JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

APPLICATION OF MACHINE LEARNING IN DATA STREAM MINING FOR DYNAMIC FEATURES SELECTION AND MODEL UPDATING

[1]Jimit Patel, [2]Meet Patel
[1]Staff Software Engineer, [2]Engineering Manager
[1]Research and Development, [2]Enterprise Business Intelligence,
[1]Very Good Security, San Francisco, USA, [2]Comcast Cable, Philadelphia, USA

Abstract: The continuous influx of data in real-time applications necessitates advanced methodologies for data stream mining, particularly focusing on dynamic feature selection and model updating. This paper explores the application of machine learning techniques in data stream environments, aiming to address the challenges posed by the non-stationary nature of data streams. Dynamic feature selection methods, which identify relevant features in real-time, are crucial for maintaining the efficiency and accuracy of predictive models. Concurrently, model updating strategies ensure that machine learning models evolve with incoming data, adapting to changes and preserving their predictive performance. We review various approaches, including data life-aware model updating, hybrid batch-stream processing, and automated machine learning, and discuss their applications in fields such as IoT data analytics and network traffic management. By integrating these advanced techniques, this study provides a comprehensive overview of maintaining robust and accurate models in dynamic data stream environments.

IndexTerms **– Dynamic Feature Selection, Data Stream Mining, Concept Drift Detection, Real-time Data Analytics.**

1. Introduction

Data stream mining is a crucial area within machine learning, focusing on the real-time analysis and processing of continuous data flows. Unlike traditional batch processing, which handles static datasets, data stream mining deals with dynamically generated data that can be potentially unbounded. This field addresses the challenges of continuously evolving data, which requires adaptive algorithms capable of real-time feature selection and model updating to maintain the accuracy and relevance of machine learning models.

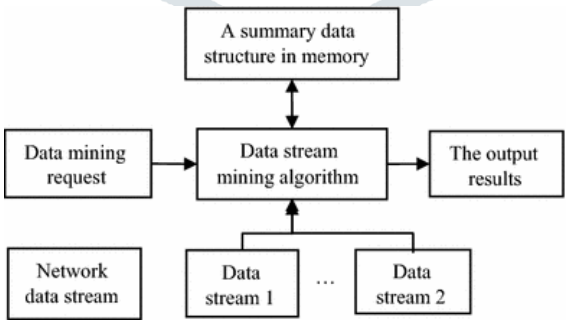

Figure 1.1: Data stream mining algorithm [23]

1.1 Challenges in Data Stream Mining

a) Dynamic Nature of Data Streams

The dynamic nature of data streams means that data distributions can change over time, a phenomenon known as concept drift. This drift poses significant challenges for traditional machine learning models, which are typically trained on static datasets and

may not adapt well to changes in the underlying data distribution. To address this, data stream mining requires algorithms that can dynamically update both features and models in response to new data.

b) Scalability and Efficiency

Another key challenge is scalability. Data streams can generate massive amounts of data at high velocity, necessitating efficient algorithms that can process data in real-time without excessive computational overhead. Traditional batch processing methods often fail in this regard because they are not designed to handle continuous, high-volume data streams.

2. Dynamic Feature Selection in Data Stream Mining

Dynamic feature selection is a pivotal process in data stream mining that addresses the challenge of managing high-dimensional streaming data. The dynamic nature of data streams means that the relevance of features can fluctuate over time, making it crucial to continually adapt the feature set to maintain model performance. Two prominent techniques in dynamic feature selection are online feature selection and incremental feature grouping.

2.1 Online Feature Selection

Online feature selection involves the continuous evaluation and updating of the feature set as new data points arrive. This method is designed to adapt to the dynamic nature of streaming data by selecting the most relevant features in real-time, thereby enhancing model accuracy and efficiency.

An innovative online streaming feature selection method that groups feature incrementally based on their relevance [8]. This approach significantly reduces computational overhead while maintaining high model accuracy. The method evaluates the importance of each feature dynamically, ensuring that only the most pertinent features are retained for model training and prediction, thus avoiding the pitfalls of feature redundancy and irrelevance.

A comprehensive review of feature selection techniques for online streaming high-dimensional data [2]. They highlight various state-of-the-art methods, including those that employ statistical and machine learning-based criteria to select features dynamically. These techniques help in handling the evolving nature of data streams by continuously adapting the feature set to reflect the most current data distribution.

2.2 Incremental Feature Grouping

Incremental feature grouping is another effective technique for dynamic feature selection. This method involves grouping features incrementally, which allows the model to adapt efficiently to changes in the data stream. By managing the high dimensionality of the data, incremental feature grouping ensures that the model remains scalable and efficient over time.

A dynamic feature selection approach that integrates intelligent model serving for hybrid batch-stream processing [3]. This method not only adapts to the changing feature relevance but also optimizes the processing of large-scale data streams by combining batch and stream processing advantages.

An extensive review of feature subset selection techniques for data and feature streams [4]. They discuss various algorithms and methodologies that incrementally group features based on their evolving importance. This incremental approach helps in maintaining a manageable feature set, thereby improving the model's scalability and computational efficiency.

3. Model Updating Strategies

Model updating strategies are essential for maintaining the performance of machine learning models in dynamic environments where data distributions may shift over time. These strategies are crucial to ensure that models remain accurate and relevant as they encounter new data patterns and concept drifts. Here, we explore three prominent model updating strategies: Data Life Aware Model Updating, Hybrid Batch-Stream Processing, and Automated Machine Learning (AutoML).

3.1 Data Life Aware Model Updating

The data life aware model updating strategy focuses on updating the model based on the lifespan of data points. This approach prioritizes the retention of the most relevant information while discarding outdated data, thereby enhancing model performance over time. Method to address the challenge of maintaining model accuracy in environments where data can quickly become obsolete [1]. By continuously evaluating the relevance of data points, the model can dynamically adjust to reflect the most current and pertinent information, ensuring that outdated or less significant data does not negatively impact model performance.

For instance, in scenarios where data streams are subject to rapid changes, such as financial markets or real-time sensor data, this strategy proves invaluable. It allows models to remain adaptive and responsive to new trends and patterns, maintaining high levels of accuracy and reliability.

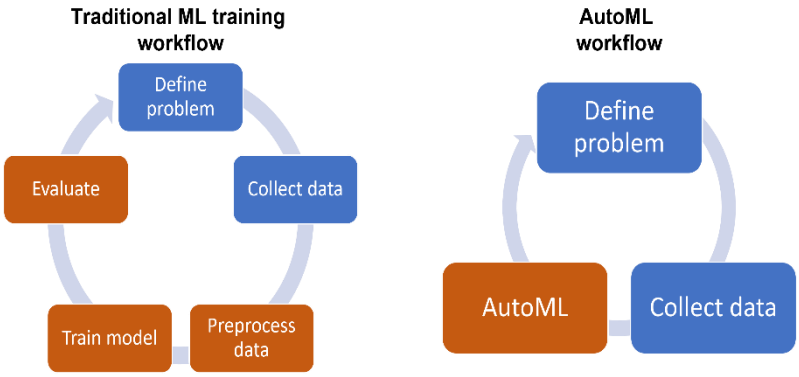

Figure 3.1: Data Life Aware Model Updating

3.2 Hybrid Batch-Stream Processing

Hybrid batch-stream processing combines the advantages of traditional batch processing with the real-time capabilities of stream processing. This method, proposed by Pishgoo et al. (2022) [3], allows for efficient handling of large volumes of data while ensuring timely updates to the model. The hybrid approach leverages the robustness of batch processing to handle historical data and the agility of stream processing to integrate new data points continuously.

This strategy is particularly effective in environments where both historical and real-time data are crucial for decision-making. For example, in predictive maintenance for industrial equipment, historical data provides context and baseline performance metrics, while real-time data enables immediate detection and response to anomalies. By integrating both data types, the model can offer more comprehensive insights and predictions, balancing long-term trends with immediate variations .

3.3 Automated Machine Learning (AutoML)

Automated Machine Learning (AutoML) is a powerful approach that automates the process of model selection, hyperparameter tuning, and feature engineering. The application of AutoML in dynamic environments, emphasizing its ability to adapt and optimize models continuously [5]. AutoML systems leverage machine learning algorithms to automatically identify the best model architecture and configuration for the given data, reducing the need for manual intervention and expertise.

In dynamic environments where data characteristics can change rapidly, AutoML offers significant advantages. It can quickly retrain and adjust models in response to new data, ensuring that the most effective and up-to-date model is always in use. This capability is particularly useful in applications such as IoT data analytics, where data streams from various sensors can exhibit diverse and evolving patterns. By automating the model adaptation process, AutoML enhances the scalability and robustness of machine learning systems in real-time settings.

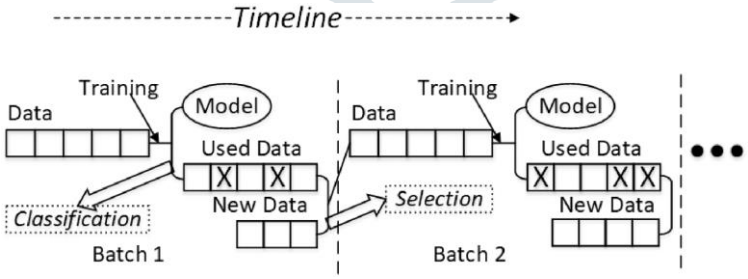

Figure 4.1: Automated Machine Learning [22]

4. Concept Drift Detection

Concept drift refers to the changes in the statistical properties of the target variable that the model is predicting, which can significantly affect model performance. Detecting and adapting to concept drift is crucial for the success of data stream mining. Here are detailed insights into two key approaches for handling concept drift:

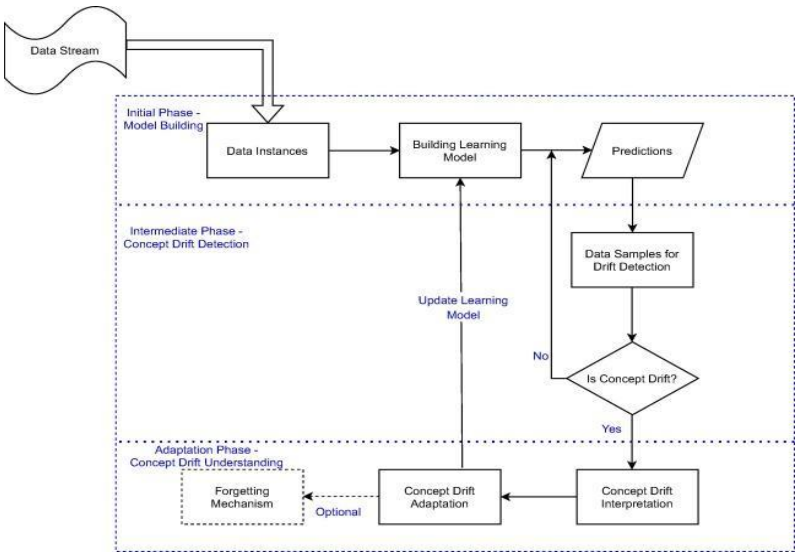

Figure 4.1: Drift Detection [21]

4.1 Recurring Drift Detection

Recurring drift detection involves identifying and responding to drifts that repeat over time. Methods for detecting recurring drifts and selecting appropriate models based on past occurrences of similar drifts [13]. This approach leverages historical data to recognize patterns in the drift, allowing for the anticipation and adaptation to future changes. By maintaining a repository of past drifts and the corresponding model adjustments, the system can quickly switch to previously successful models when similar drifts occur again. This method enhances model robustness and stability, ensuring that performance remains consistent even in the face of recurring changes in the data stream.

4.2 Dynamic Multi-Objective Evolutionary Algorithms

A framework based on dynamic multi-objective evolutionary algorithms (MOEAs) to handle feature drifts in data streams [17]. This method allows for continuous adaptation of the feature set and the model to ensure optimal performance. The framework operates by simultaneously optimizing multiple objectives, such as accuracy and processing time, under changing data conditions. By employing evolutionary strategies, the algorithm can evolve the feature selection and model parameters over time, ensuring that the system adapts to new data distributions effectively. The dynamic nature of MOEAs makes them particularly well-suited for environments with frequent and unpredictable changes, as they can explore a wide range of solutions and converge on the most effective configurations for the current data stream.

5. Methodology

The methodology for applying machine learning in data stream mining involves several key steps, each essential for maintaining the integrity and accuracy of the models in dynamic environments.

Data Preprocessing: Initial preprocessing of data streams is critical for handling noise, missing values, and outliers, which ensures data integrity. Effective preprocessing techniques are necessary to prepare the data for further analysis and model training. These steps include filtering out irrelevant data, imputing missing values, and normalizing the data to ensure consistency across the stream [10].

Dynamic Feature Selection: Implementing algorithms for dynamic feature selection helps identify the most relevant features in real-time, improving model performance and reducing computational overhead. Dynamic feature selection algorithms [7], enable the model to focus on the most informative features, thereby enhancing predictive accuracy and efficiency.

Model Updating: Continuously updating the machine learning model as new data arrives ensures the model adapts to changes in data distribution, maintaining its predictive accuracy. A data life-aware model updating strategy that adjusts the model based on the lifecycle of the incoming data, ensuring the model remains relevant over time [18].

Handling Concept Drift: Detecting and adapting to concept drift is essential for maintaining the relevance of the model in a changing environment. Concept drift occurs when the statistical properties of the target variable change over time, which can degrade model performance. Techniques such as incremental learning and ensemble methods are often used to address this challenge. An extensive review of methods for managing recurring concept drifts, emphasizing the need for adaptive strategies to maintain model accuracy in the presence of evolving data streams [6].

6. Applications of Dynamic Feature Selection and Model Updating in Data Stream Mining

IoT Data Analytics

The application of dynamic feature selection and model updating techniques in Internet of Things (IoT) environments facilitates the efficient processing and analysis of vast amounts of data generated by IoT devices. These techniques enable real-time decision-making and predictive maintenance, crucial for optimizing performance and preventing potential failures. By leveraging automated machine learning (AutoML) frameworks [5], IoT systems can adaptively select relevant features and update models to reflect the latest data trends, thus enhancing their responsiveness and accuracy.

Network Traffic Management

In software-defined networks, the implementation of dynamic feature selection and model updating significantly improves the classification and management of network traffic. This leads to enhanced network performance and security. Improved feature selection methods could boost the accuracy and efficiency of stream traffic classification, ensuring that network operations remain robust and secure against evolving threats [7].

Real-Time Data Analysis

Industries such as finance, healthcare, and manufacturing benefit immensely from real-time data analysis enabled by dynamic feature selection and model updating. These techniques provide timely insights and facilitate proactive measures, which are critical in dynamic and high-stakes environments. For instance, A hybrid batch-stream processing framework that optimizes feature extraction and model deployment, proving particularly beneficial for real-time analytics in these sectors [11]. Furthermore, the ability to handle concept drift [9], ensures that models remain accurate and relevant despite changes in data patterns, which is essential for maintaining the reliability of real-time decision-making systems.

7. Methodologies for Dynamic Feature Selection and Model Updating in Data Stream Mining

Table 7.1: Summary of Methodologies for Dynamic Feature Selection and Model Updating in Data Stream Mining

References	Methodology/Dataset	Tasks/Objectives	Techniques/Approaches	Accuracy/Performance
Agrahari S, Singh AK. (2022) [12]	Concept Drift Detection	Feature Selection and Drift Detection	Literature Review on Data Stream Mining	Various accuracies reported
Li P, Wu M, He J, Hu X. (2021) [13]	Ensemble Classification for Data Streams	Model Selection with Unlabeled Data	Recurring Drift Detection and Ensemble Methods	High accuracy in different scenarios
Almusallam N, Tari Z, Chan J, et al. (2021) [14]	Unsupervised Feature Selection	Dynamic Feature Selection	Unsupervised Methods for Feature Selection	Improved feature relevance
Li Y, Zhang M, Wang W. (2018) [15]	Incremental High-Order Deep Learning	Real-time Stream Analysis	Incremental Learning with Deep Models	Effective for real-time analysis
Gomes HM, Bifet A, Read J, et al. (2017) [16]	Adaptive Random Forests	Evolving Data Stream Classification	Adaptive Random Forest Algorithms	High adaptability and accuracy
Din SU, Shao J, Kumar J, et al. (2021) [18]	Novel Class Detection	Data Stream Classification	Review and Comparison of Novel Class Detection Methods	Various accuracies reported
Singh MN, Khaiyum S. (2021) [19]	Meta-Learning with Concept Drift	Continuous Learning and Classification	Optimized Weight Updated Meta-Learning	Enhanced accuracy with concept drift
Benczúr AA, Kocsis L, Pálovics R. (2018) [20]	Online Machine Learning	Big Data Streams Analysis	Online Learning Algorithms for Big Data	Effective for large-scale data streams

8. CONCLUSION

The application of machine learning in data stream mining, particularly through dynamic feature selection and model updating, offers significant advantages in handling the continuous and evolving nature of data streams. By adopting these techniques, organizations can enhance their data processing capabilities, maintain model accuracy, and derive actionable insights in real-time. This is particularly crucial in domains such as IoT data analytics, where the ability to process and analyze vast amounts of data generated by IoT devices in real time can enable timely decision-making and predictive maintenance. In the realm of network traffic management, dynamic feature selection and model updating can significantly improve the classification and management of network traffic, leading to enhanced network performance and security.

Furthermore, industries such as finance, healthcare, and manufacturing benefit immensely from real-time data analysis enabled by these advanced techniques. In finance, for example, the ability to process streaming data allows for the detection of fraudulent activities as they occur, providing a robust mechanism for safeguarding financial transactions. In healthcare, real-time analysis can facilitate timely diagnosis and treatment by continuously monitoring patient data. In manufacturing, dynamic feature selection and model updating can optimize production processes, predict equipment failures, and minimize downtime.

By leveraging machine learning for dynamic feature selection and model updating, organizations can stay ahead in a data-driven world, ensuring that their systems are adaptive, resilient, and capable of making informed decisions swiftly. This not only improves operational efficiency but also enhances the ability to respond to emerging challenges and opportunities in various sectors.

References

[1] Rang W, Yang D, Cheng D, Wang Y. Data life aware model updating strategy for stream-based online deep learning. IEEE Transactions on Parallel and Distributed Systems. 2021 Apr 8;32(10):2571-81.

[2] Zaman EA, Mohamed A, Ahmad A. Feature selection for online streaming high-dimensional data: A state-of-the-art review. Applied Soft Computing. 2022 Sep 1;127:109355.

[3] Pishgoo B, Azirani AA, Raahemi B. A dynamic feature selection and intelligent model serving for hybrid batch-stream processing. Knowledge-Based Systems. 2022 Nov 28;256:109749.

[4] Villa-Blanco C, Bielza C, Larrañaga P. Feature subset selection for data and feature streams: a review. Artificial Intelligence Review. 2023 Oct;56(Suppl 1):1011-62.

[5] Yang L, Shami A. IoT data analytics in dynamic environments: From an automated machine learning perspective. Engineering Applications of Artificial Intelligence. 2022 Nov 1;116:105366.

[6] Suárez-Cetrulo AL, Quintana D, Cervantes A. A survey on machine learning for recurring concept drifting data streams. Expert Systems with Applications. 2023 Mar 1;213:118934.

[7] Eldhai AM, Hamdan M, Abdelaziz A, Hashem I, Marsono MN, Hamzah M, Jhanjhi NZ. Improved Feature Selection and Stream Traffic Classification Based on Machine Learning in Software-Defined Networks. IEEE Access. 2024 Feb 26.

[8] Al Nuaimi N, Masud MM. Online streaming feature selection with incremental feature grouping. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2020 Jul;10(4).

[9] Bifet A. Adaptive stream mining: Pattern learning and mining from evolving data streams. Ios Press; 2010.

[10] Lughofer E. On-line active learning: A new paradigm to improve practical usability of data stream modeling methods. Information Sciences. 2017 Nov 1;415:356-76.

[11] Parker BS, Khan L, Bifet A. Incremental ensemble classifier addressing non-stationary fast data streams. In2014 IEEE International Conference on Data Mining Workshop 2014 Dec 14 (pp. 716-723). IEEE.

[12] Agrahari S, Singh AK. Concept drift detection in data stream mining: A literature review. Journal of King Saud University-Computer and Information Sciences. 2022 Nov 1;34(10):9523-40.

[13] Li P, Wu M, He J, Hu X. Recurring drift detection and model selection-based ensemble classification for data streams with unlabeled data. New generation computing. 2021 Aug;39(2):341-76.

[14] Almusallam N, Tari Z, Chan J, Fahad A, Alabdulatif A, Al-Naeem M. Towards an unsupervised feature selection method for effective dynamic features. IEEE Access. 2021 May 21;9:77149-63.

[15] Li Y, Zhang M, Wang W. Online real-time analysis of data streams based on an incremental high-order deep learning model. IEEE Access. 2018 Nov 28;6:77615-23.

[16] Gomes HM, Bifet A, Read J, Barddal JP, Enembreck F, Pfharinger B, Holmes G, Abdessalem T. Adaptive random forests for evolving data stream classification. Machine Learning. 2017 Oct;106:1469-95.

[17] Sahmoud S, Topcuoglu HR. A general framework based on dynamic multi-objective evolutionary algorithms for handling feature drifts on data streams. Future Generation Computer Systems. 2020 Jan 1;102:42-52.

[18] Din SU, Shao J, Kumar J, Mawuli CB, Mahmud SH, Zhang W, Yang Q. Data stream classification with novel class detection: a review, comparison and challenges. Knowledge and Information Systems. 2021 Sep;63:2231-76.

[19] Singh MN, Khaiyum S. Enhanced data stream classification by optimized weight updated meta-learning: Continuous learning-based on concept-drift. International Journal of Web Information Systems. 2021 Dec 1;17(6):645-68.

[20] Benczúr AA, Kocsis L, Pálovics R. Online machine learning in big data streams. arXiv preprint arXiv:1802.05872. 2018 Feb 16.

[21] S. Agrahari and A. K. Singh, "Concept Drift Detection in Data Stream Mining: A literature review," Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 10, pp. 9523-9540, 2022. doi: 10.1016/j.jksuci.2021.11.006.

[22] https://learn.microsoft.com/en-us/dotnet/machine-learning/automated-machine-learning-mlnet

[23] https://www.researchgate.net/figure/Data-stream-mining-algorithm-architecture-diagram_fig1_318461827