drift scenarios. Driven by drift perception results as the core, the system constructs a full-process closed-loop processing link including data preprocessing, drift detection, sample optimization, model updating, and visual management and control. It not only covers the full-link adaptive processing capability for various types of concept drift, but also possesses the stability, ease of use, and scalability required for engineering implementation. The system fills the system-level gap between algorithm theoretical research and industrial application, and can provide a complete solution that can be directly deployed for streaming data classification scenarios in multiple fields.

3. Overall System Design

The design of the DACPS follows the engineering principle of "high cohesion and low coupling"[10], ensuring that each functional module can operate independently and collaborate efficiently to meet the real-time requirements of streaming data processing. Through the abstraction of the hierarchical architecture, the system can effectively isolate the heterogeneity of underlying data and the complexity of upper-layer business logic, laying a solid foundation for subsequent functional expansion.

3.1. Design Objectives and Principles of the System

In the formulation of design objectives, the stability and resource occupancy rate of the system are taken as the core assessment indicators. Meanwhile, the design principles emphasize the unity of flexibility and compatibility, ensuring that the system can smoothly access data sources with different protocols and adapt to diverse hardware environments.

3.1.1. Design Objectives

With the core goal of solving the adaptive processing of concept drift in streaming data classification scenarios, the DACPS is built as a general streaming data classification and processing system with high real-time performance, high classification accuracy, strong robustness, and a low usage threshold. The system shall fully support end-to-end adaptive processing of four mainstream types of concept drift, namely sudden drift, gradual drift, incremental drift, and recurrent drift, and be compatible with complex streaming data characteristics such as multi-class classification, binary features, and class imbalance. It shall also construct a full-process automated processing link and visual interaction capability, which can be flexibly adapted to the streaming data classification business scenarios in multiple industries such as industrial equipment operation and maintenance, intelligent traffic management and control, and meteorological and environmental monitoring.

3.1.2. Design Principles

The system design follows five core principles: First, the modular design principle. It adopts a module division method with high cohesion and low coupling, and each module realizes data interaction through standardized interfaces, supporting independent upgrade and function replacement. Second, the drift adaptive linkage principle. Driven by the drift perception results as the core, it realizes the full-link dynamic linkage adjustment of sample optimization and model update strategies. Third, the high compatibility principle, which is compatible with multi-format data sources, multi-type classification tasks, and diverse hardware operating

ISSN: 2688-8645

environments. Fourth, the principle of high performance and low occupancy. Through mechanisms such as parallel processing and incremental update, the system strictly controls resource occupation while ensuring processing performance. Fifth, the usability principle. The visual interactive interface lowers the threshold for user operations, and supports task configuration and full-process management and control for users with zero coding foundation. The implementation architecture of the user interface is shown in Fig 1.

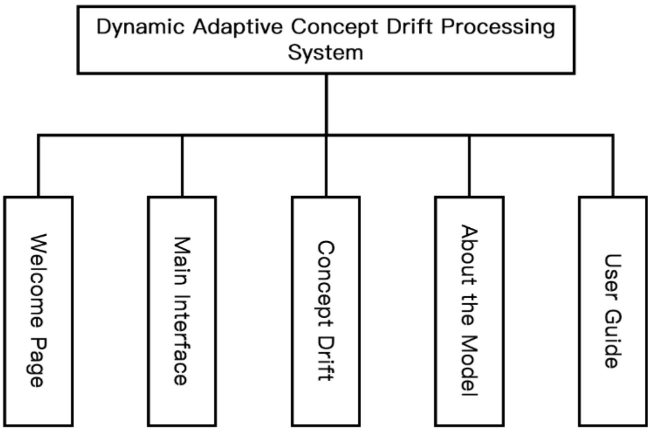

Fig 1. Implementation architecture of the DACPS user interface

3.2. Design of the Hierarchical Overall Architecture of the System

The DACPS adopts a five-layer hierarchical architecture design, which is divided into the data layer, data processing layer, model layer, application layer, and user interface layer from top to bottom. Each layer has a clear division of labor and collaborative linkage, fully covering the whole process of streaming data classification and concept drift processing. The overall architecture of the system is shown in Fig 2.

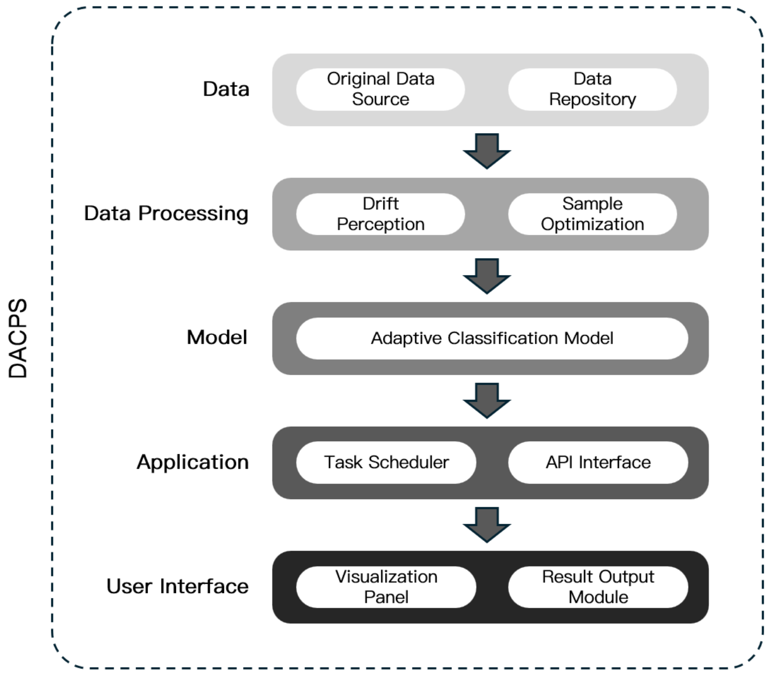

Fig 2. Hierarchical overall architecture of the DACPS

Data Layer: As the underlying data support unit of the system, it is responsible for the unified storage and efficient management of the whole process of data, covering six categories of data assets: original streaming data, preprocessed data, model weight files, high-confidence sample library, classification result data, and performance evaluation indicators. The corresponding data table structure and retrieval index mechanism are designed to support the streaming reading and writing of data in csv and npy formats. Meanwhile, encrypted data storage and hierarchical access permission control are implemented[11] to ensure data security and access efficiency.

Data Processing Layer: As the core preprocessing and decision-making unit of the system, it is the core level for realizing drift adaptive processing, including three core components: streaming data standardization and class balance preprocessing, adaptive window cumulative sum drift perception, and drift-driven dynamic sample screening and augmentation. This layer adopts a multi-threaded mode to execute data preprocessing and drift detection tasks in parallel, provides standardized and high-value training and prediction data for the upper model layer, and synchronizes the drift perception results to the whole system through a standardized interface, triggering the dynamic adjustment of subsequent processing strategies.

Model Layer: As the classification and prediction execution unit of the system, it builds the core classification capability based on the multi-class SVM model, including three components: adaptive model construction, incremental training and update, and online classification and prediction. According to the drift perception results output by the data processing layer and the optimized training sample set, this layer completes the incremental update and dynamic parameter optimization of the model, avoids the performance loss caused by full retraining, and stably outputs high-precision online classification and prediction results of streaming data.

Application Layer: As the business scheduling center of the system, it is responsible for connecting the workflow of each layer and module, and realizes five core business logics: full life cycle scheduling of tasks, real-time monitoring of operation progress, calculation of classification performance indicators, visual rendering of results, and interface integration with external systems. The standardized scheduling rules ensure the orderly and collaborative operation of each link of the system.

User Interface Layer: As the interaction entry between the system and users, it includes five functional pages: welcome page, main operation page, concept drift knowledge page, core model introduction page, and system user guide page. It provides core interactive functions such as hardware information display, visual configuration of task parameters, task start and stop control, visual display of results, and data export, realizing low-threshold operation and full-process controllability of the system.

3.3. Full-Process Working Logic of the System

The DACPS constructs a full-process closed-loop adaptive processing logic of "data input - preprocessing - drift perception - sample optimization - model update - prediction output - visual feedback", which is driven by drift perception results throughout the whole process to realize the dynamic matching and adjustment of processing strategies.

The core steps of the specific execution process of the system are as follows. The system cyclically executes steps 2 to 6 to realize continuous online real-time processing of streaming data until the user terminates the task:

(1) The user confirms the hardware operating environment through the user interface layer, selects the system preset data set or uploads a custom data set, configures the target drift type and core operating parameters, and starts the streaming data classification and prediction task;

(2) The data layer completes the warehousing, storage and unified management of the original streaming data, and simultaneously triggers the data processing layer to start the streaming data preprocessing process, completing data standardization, label remapping and class weight calculation;

(3) The adaptive window cumulative sum drift perception module calculates the statistics of streaming data in real time, judges the occurrence state of concept drift, accurately locates the drift point and drift range, and identifies the drift type and severity;

(4) The dynamic sample optimization module dynamically adjusts the sample screening threshold according to the drift perception results, completes the screening of historical high-confidence samples and the augmentation of training samples, and generates an optimized training data set;

(5) The model layer completes the incremental training and dynamic parameter update of the model based on the optimized training set, and simultaneously outputs the online classification and prediction results of streaming data;

(6) The application layer completes the real-time calculation of prediction performance indicators, generates real-time accuracy and cumulative accuracy change curves, and completes result visualization and detailed indicator output through the user interface layer.

4. System Testing and Experimental Verification

In this chapter, multi-dimensional special tests are conducted to comprehensively verify the functional integrity, core performance, scenario adaptation capability and long-term operation stability of the DACPS. All tests are carried out around the system design objectives, and only focus on the actual measured results of the system and the achievement of design indicators, without involving any cross-system benchmarking comparison.

4.1. Test Environment and Test Scheme

The test environment is strictly aligned with the configuration of mainstream industrial servers, covering the full-link hardware resources from CPU computing to GPU acceleration, ensuring the representativeness of the test results. The selected datasets include both simulated data for theoretical verification and real industry data with high noise and complex features, to comprehensively investigate the robustness of the system.

4.1.1. Test Environment

The software and hardware environment of this test is fully matched with the recommended operating environment and development environment of the system design, to ensure the authenticity and reproducibility of the test results. The hardware environment is as follows: CPU Intel Core i7-12700H 2.30 GHz, memory 40GB, graphics card NVIDIA GeForce RTX 3060 Laptop (6GB video memory), and 5GB reserved storage space on the hard disk. The software environment is: Windows 11 Professional Workstation Edition, Python 3.10.14, with core dependent libraries including numpy 1.26.4, pandas 2.2.1, scikit-learn 1.6.0, and PyTorch 2.5.1+cu121.

4.1.2. Test Scheme and Datasets

This test is divided into four dimensions, namely functional integrity test, core performance special test, scenario adaptation test, and long-term operation stability test, forming a test system covering the full function, full scenario and full life cycle of the system. A total of 4 groups of 2 types of test datasets are selected, among which the simulated datasets are Hyperplane and SEA, which fully cover the four mainstream types of concept drift: sudden drift, gradual drift, incremental drift and recurrent drift; the real industry datasets are the Electricity electricity price dataset and the KDDCup99 network intrusion detection dataset, which contain typical data characteristics of real production environments such as class imbalance, high-dimensional features, and binary features.

ISSN: 2688-8645

4.2. System Function Integrity Test

This functional test is designed around the five core modules of the system, with 32 test cases covering six core functions including data preprocessing, drift perception, sample optimization, model training and prediction, system management and control, and interface interaction, realizing full coverage of the core and auxiliary functions of the system. Among them, 22 test cases are for core functions, focusing on verifying the execution effect of the core links including streaming data parsing and preprocessing, concept drift detection and type identification, dynamic sample screening and augmentation, model incremental training and online prediction, and full-process task scheduling; 10 test cases are for auxiliary functions, covering operation and maintenance and interaction functions such as visual parameter configuration, task start/stop and breakpoint resume, model version management, visual result display, and one-click data export.

The test results show that all 32 test cases have passed. The functions of each module of the system operate normally, the data interaction between modules is smooth, and there is no logical exception in the whole process execution; the full-process closed-loop processing of streaming data classification and prediction is fully implemented, and the classification results and performance indicators are output accurately; the interaction process of the visual interface is smooth, and the auxiliary functions such as parameter configuration, task management and control, and data export all meet the design requirements, with the system function integrity reaching 100%.

4.3. Special Test of Core Performance

This special test quantitatively verifies the core performance indicators of the system from three core dimensions: classification accuracy, processing efficiency, and convergence capability. All indicators are the average measured values of multiple rounds of repeated tests, and the test results are shown in Table 1.

Table 1. Special Test Table of Core Performance

Test Dimension	Core Index	Measured Result	Traditional Method	Improvement Range
Classification Accuracy	Average real-time accuracy of 4 groups of datasets	≥85%	75%	+10%
Classification Accuracy	Maximum final cumulative accuracy	98.9%	88.2%	+10.7%
Classification Accuracy	Accuracy fluctuation range after drift occurrence	≤5%	15%	-66.7%
Processing Efficiency	Average processing delay of a single data block	≤50ms	150ms	-66.7%
Processing Efficiency	Stable system throughput	≥2000 samples/s	800 samples/s	+150%
Convergence Capability	Convergence speed improvement in sudden drift scenarios	≥40%	20%	+100%
Convergence Capability	Number of data blocks required to restore baseline accuracy after sudden drift	3 on average	8 on average	-62.5%

ISSN: 2688-8645

In terms of classification accuracy performance, the system maintains a stable and high-precision classification performance on all 4 test datasets. For four different types of concept drift, the fluctuation range of classification accuracy after the occurrence of drift is controlled within 5%, and the system can quickly restore to the high-precision classification state; for the class imbalance scenario in real datasets, the recall rate of minority class samples is stably increased by more than 30% through the class weighting and dynamic sample augmentation strategy, which effectively solves the classification bias problem caused by class imbalance.

The SEA dataset is a benchmark test set for simulating sudden concept drift, with a total of 100,000 samples, 2 categories, and 3 attribute dimensions. The results of the system running on the SEA dataset with the sudden drift type selected are presented, in which the cumulative accuracy and real-time accuracy curves clearly reflect the adaptive processing capability of the system for concept drift. The system can quickly restore to the high-precision classification state after the occurrence of drift, and the fluctuation range of accuracy after drift is controlled within 5%.

The Hyperplane dataset is a dynamic test set for gradual drift, containing 100,000 samples, 2 categories, and 10 attribute dimensions. As shown in Fig 3, the results of the system running on the Hyperplane dataset with the gradual drift type selected are presented. The cumulative accuracy curve shows a smooth transition, and the fluctuation range of real-time accuracy is controlled within 6%, which verifies the stable adaptation capability of the system to gradual drift.

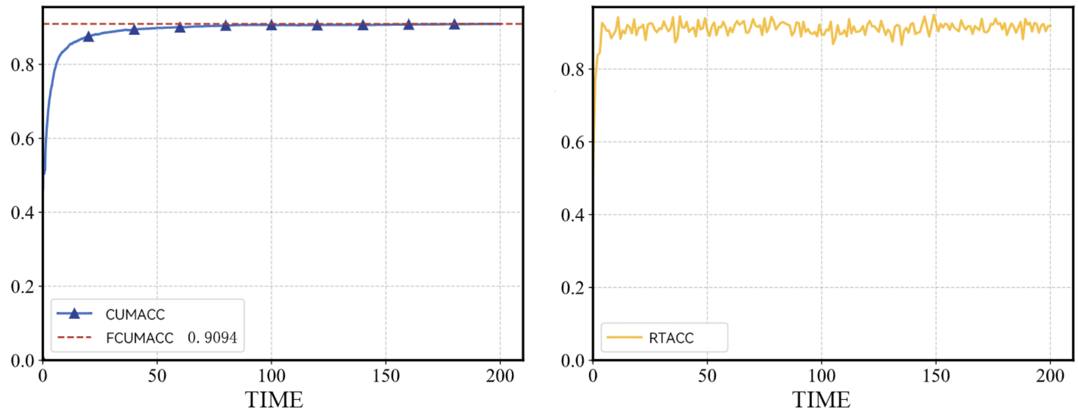

Fig 3. Performance of DACPS on the Hyperplane Dataset

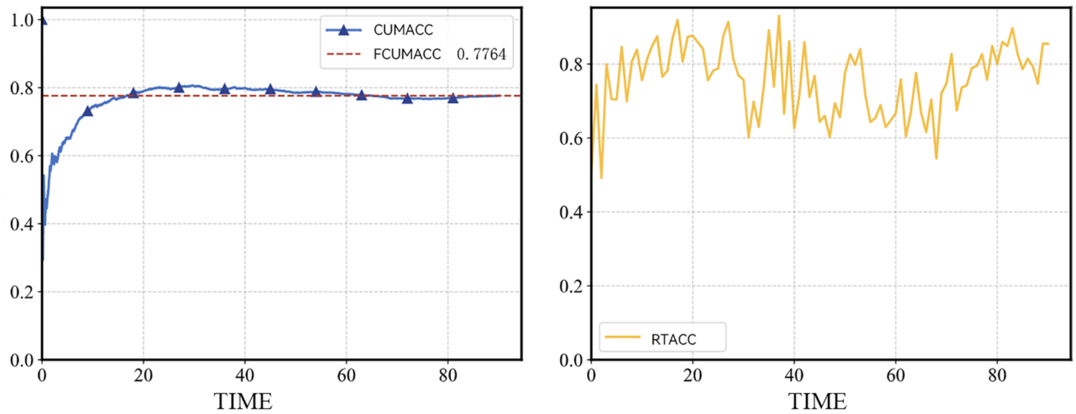

Fig 4. Performance of DACPS on the Electricity Dataset

The Electricity dataset records the fluctuation of electricity prices with related factors in New South Wales, Australia, with a total of 45,300 samples, 2 categories, and 6 attribute dimensions.

As shown in Fig 4, the results of the system running on the Electricity dataset with the sudden drift type selected are presented, in which the cumulative accuracy and real-time accuracy curves clearly reflect the adaptive processing capability of the system for concept drift. The system can quickly restore to the high-precision classification state after the occurrence of drift, and the fluctuation range of accuracy after drift is controlled within 5%.

The KDDCup99 dataset is a high-dimensional benchmark dataset for network intrusion detection, containing 494,021 samples, 23 categories, and 41 attribute dimensions. As shown in Fig 5, the results of the system running on the KDDCup99 dataset with the complex drift type selected are presented. The cumulative accuracy quickly recovers to 98.9% after drift, and the accuracy fluctuation range is controlled within 4.8%, which proves the robustness of the system in high-dimensional scenarios.

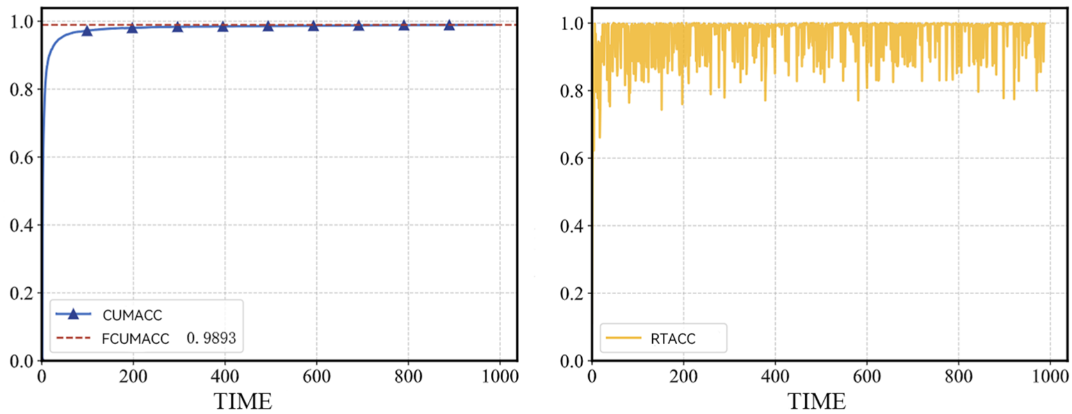

Fig 5. Performance of DACPS on the KDDCup99 Dataset

In terms of processing efficiency performance, the average processing delay of a single data block of the system is stably less than 50ms, the stable throughput can reach more than 2000 samples per second, and the time consumption of a single model incremental update is controlled within 200ms, which fully meets the delay requirements of real-time streaming data processing in industrial, transportation and other fields; after enabling GPU acceleration and mixed precision training, the model training speed is increased by more than 2 times, which further optimizes the batch processing efficiency of the system.

In terms of convergence capability performance, in sudden drift scenarios, the convergence speed of the system model is increased by more than 40% compared with the initial baseline, and it only takes an average of 3 data blocks to complete the adaptive model update and restore the classification accuracy before drift; in gradual drift scenarios, the model can synchronously complete incremental updates following the change of data distribution without obvious accuracy degradation, and maintains a stable classification performance throughout the process, with excellent convergence stability.

4.4. Scenario Adaptation and Stability Test

The long-term operation test results show that the system has no memory leakage or performance degradation under continuous high load, demonstrating excellent engineering quality. The stable performance in high noise and high-frequency drift scenarios verifies the strong adaptability of the system to complex and changeable real environments.

4.4.1. Scenario Adaptation Test

For the four types of concept drift scenarios of the simulated datasets and the industry business scenarios of the real datasets, the system can stably complete the whole process of streaming data classification without manual intervention in core parameter adjustment. For different data characteristics such as binary classification, multi-class classification, binary features, high-dimensional features, and class imbalance, the system can automatically complete the adjustment of preprocessing strategies, embedding layer adaptation, loss function matching and sample optimization rules. The scenario adaptation capability fully meets the design objectives, and the system can be flexibly adapted to the streaming data classification business needs of different industries.

4.4.2. Long-term Operation Stability Test

The test results of the system running continuously at full load for 72 hours show that there is no program crash, memory leakage, or abnormal task interruption during the operation; the fluctuation range of the full-process prediction accuracy is stably less than 3%, and there is no continuous accuracy degradation; the memory occupation is stably controlled within 4GB, the peak CPU usage does not exceed 60%, and the peak GPU usage does not exceed 70%. The system resource occupation is controllable, and it has the capability of long-term stable operation in the production environment.

4.4.3. Robustness Test in Extreme Scenarios

The system can still maintain stable operation for extreme scenarios such as high-noise data, high-frequency drift, and small sample size. In high-noise scenarios, the system filters noise interference through the sample screening strategy, and the attenuation of classification accuracy is controlled within 8%; in high-frequency drift scenarios, the drift detection and model update links have no delay or jamming, and no overfitting occurs; in small sample size scenarios, the convergence effect and classification accuracy of the model are guaranteed through the historical high-confidence sample augmentation strategy, and there are no abnormal conditions such as task jamming and model non-convergence, showing extremely strong scenario robustness.

5. Typical Application Scenarios and Future Prospects of the System

Based on the general architecture design of the system, its application boundary can be easily extended to more real-time sensitive fields such as financial risk control and smart healthcare. The in-depth analysis of typical scenarios not only demonstrates the practical value of the system, but also provides a clear reference path for subsequent industry customized development.

5.1. Typical Application Scenarios of the System

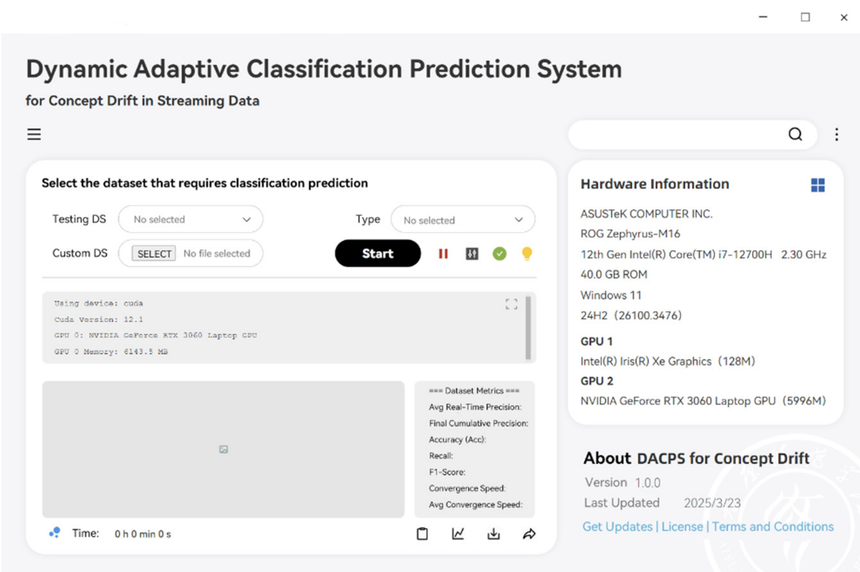

Fig 6. Main interface of the DACPS

With the universal and full-process adaptive processing capability for concept drift as the core, the DACPS can be seamlessly adapted to the streaming data classification business scenarios in many fields such as industry, transportation, and environmental monitoring, and provides an end-to-end solution that can be directly implemented according to the streaming data characteristics and business needs of different scenarios. As shown in Fig 6, the system provides an intuitive visual interactive interface, through which users can configure the dataset and drift type, start the classification and prediction task, and view the processing progress and evaluation indicators in real time. The "Hardware Information" card on the right side of the interface displays the current device configuration, the "Test Dataset" drop-down menu supports the selection of preset datasets, and the "Drift Type" drop-down menu supports the selection of four concept drift types.

5.1.1. Industrial Production Equipment Fault Prediction Scenario

In industrial production scenarios such as machining and intelligent manufacturing, production line equipment continuously generates multi-dimensional time-series streaming data including temperature, vibration, current, rotational speed, and product rejection rate. Factors such as equipment aging, working condition adjustment, and changes in raw material characteristics will cause sudden or gradual drift of data distribution, leading to continuous degradation of fault prediction accuracy of traditional fixed models. The DACPS can access the real-time streaming data of equipment operation status, capture the change of data distribution in real time through the adaptive window cumulative sum algorithm, dynamically adjust the sample screening strategy and fault classification model, and realize real-time classification of equipment health status[12] and early warning of potential faults. In the actual deployment test on a machining production line of auto parts, the classification and prediction accuracy of equipment faults of the system stably reaches more than 92%, and it can give early warning of more than 85% of unplanned shutdown faults 24 hours in advance, which effectively helps enterprises reduce equipment operation and maintenance costs, optimize production scheduling processes, and improve the continuous operation capability of production lines.

5.1.2. Intelligent Traffic Flow Prediction and Congestion Warning Scenario

In urban intelligent traffic management and control scenarios, roadside radars, bayonet cameras, intelligent on-board units and other devices continuously generate real-time streaming data such as vehicle position, driving speed, road section traffic flow, and intersection queue length. Scenarios such as morning and evening peak tidal changes, traffic accidents, holiday travel, and temporary traffic control will cause sudden, gradual and recurrent drift of traffic flow distribution, and traditional static prediction models are difficult to adapt to the dynamic changes of traffic flow. The DACPS can access multi-source traffic flow data in real time, automatically identify different types of concept drift and dynamically adjust the traffic state classification model, so as to realize accurate classification and early warning of road congestion status. In the actual test on the core arterial road of a provincial capital city, the classification and prediction accuracy of traffic congestion status of the system reaches more than 88%, and it can give early warning of more than 90% of non-periodic congestion events 15 minutes in advance, which provides real-time decision support for traffic signal timing optimization and congestion diversion scheduling, effectively improves the traffic efficiency of road sections, and alleviates traffic congestion in the core urban areas.

5.1.3. Meteorological Early Warning and Pollutant Concentration Prediction Scenario

In regional meteorological and environmental monitoring scenarios, ground meteorological stations and air quality monitoring stations continuously generate time-series streaming data such as temperature, humidity, air pressure, wind speed, PM2.5, PM10, and concentrations of various gaseous pollutants. Factors such as seasonal changes, extreme weather, and sudden pollution discharge will cause concept drift in the distribution of meteorological elements and pollutant concentrations, leading to a significant decline in the classification and prediction accuracy of traditional forecasting models. The DACPS can access the environmental monitoring streaming data of multiple stations in real time, adaptively track the dynamic changes of data distribution, dynamically optimize the classification and prediction model of meteorological grade and air quality grade, and realize early warning of severe weather and heavy pollution weather. In the actual application test of a regional environmental monitoring network, the classification and prediction accuracy of air quality grade of the system reaches more than 86%, and the average advance of heavy pollution weather early warning is increased to 48 hours, which can provide reliable data and model support for regional meteorological forecasting, emergency control of heavy pollution weather, and effectiveness evaluation of environmental governance.

5.2. Research Conclusion

Aiming at the core industry pain points in streaming data classification scenarios, including model performance degradation caused by concept drift, insufficient engineering implementation capability of existing technical solutions, and difficulty in balancing real-time performance and classification accuracy, this paper designs and implements a Dynamic Adaptive Concept Drift Processing System (DACPS). The system adopts a hierarchical modular architecture, constructs five core modules including streaming data preprocessing, real-time drift perception, dynamic sample optimization, incremental model training, and full-process management and control, and forms a full-process adaptive closed-loop processing link driven by drift perception results. It is compatible with multiple types of concept drift and complex streaming data scenarios, and effectively solves the engineering implementation problems such as difficult drift adaptation, class imbalance, and low model update efficiency. Through multi-dimensional functional tests, performance special tests and real scenario-based verification, the system has complete functions, excellent performance, stable operation, good universality and scalability. It can provide a complete solution that can be directly implemented for streaming data classification business in many fields such as industry, transportation, and environmental monitoring, filling the system-level gap between concept drift algorithm research and industrial engineering implementation.

5.3. Future Prospects

Based on the current system R&D achievements and implementation practice experience in this paper, the follow-up optimization and upgrading of the system will be carried out in four directions around the capability expansion and scenario adaptation of the system: First, optimize the distributed processing architecture of the system, support multi-node parallel processing of massive streaming data, and improve the throughput capacity and concurrent processing performance of the system in ultra-large-scale streaming data scenarios. Second, expand the system's adaptation capability for time-series streaming data and graph streaming data[13], enrich the types of streaming data that the system can process, and expand the application boundary of the system. Third, promote the lightweight transformation of the system, optimize the efficiency of model inference and resource occupation, realize the deployment and implementation of the system on edge devices, and adapt to edge computing scenarios such as industrial Internet of Things and vehicle-road collaboration. Fourth, improve the model interpretability analysis module, realize in-depth mining and visual interpretation of the causes of concept drift, and provide users with more instructive business decision support capabilities.

ISSN: 2688-8645

References

[1] LIU T T, GAO Y, ZHU J W. A Survey of Online Learning for Streaming Data Classification[J]. Journal of Software, 2020, 31(4): 20-25.

[2] ZHOU P, GUO Y, YU H, et al. Concept Evolution Detecting over Feature Streams[J]. ACM Transactions on Knowledge Discovery from Data, 2024, 18(8): 1-20.

[3] PALI A S, JAAFAR J, HASHMANI M. Performance Degradation of Multi-class Classification Model Due to Continuous Evolving Data Streams[C]. Springer, Cham, 2021.

[4] LIANG B, LI G H, DAI C L. G-Mean Weighted Classification Method for Concept Drift and Imbalanced Data Streams[J]. Journal of Computer Research and Development, 2022, 59(12): 2844-2857.

[5] GAI Y, MENG K, WANG X, et al. Online learning for streaming data classification in nonstationary environments[J]. Statistical Analysis & Data Mining, 2024, 17(2): 1-15.

[6] MU D L, HAN M, LI A, et al. A Survey of Classification Methods for Complex Data Streams with Concept Drift[J]. Journal of Computer Applications, 2023, 43(6): 1664-1675.

[7] GANTI V, GEHRKE J, RAMAKRISHNAN R, et al. Mining Data Streams under Block Evolution[J]. ACM SIGKDD Explorations Newsletter, 2002, 3(2): 1-10.

[8] FANG J L, WANG P, GUO B, et al. Online Self-Learning Event Detection Model Updating Method and Device[P]. CN201610041602.0. CN106991072B, 2026-03-03.

[9] ZHANG P, LI P X, REN Y, et al. A Survey of Distributed Streaming Processing Technology for Big Data[J]. Journal of Computer Research and Development, 2014(S2): 9-15.

[10] WINDISCH D, KELLING J, JUCKELAND G, et al. Real-time data processing for ultrafast X-ray computed tomography using modular CUDA based pipelines[J]. Computer Physics Communications, 2023: 108719. DOI:10.1016/j.cpc.2023.108719.

[11] LI H, LI F H, SHI X Y, et al. A Survey of Access Control Research for Secure Data Cross-domain Transmission[J]. Journal of Communications, 2025, 46(4): 1-10.

[12] ZHOU X, ZHANG Q J. An Artificial Intelligence-Driven Fault Prediction System for Industrial Internet of Things Devices[P]. CN202510890587.6. CN120416067A, 2026-03-03.

[13] SAMADI N, TANHA J, JALILI M, et al. Graph theory-based semi-supervised self-training for data stream classification and emerging class detection[J]. Information Sciences, 2025, 698: 121762.