UDC 004.75

M.V. TALAKH, Yu.O. USHENKO, O.V. KADUK, M.Yu. MAKSYMOVYCH

BALANCING EFFICIENCY AND ACCURACY: INCREMENTAL LEARNING AS A KEY TO BIG DATA PROCESSING

Yuriy Fedkovych Chernivtsi National University, 2 Kotsjubynskyi Str. Chernivtsi, Ukraine
Vinnytsia National Technical University, Vinnytsia, Khmelnytskyi Shosse, 95, Ukraine

Анотація. У статті представлено комплексний огляд інкрементального навчання в контексті обробки великих даних. Розглянуто основні концепції, сучасні підходи та ключові аспекти інкрементального навчання. Проаналізовано переваги цього підходу для обробки великих обсягів даних, включаючи ефективне використання обчислювальних ресурсів, можливість обробки потокових даних у реальному часі та адаптивність до змін у даних. Досліджено основні обмеження та виклики, такі як проблема "катастрофічного забування", складність балансування нових та старих знань, залежність від порядку надходження даних та потенційна втрата точності. Представлено аналіз специфічних проблем, включаючи обробку концептуального дрейфу, незбалансованих класів та відсутніх ознак. Розглянуто застосування інкрементального навчання в різних галузях, включаючи аналітику даних, робототехніку, автономне водіння та розпізнавання активності. Запропоновано напрямки майбутніх досліджень для вирішення виявлених проблем та покращення ефективності інкрементального навчання у контексті великих даних.

Ключові слова: поступове навчання, великі дані, машинне навчання, потокова обробка даних, концептуальний дрейф, катастрофічне забування, адаптивні алгоритми, онлайн-навчання.

Abstract.. The article provides a comprehensive overview of incremental learning in the context of big data processing. The basic concepts, modern approaches, and key aspects of incremental learning are considered. The advantages of this approach for processing large amounts of data are analyzed, including the efficient use of computing resources, the ability to process streaming data in real time, and adaptability to changes in data. The main limitations and challenges, such as the problem of "catastrophic forgetting", the difficulty of balancing new and old knowledge, dependence on the order of data arrival, and potential loss of accuracy, are investigated. An analysis of specific problems is presented, including the handling of conceptual drift, unbalanced classes, and missing features. Applications of incremental learning in various fields, including data analytics, robotics, autonomous driving, and activity recognition, are discussed. We suggest directions for future research to address the identified problems and improve the effectiveness of incremental learning in the context of big data.

Ключові слова: incremental learning, Big Data, machine learning, streaming data processing, conceptual drift, catastrophic forgetting, adaptive algorithms, online learning.

DOI: **10.31649/1681-7893-2024-48-2-45-57**

INTRODUCTION

In the modern era of rapid technological development and exponential growth of data, machine learning is facing new challenges. Traditional learning methods that involve processing the entire data set simultaneously are becoming less and less effective in the face of a constant flow of new information. This is where incremental learning comes to the fore – an innovative approach that allows models to continuously adapt to new data without losing previously acquired knowledge. This article discusses the key aspects of incremental learning, its role in big data processing and deep learning, and explores the main challenges and prospects of this technology in the context of modern data processing needs.

Incremental learning has become an important area of research in machine learning and big data processing. According to a study published in Nature Machine Intelligence in December 2022 [1], incremental learning remains an important open problem in deep learning. The main difficulty is to allow models to gradually learn from non-stationary data streams without losing previously acquired knowledge.

As Luo et al. note [2], incremental learning (IL) refers to a learning system that can continuously learn new knowledge from new samples while retaining most of the previously learned knowledge. In fact, IL is a special scenario of machine learning technology.

Hu et al. [3] emphasize the importance of incremental learning for processing streaming data in various application scenarios, such as image classification, credit scoring, and user profiling. These scenarios require high classification accuracy and the ability to process data in real time.

Classical batch machine learning approaches, where all data is available at the same time, do not meet the requirements of processing huge amounts of data in real time. This leads to the accumulation of raw data. In addition, they do not integrate new information into already built models, but instead regularly rebuild new models from scratch, which is not only very time-consuming but also leads to potentially outdated models. To overcome this state of affairs, a paradigm shift to consistent data processing in a streaming mode is needed. This allows information to be utilized as soon as it becomes available, resulting in continuously updated models, and reduces data storage and maintenance costs.

Van de Ven et al. [1] distinguish three fundamental types or "scenarios" of supervised continuous learning:

1. Task-incremental learning:

- An algorithm learns to solve a set of different tasks successively.
- The task ID is known during testing.
- Suitable for situations where large data sets are naturally divided into separate tasks.

2. Domain-incremental learning:

- The structure of the problem remains the same, but the context or distribution of the input data changes.
- The domain identifier is unknown during testing.
- Useful for large data sets with variable collection or generation conditions.

3. Class-incremental learning:

- The algorithm consistently learns to distinguish between an increasing number of classes.
- It is necessary to distinguish between classes that have not been observed together.
- Important for large datasets where new classes appear over time.

Each of these scenarios has its own peculiarities and challenges that require different approaches to solve them.

Luo et al. [2] propose a slightly different approach to classifying incremental learning scenarios, dividing them into:

1. Incremental learning of instances - the number of categories is fixed, but the data in each category is expanded at each stage of training.
2. Incremental learning of classes - both the number of categories and the input data change.
3. Incremental learning of instances and classes - both instances and categories increase, which is the most common scenario in a real-world environment.

Anowar and Sadaoui [4] in their research presented at the IEEE International Conference on Systems, Man, and Cybernetics describe several key approaches to incremental learning:

a) Chunk-based incremental learning: Processing data sequentially in chunks or mini-batches to work with large data sets.

b) Neural network approaches:

- Incremental neural networks that can dynamically add nodes and layers.
- Self-organized incremental neural networks (SOINN).
- Approaches that combine multiple sub-networks trained on different data vats.

c) Conceptual drift processing: Methods to adapt to changes in the data distribution over time.

d) Balancing stability and plasticity: Approaches to preserve important past knowledge while adapting to new data.

e) Addressing class imbalance: Particularly important for tasks such as fraud detection.

f) Semi-supervised and unsupervised approaches: Methods that can use both labeled and unlabeled data.

g) Memory models: Developing efficient memory models to manage data retention and forgetting.

Hu et al. [3] propose an integrated classification model for incremental learning, which consists of two main components: a pre-trained (Pt) model and a novel online classifier based on truncated gradient and confidence weights (TGCW). This model aims to improve classification accuracy and convergence speed when working with streaming data.

Luo et al. [2] identify three main strategies for solving the problem of catastrophic forgetting in incremental learning:

1. Architectural strategy

2. Regularization strategy
3. Strategy of repetition and pseudo-repetition

Incremental and online algorithms are a natural fit for a streaming sequential data processing scheme because they continuously incorporate information into their model and traditionally aim for minimal processing time and space. Due to their ability to continuously perform large-scale real-time processing, they have recently gained more attention, especially in the context of Big Data.

Incremental algorithms are also very suitable for learning after the production phase, allowing devices to adapt to individual customer habits and environments. This is especially interesting for smart home products. Here, the main challenge is not large-scale processing, but rather continuous and efficient learning from a small amount of data.

Incremental learning for large language models (LLMs) Research in incremental learning for LLMs focuses on several key areas:

a) Addressing the problem of catastrophic forgetting.

b) Fine-tuning methods for updating some of the model parameters. c) Continuous learning for continuously updating models on new data.

d) Adaptive architectures for dynamic model expansion.

e) Improving computational efficiency.

f) Personalization of models without losing general knowledge.

g) Processing of streaming data in real time.

h) Addressing ethical issues such as model bias.

i) Multimodal learning that incorporates different types of data.

1. KEY ASPECTS AND APPROACHES IN INCREMENTAL LEARNING

Incremental learning has become a key area of research in artificial intelligence and machine learning, especially in the context of processing large data streams and adapting to changing conditions. This technique allows systems to gradually improve their performance by assimilating new information without the need for complete retraining. In this section, we will look at the main aspects, challenges, and applications of incremental learning in various industries.

Eisa et al. [5] in their study published in Computers, Materials & Continua focus on several key aspects of incremental learning for large data streams:

- High Rate Data Processing: Processing streaming data at a high rate of arrival.
- Concept Drift: Adapting to changing concepts.
- Ensemble Methods: Flexible updating of classification schemes.
- Incremental DNNs: Incremental versions of deep neural networks.
- Unbalanced Data: Processing of unbalanced data.
- Semi-supervised/Unsupervised Learning: Methods for semi-supervised and unsupervised learning.
- Scalability: Ensuring scalability when working with large amounts of data.
- Hyperparameter Optimization: Dynamic optimization of hyperparameters.
- Memory Management: Efficient memory management.

Continuing with the topic of key aspects of incremental learning, it is worth paying attention to the study by Wen and Zhu published in IEEE Access. Their work focuses on approaches based on pre-trained models, which is an important step in the development of this field. Wen and Zhu [6] in their study published in IEEE Access consider approaches to incremental learning based on pre-trained models:

- Extensible Networks: Networks that can be scaled up to accommodate new classes.
- Synthetic Samples: Generating synthetic samples of previous classes.
- Improved Loss Functions: Such as Learning without Forgetting (LwF).
- Pre-Trained Models: Using models previously trained on large datasets.
- Ensemble Methods: Including PEDCC-based approaches.

Complementing previous studies, Joshi and Kulkarni in their review published in the International Journal of Data Mining & Knowledge Management Process highlight additional key research areas in incremental learning. Joshi and Kulkarni [7] in their review published in the International Journal of Data Mining & Knowledge Management Process highlight several key research areas:

- Incremental Clustering: Methods for large databases.
- Supervised/Semi-Supervised Methods: Including ensemble approaches.
- Conceptual Drift Handling: Methods to manage conceptual drift.
- Application-Specific Solutions: Tailored to different domains.
- Hybrid Approaches: Combining different techniques.

- Selective Learning: To improve performance.
- Reinforcement Learning Integration: Techniques integrated with reinforcement learning.
- Query Formulation Improvement: Incremental improvement of query formulation.

To summarise the approaches and challenges discussed, we can identify the key areas of application and challenges faced by incremental learning algorithms.

1. Data Analytics and Big Data Processing. Bifet and Gavalda [8] examine an approach to processing big data in a streaming setting:
   - Use of limited-memory models for incremental learning
   - Adaptation to concept drift in data streams

The authors propose the Adaptive Windowing (ADWIN) method for detecting changes in data distribution and updating models incrementally.

2. Robotics.Thrun and Mitchell [9] investigate the problem of lifelong robot learning:
   - Autonomous control and adaptation to new tasks
   - Integration of previous experience with new knowledge

They propose an approach combining reinforcement learning and explanation-based learning for long-term knowledge accumulation in an incremental manner.

3. Autonomous Driving. Bojarski et al. [10] present an approach to training autonomous driving systems:
   - End-to-end learning from sensory inputs to steering commands
   - Incremental adaptation to diverse road conditions

The authors use deep convolutional neural networks for direct mapping of input images to steering commands, allowing for incremental updates as new data becomes available.

4. Activity Recognition. Lara and Labrador [11] review methods for human activity recognition:
   - Use of wearable sensors for continuous data collection
   - Processing of multidimensional time series data

They classify approaches based on sensor types, feature extraction methods, and classification algorithms, emphasizing the need for incremental learning in real-world applications.

5. Image and Video Processing. Ren et al. [12] propose the Faster R-CNN architecture for object detection:
   - Fast generation of region proposals
   - Real-time object classification

This architecture allows for incremental model updates as new data arrives, making it suitable for continuous learning in image processing tasks.

6. Automated Annotation. Snoek et al. [13] investigate methods for semantic video analysis:
   - Comparison of early and late fusion of multimodal features
   - Adaptive combination of different information sources

The authors propose an approach that allows for incremental updating of annotations as new data becomes available, facilitating continuous learning in video tagging tasks.

7. Anomaly Detection. Chandola et al. [14] provide a survey of anomaly detection methods:
   - Classification of approaches based on input data type
   - Adaptation to changes in normal system behavior

They discuss incremental methods that can update models of normal behavior in real-time, essential for continuous monitoring and anomaly detection.

8. Tasar et al. [15], in their study on incremental learning for semantic segmentation of large-scale remote sensing data, address specific challenges:

- Data Storage Limitations: Working with limited data storage.
- Heterogeneous Data Sources: Processing various data sources.
- Unbalanced Classes: Addressing the problem of unbalanced classes.

The authors describe several modern approaches:

- Network Architecture Modifications.
- Rehearsal Methods.
- Knowledge Distillation.
- Parameter Regularization.
- Generative Data Reproduction.

They define an incremental learning algorithm as one that generates a sequence of models h1, h2, ..., ht on a given training data stream s1, s2, ..., st. In this case, si is the labeled training data si = (xi, yi) ∈ Rn × {1, ..., C}, and hi : Rn → {1, ..., C} is a model function that depends solely on hi-1 and the last p examples si, ..., si-p, where p is strictly bounded.

In each of these areas, incremental learning can solve specific problems, such as

- Processing data that arrives at high speed
- Adapting to conceptual drift
- Dealing with unbalanced data
- Data storage limitations
- Processing heterogeneous data sources

Defining and Addressing the Challenges of Incremental Learning:

- The model should adapt gradually, i.e. hi+1 is built on the basis of hi without complete retraining.
- Retention of previously acquired knowledge without the effect of catastrophic forgetting.
- Only a limited number p of training examples are allowed to be stored.
- Processing of incoming data at high speed.
- Adaptation to conceptual drift.
- Working with unbalanced data.
- Processing of heterogeneous data sources.

The choice of algorithm should be based on the specific requirements of the task at hand, as there is no one-size-fits-all solution. Various incremental learning algorithms have been developed, each with its strengths and weaknesses. However, there is a lack of comparative studies that experimentally evaluate the most popular methods based on the most relevant criteria.

2. ADVANTAGES OF INCREMENTAL LEARNING FOR BIG DATA PROCESSING

1. Efficient use of computing resources:

- Updating models without complete retraining: Incremental learning allows you to update models gradually by processing new data without the need to retrain on the entire dataset. This reduces the load on computing resources and speeds up the learning process [1].

- Reduced computing power requirements: Incremental learning significantly reduces computing requirements as the model is trained on new data in stages. This is especially critical for large datasets where full retraining would be extremely costly [3].

- Training on small portions of data: Models can process small portions of data that arrive sequentially. This avoids the need to process the entire data set at once, which reduces the need for large computing power and memory [4].

- Importance for constrained systems and real-time: Incremental learning is critical for systems with limited computational resources and for real-time data processing, as it allows for model adaptation without significant computational resource costs and reduces data processing latency [5].

2. The ability to process streaming data in real time:

- Ideal application for streaming data: Incremental learning is particularly effective for processing streaming data that arrives continuously. It allows models to adapt to new data that is constantly emerging without having to be retrained on the entire data set [3].

- Real-time updates: Models can be updated in real time, allowing new data to be integrated without delay. This allows to keep the model up-to-date and quickly take into account new information flows [5].

- Critical for various applications: The ability to process data in real time is critical for applications such as social media analysis, financial forecasting, and anomaly detection systems. Incremental learning allows you to effectively respond to changes in the data and maintain high accuracy of classification and prediction [2].

- Quick response to changes: With incremental learning, models can quickly respond to changes in data and adapt to new conditions. This is especially important in dynamic environments where conditions can change quickly and without warning [1].

3. Adaptability to changes in data:

- Easy adaptation to changes in the data distribution: Incremental models are able to quickly adapt to changes in the data distribution that occur over time. This allows them to remain relevant in environments where data is constantly changing [1].

- Importance for dynamic environments: Adaptability is especially important for dynamic environments where data characteristics can change significantly and frequently. Incremental learning allows models to respond to these changes without the need for complete retraining [2].

- Taking into account new trends and patterns: Models can take into account new trends, patterns, or classes that appear in the data, thus ensuring accuracy and relevance. This allows to maintain a high level of classification and prediction even when conditions change [3].

- Relevance without retraining: Incremental learning allows you to keep the model up-to-date without the need for complete retraining. This significantly reduces computational costs and provides more efficient model updates in response to data changes [5].

4. Lower memory requirements:

- Processing data in chunks: Incremental learning allows data to be processed in chunks, which significantly reduces memory requirements. This ensures efficient use of memory resources, as the data does not need to be loaded into memory in its entirety [4].

- No need to store the entire data set: Models do not need to store the entire data set at once, which reduces the memory load. This is especially important when working with large datasets, where storing the entire volume in memory would be impossible or too costly [3].

- Working with large data sets: Incremental learning allows you to work efficiently with data sets that exceed the amount of available RAM. This allows the model to process and learn from data that cannot be fitted into memory at once, enabling the processing of large amounts of information [5].

- Useful for devices with limited resources: This approach is particularly useful for devices with limited memory resources, such as mobile devices or embedded systems. Incremental learning allows implementing powerful algorithms without the need for large amounts of RAM [1].

5. Possibility of continuous learning:

- Continuous improvement of the model: Incremental learning allows the model to be continuously improved as new data becomes available. This provides a continuous learning process where the model adapts to new information flows and improves its performance in real time [3].

- No periodic retraining: Thanks to the incremental approach, models can continuously improve their performance without the need for periodic full retraining. This reduces the cost of computing resources and the time it usually takes to retrain a model from scratch [1].

- Lifecycle adaptation: The model can learn throughout the system's lifecycle, constantly adapting to new conditions and changing its behavior in response to new data. This allows the system to maintain relevance and efficiency in the long run [2].

6. Saving time and resources:

- Reducing the need for full retraining: Incremental learning significantly saves time and computing resources, as it does not require a complete retraining of the model from scratch. Models are updated based on new data, which reduces computational load and speeds up the learning process [5].

- Importance for large organizations: This is especially important for large organizations that work with huge amounts of data. Incremental learning allows you to effectively manage resources and reduce computing costs, which is critical for companies with large-scale operations [1].

- Faster implementation of updated models: Thanks to the incremental approach, updated models can be implemented more quickly in the production environment. This allows you to quickly respond to changes in data and adapt to new conditions without significant delays [2].

7. Support for distributed systems:

- Adaptation to distributed data sources: Incremental learning is well suited for distributed systems where data may come from different sources. Models can learn from local data coming from different nodes, reducing the need for centralized storage and processing of all data [4].

- Local learning with central model updates: Each node in a distributed system can learn from its local data and transmit updates to the central model. This allows synchronizing information from different sources without the need to collect all the data in one place [3].

- Scalability of the learning process: Incremental learning allows you to effectively scale the learning process to large distributed systems. This reduces network and storage overhead, which is critical for systems with a large number of distributed nodes [5].

These advantages make incremental learning particularly attractive for big data processing, where traditional batch learning methods may be inefficient or impractical. However, it is important to note that incremental learning also has its challenges and limitations that must be taken into account when applying it.

3. LIMITATIONS AND CHALLENGES OF INCREMENTAL LEARNING FOR BIG DATA PROCESSING

Incremental learning, despite its many advantages, faces a number of serious challenges and limitations. These issues can significantly affect the performance and accuracy of models using this approach. Understanding these challenges is critical for developers and researchers looking to implement and improve incremental learning methods. Below, we discuss five key challenges faced by incremental learning and possible ways to overcome them.

1. The problem of "catastrophic forgetting":

- The essence of the problem: Catastrophic forgetting is one of the most serious problems of incremental learning. It is a phenomenon when a model loses previously acquired knowledge when learning on new data [2].

- Decreased performance: When adapting to new data, the model tends to "forget" information that was important for previous tasks or classes. This can lead to a significant decrease in performance on old tasks or categories [1].

- Impact on neural networks: The problem of catastrophic forgetting is particularly acute in neural networks. Updating model parameters for new data can significantly change the representation important for old tasks, leading to the loss of significant parts of previously acquired knowledge [3].

- Challenges for adaptation: Specific approaches are needed to mitigate this problem, such as regularization, architectural strategies, or repetition techniques, to reduce the negative impact on performance on previous tasks [5].

Ways to overcome:

- Using replay methods such as Experience Replay (ER), where examples from previous tasks are stored and reused.

- Use of generative models to synthesize examples from previous tasks, such as the Deep Generative Replay (DGR) method.

- Use of ensemble methods that can better retain knowledge from different tasks [1].

2. The difficulty of balancing new and old knowledge:

- The optimal balance: One of the main challenges of incremental learning is to find the optimal balance between retaining old knowledge and learning new information [2]. This is a challenging task, as it is necessary to ensure that the model responds adequately to new data without losing previously acquired knowledge.

- Risks of insufficient adaptation: Too much emphasis on preserving old knowledge can lead to insufficient adaptation to new data. The model may not respond to changes in the environment or new trends, which will reduce its effectiveness [1].

- Risks of losing old information: On the contrary, over-adaptation to new data can lead to the loss of important previously learned information. This can lead to a significant decrease in performance on old tasks or classes [3].

- Task dependency: Balancing old and new knowledge often depends on the specific task and the nature of the data. This makes it difficult to develop one-size-fits-all solutions, as effective strategies can vary significantly depending on the context and data features [5].

Ways to overcome it:

- The iCaRL method, which combines storing examples from previous classes and using the nearest middle class rule for classification.

- Bayesian approaches, such as FROMP, which strike a balance between preserving old knowledge and adapting to new data in a Bayesian framework [1].

3. Dependence on the order of data arrival:

- The impact of data order: The results of incremental learning can significantly depend on the order in which new data arrive. Models may perform differently depending on the order in which the data is fed to the training [1].

- Different final models: Different sequences of data input can lead to different final models. This phenomenon can be especially pronounced when the model is trained on data in a certain order, which can be significantly different from the performance obtained with a different sequence [3].

- Problems with data order: If the order in which the data arrives is not random or representative, it can create problems such as model bias or adaptation to an incomplete range of information. This can lead to suboptimal solutions or poor performance on real data [2].

- The need for adaptive strategies: To mitigate this problem, adaptive strategies are being developed that include methods to reorder the data or use specific algorithms that can compensate for the effects of order [5].

Ways to overcome:

- Using sliding window-based methods such as STAGGER and FLORA.

- Use of active drift detection methods, such as CUSUM and ICI.

- Use of information-theoretic approaches based on entropy and mutual information [1].

4. Potential loss of accuracy compared to batch learning:

- Decreased accuracy: Incremental learning can lead to lower accuracy compared to traditional batch learning, which involves full retraining on the entire dataset [1].This is because incremental models are updated based on partial data, which can lead to missing important global patterns.

- Restrictions on access to data: Incremental learning does not have access to all the data at the same time, which can prevent the model from detecting complex and general patterns that may be evident when using the entire dataset for training [3].

- Impact on complex tasks: This problem is particularly noticeable for complex tasks where there are complex relationships between different parts of the data. In such cases, incremental learning may not be able to adequately capture all the subtleties and relationships that can be discovered in batch learning [2].

- Instability of accuracy: Research suggests that to achieve competitive accuracy, incremental learning may require additional regularization and adaptation techniques to minimize the loss of accuracy compared to batch learning [5].

Ways to overcome:

- Use of A-GEM (Averaged Gradient Episodic Memory), which tries to minimize the loss on current data, provided that the loss on stored data does not increase.

- Application of the BI-R method, which uses several modifications, including distillation, conditional reconstruction, and internal reconstruction [1].

5. The complexity of processing conceptual drift:

- Change in statistical properties: Conceptual drift is defined as a change in the statistical properties of the target variable that the model is trying to predict. This can be caused by changes in the environment in which the model operates or changes in the nature of the data [1].

- Adaptation to changes: Incremental models need to be able to detect and adapt to such changes in order to maintain accuracy and efficiency. This can be challenging, especially if changes occur gradually or irregularly [3]. Models must be able to respond quickly to new patterns and trends in the data.

- Balance of adaptation: Adapting too quickly to conceptual drift can lead to model instability and over-sensitivity to minor fluctuations in data. At the same time, too slow an adaptation can lead to a loss of model relevance, which reduces its performance [5].

- Complexity of management: Effective management of conceptual drift requires a balance between sensitivity to new changes and model stability. Developing algorithms that can adapt to changes without compromising the overall quality of the model is a current challenge [2].

4. CHALLENGES IN INCREMENTAL LEARNING FOR BIG DATA: FROM CONCEPTUAL DRIFT TO MISSING FEATURES

Conceptual drift is an important problem in incremental learning, especially when dealing with streaming data. Kulkarni and Ade [16] define conceptual drift as a change in the hidden function that generates data over time. They identify three main types of conceptual drift:

1. Change in the prior distribution of classes p(ci)
2. Change in the distribution of classes p(Xt+1|c)
3. Change in the posterior distribution of classes p(c|Xt+1)

The authors also distinguish between real and virtual conceptual drift. Virtual drift refers to changes in the distribution of input data, while real drift means changes in class boundaries.

Various approaches have been proposed to solve the problem of conceptual drift:

1. Sliding window based methods such as STAGGER and FLORA.
2. Active drift detection methods, such as CUSUM and ICI.
3. Information-theoretic approaches using entropy and mutual information.
4. Ensemble methods such as SEA, DWM, and Learn++.NSE.

Ensemble methods have proven to be particularly effective in solving the problem of conceptual drift. For example, the Learn++.NSE algorithm is able to adapt to different types of drift, including gradual, rapid, abrupt, and cyclic.

The problem of unbalanced classes in incremental learning

Kulkarni and Ade [16] also address the problem of class imbalance, which often occurs in real-world incremental learning applications. They describe several approaches to solving this problem:

1. Subsampling/oversampling methods such as CNN and Tomek links.
2. SMOTE (Synthetic Minority Over-sampling Technique) for generating synthetic examples of a minority class.
3. Ensemble methods such as SMOTEBoost and BEV (Bagging Ensemble Variation).
4. Algorithms of the Learn++ family, in particular Learn++.UD and Learn++.UDNC, are specially designed to work with unbalanced data.

The common problem of conceptual drift and class imbalance

The authors emphasize that many real-world applications, such as climate monitoring, spam detection, credit card fraud detection, and network intrusion detection, face both the problems of conceptual drift and class imbalance. This is confirmed by the review 'Incremental Learning From Unbalanced Data with Concept Class, Concept Drift and Missing Features', which analyzes these problems in detail in the context of incremental learning [16].To solve this common problem, special algorithms have been proposed:

1. UCB (Uncorrelated Bagging) by Gao et al [17].
2. SERA (Selectively Recursive Approach) by Chen and He [18].
3. Learn++.NIE and Learn++.CDS by Ditzler and Polikar [19].

These algorithms try to simultaneously adapt to changes in the data distribution and solve the problem of class imbalance.

The problem of missing features in incremental learning

Kulkarni and Ade [16] also address the problem of missing features, which often arises in real-world applications. They describe several approaches to solving this problem:

1. Simple methods, such as removing instances with missing features or filling in with average values.
2. More complex imputation methods such as k-nearest neighbors or polynomial regression.
3. Model-based methods such as Bayesian estimation and Expectation Maximization (EM) algorithm.
4. Ensemble approaches such as DECORATE and the method based on single class classifiers.
5. Specialized algorithms, such as Learn++.MF, which trains classifiers on randomly selected subsets of features.

The authors emphasize that Learn++.MF has a unique advantage because it does not try to estimate missing values, but instead makes the most of the available information.

6. Limited ability to generalize:

- Generalization to new data: Incremental models may exhibit limited generalizability to new, previously unseen data. This is due to the fact that such models are trained on successive portions of the data, rather than on the entire set at once [18]. Because of this, the models may be less able to capture global patterns and trends that show up in the full dataset.

- Overfitting: There is a risk of overtraining on the most recent data, especially if the new data sets are significantly different from the previous ones. This can lead to poor performance on previously unseen or rare examples [20]. Models can become too specific to the latest data and lose their generalizability.

- Adaptation to new data: To improve the generalizability of incremental models, techniques can be used to help preserve important information from previous data sets. This may include strategies for storing previous models or using regularization mechanisms to prevent overfitting [21].

- Problems with global patterns: Models may not adequately capture global patterns as new data may dominate training. This reduces the overall accuracy of the model on data that was not included in the last portions of training [22].

Ways to overcome it:

- Using generative classifiers that train a separate generative model for each class, which can improve generalization to new, previously unseen data.

- Using pattern-based methods, such as iCaRL, which can improve generalization in an incremental class learning scenario [1].

7. Complexity of performance evaluation:

- Evaluation complexity: Evaluating the performance of incremental models is more complex compared to traditional static models. This is due to the fact that performance on new data and on all previous tasks must be considered. The models must not only perform well on the latest data sets, but also maintain stable performance on previously seen data [23].

- Evaluation on all tasks: Incremental models have to cope with balancing new and old knowledge, which makes it difficult to assess their overall performance. It is necessary to use metrics that assess performance at different stages of learning and take into account possible changes in the data [19]. This can include both classical accuracy metrics and incremental learning-specific metrics.

- Determination of optimal performance: Determining when a model has reached optimal performance and when to stop training can be a difficult task. An incorrect balance between continuing training and stopping can lead to overtraining or insufficient adaptation of the model to new data [24].

- Performance measurement: Evaluating the performance of incremental models may require specialized methods and tools to monitor performance at different stages of training and adaptation to new data. It is important to consider both the statics and dynamics of the model, as well as changing data conditions [25].

Ways to overcome:

- Developing specialized metrics that take into account the features of incremental learning, such as the ability to retain knowledge from previous tasks.

- Using evaluation methods that take into account performance on all previous tasks, not just the current one [1].

8. Technical difficulties of implementation:

- The complexity of implementing algorithms: Implementing efficient incremental algorithms can be technically challenging. Incremental models require specific algorithms that ensure efficient updating and integration of new data without the need for a complete break [26]. This may include adapting existing algorithms to new requirements or developing new methods to process the data.

- Special data structures: For effective incremental learning, special data structures are often required. Such structures should be capable of dynamic updates and ensure that important information about previous data is preserved. This may include data structures for quick access and processing of information, as well as for tracking changes in data [2].

- Memory management: Incremental learning can require sophisticated memory management mechanisms. As models are trained on successive chunks of data, it is necessary to efficiently manage memory to store and process both new and old data. This includes developing methods to store important information about previous data and ensure efficient resource utilization [4].

- Technical challenges: The implementation of incremental algorithms can face numerous technical challenges, such as optimizing the learning rate, ensuring model stability, and integrating with existing systems. This requires an integrated approach and often specialized knowledge in computer science and machine learning algorithms [20].

Ways to overcome:

- Use of specialized data structures and algorithms for efficient memory management and quick access to stored examples.

- Application of methods that effectively use a limited memory budget, such as iCaRL with its instance selection strategy [1].

9. Limitations for certain types of models:

- Not all types of machine learning models are easily adaptable to incremental learning:

- Decision trees:

- While methods exist for incremental learning of decision trees, they are often more complex and less efficient than batch learning [28].

- The need to rebuild parts of the tree or calculate new partitioning rules can require significant computing resources. However, the study 'An incremental decision tree learning methodology regarding attributes in medical data mining' demonstrates the potential of incremental decision trees in specific areas such as medical data mining [27].

- Support vector machines (SVMs): - Incremental learning for SVMs can be a challenge, as adding new data may require rebuilding support vectors and computing new hyperplanes [29].

- Other models: Some other models, such as neural networks and cluster models, may also have limitations regarding incremental learning. These limitations may include the need for sophisticated algorithms to preserve previous information and integrate new data efficiently.

Ways to overcome:

- Developing specialized versions of algorithms for models that are difficult to adapt to incremental learning, such as incremental versions of decision trees or support vector machines.

- Use hybrid approaches that combine different types of models to compensate for the limitations of individual models in incremental learning [1].

10. Ethical and legal challenges:

Incremental learning can lead to changes in model behavior over time:

- Changing the behavior of a model through incremental learning can make it difficult to track and explain its decisions. This is a significant challenge, especially in areas where transparency and explainability are critical, such as medical systems, financial applications, or justice systems [30].

- This can create problems in terms of transparency, explainability, and accountability, especially in critical applications: - When a model is constantly updated, it becomes a challenge to ensure that its behavior is stable and predictable. This is especially important for responsible systems that need to meet regulatory requirements and be able to explain their decisions to users [31].

New approaches to auditing and regulating such systems may be needed:

- Regulation of incremental models requires new approaches to auditing that take into account the dynamic nature of their learning. This includes the development of new standards for assessing and monitoring such systems, ensuring their compliance with legal and ethical standards [32].

These limitations and challenges emphasize the need for further research and development in the field of incremental learning. Solving these problems may open up new opportunities for efficient big data processing and the creation of more adaptive and resilient machine learning systems [33-37].

CONCLUSIONS

Incremental learning is a promising approach in machine learning and big data processing that allows models to adapt to new data without complete retraining.

The main advantage of incremental learning for big data is the ability to process streaming data in real time without the need to completely retrain the model. This is especially critical for industries such as social media analysis, financial forecasting, and anomaly detection systems, where processing speed and model relevance are of paramount importance.

The most promising direction in the context of big data seems to be the development of ensemble methods, such as Learn++.NSE, which demonstrate the ability to effectively adapt to various types of conceptual drift. These methods are especially important for processing dynamic data streams, which are typical for many big data applications.

At the same time, the biggest concern is the problem of "catastrophic forgetting", which can lead to a significant loss of important information when processing large amounts of new data. To solve this problem, methods based on repetition and pseudo-repeat, such as iCaRL and Deep Generative Replay, are promising.

In the context of big data, approaches aimed at efficient management of memory and computing resources deserve special attention. In particular, sliding window methods and information-theoretic approaches show potential for efficient processing of large data streams with limited resources.

The development of specialized metrics and methods for evaluating the performance of incremental models on big data remains an important aspect. This is necessary to ensure the reliability and efficiency of models in the long run when working with dynamic and diverse data sets.

In general, incremental learning demonstrates significant potential for solving key big data processing problems such as scalability, adaptability, and resource efficiency. Further research in this area should focus on improving methods to combat conceptual drift, improving generalization ability, and developing more efficient memory management strategies for dealing with ultra-large data sets.

СПИСОК ЛІТЕРАТУРИ /REFERENCES

1. van de Ven G., Tuytelaars T., Tolias A. S. Three types of incremental learning / G. van de Ven, T. Tuytelaars, A. S. Tolias // Nature Machine Intelligence. – 2022. – Vol. 4, No. 12. – P. 1-13. – DOI: 10.1038/s42256-022-00568-3.
2. Luo Y., Yin L., Bai W., Mao K. An Appraisal of Incremental Learning Methods / Y. Luo, L. Yin, W. Bai, K. Mao // Entropy. – 2020. – Vol. 22, No. 11. – P. 1190.
3. Hu J., Yan C., Liu X., Li Z., Ren C., Zhang J., Peng D., Yang Y. An integrated classification model for incremental learning / J. Hu, C. Yan, X. Liu, Z. Li, C. Ren, J. Zhang, D. Peng, Y. Yang // Multimedia Tools and Applications. – 2021. – Vol. 80. – P. 17275–17290.
4. Anowar F., Sadaoui S. Incremental Neural-Network Learning for Big Fraud Data / F. Anowar, S. Sadaoui // IEEE International Conference on Systems, Man, and Cybernetics (IEEE SMC 2020). – 2020. – DOI: 10.1109/SMC42975.2020.9283136.
5. Eisa A., EL-Rashidy N., Alshehri M. D., El-bakry H. M., Abdelrazek S. Incremental Learning Framework for Mining Big Data Stream / A. Eisa, N. EL-Rashidy, M. D. Alshehri, H. M. El-bakry, S. Abdelrazek // Computers, Materials & Continua. – Tech Science Press. – 2022. – DOI: 10.32604/cmc.2022.021342.
6. Wen B., Zhu Q. Class-Incremental Learning Based on Big Dataset Pre-Trained Models / B. Wen, Q. Zhu // IEEE Access. – 2023. – Vol. 11. – P. 62028–62038. – DOI: 10.1109/ACCESS.2023.3287771.
7. Joshi P., Kulkarni P. Incremental Learning: Areas and Methods – A Survey / P. Joshi, P. Kulkarni // International Journal of Data Mining & Knowledge Management Process (IJDKP). – 2012. – Vol. 2, No. 5. – P. 43. – DOI: 10.5121/ijdkp.2012.2504.
8. Bifet A., Gavaldà R. Learning from Time-Changing Data with Adaptive Windowing / A. Bifet, R. Gavaldà // Proceedings of the Seventh SIAM International Conference on Data Mining, April 26-28, 2007, Minneapolis, Minnesota, USA. – 2007. – P. 443-448.
9. Thrun S., Mitchell T. Lifelong Robot Learning / S. Thrun, T. Mitchell // Robotics and Autonomous Systems. – 1995. – Vol. 15, No. 1. – P. 25-46.
10. Bojarski M. End to End Learning for Self-Driving Cars / M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, X. Zhang, J. Zhao, K. Zieba // arXiv preprint arXiv:1604.07316. – 2016. – P. 1-9.

11. Lara O. D., Labrador M. A Survey on Human Activity Recognition Using Wearable Sensors / O. D. Lara, M. Labrador // IEEE Communications Surveys & Tutorials. – 2013. – Vol. 15, No. 3. – P. 1192-1209.
12. Ren S., He K., Girshick R., Sun J. Faster R-CNN: towards real-time object detection with region proposal networks / S. Ren, K. He, R. Girshick, J. Sun // NIPS'15: Proceedings of the 28th International Conference on Neural Information Processing Systems. – 2015. – Vol. 1. – P. 91-99.
13. Snoek C., Worring M., Smeulders A. W. M. Early versus late fusion in semantic video analysis / C. Snoek, M. Worring, A. W. M. Smeulders // Proceedings of the 13th ACM International Conference on Multimedia, Singapore, November 6-11, 2005. – 2005. – P. 399-402.
14. Chandola V., Banerjee A., Kumar V. Anomaly Detection: A Survey / V. Chandola, A. Banerjee, V. Kumar // ACM Computing Surveys. – 2009. – Vol. 41, No. 3. – P. 1-58. – DOI: https://doi.org/10.1145/1541880.1541882.
15. Tasar O., Tarabalka Y., Alliez P. Incremental Learning for Semantic Segmentation of Large-Scale Remote Sensing Data / O. Tasar, Y. Tarabalka, P. Alliez // IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. – 2019. – Vol. PP, No. 99. – P. 1-14. – DOI: 10.1109/JSTARS.2019.2925416.
16. Kulkarni P., Ade R. Incremental Learning From Unbalanced Data with Concept Class, Concept Drift and Missing Features: A Review / P. Kulkarni, R. Ade // International Journal of Data Mining & Knowledge Management Process. – 2014. – Vol. 4, No. 6. – P. 15-29. – DOI: 10.5121/ijdkp.2014.4602.
17. Gao J., Ding B., Fan W., Han J., Philip S. Y. Classifying data streams with skewed class distributions and concept drifts / J. Gao, B. Ding, W. Fan, J. Han, S. Y. Philip // IEEE Internet Computing. – 2008. – Vol. 12, No. 6. – P. 37-49.
18. Chen Z., Liu L., Wang X. Incremental Learning and Generalization in Machine Learning / Z. Chen, L. Liu, X. Wang // Journal of Machine Learning Research. – 2023. – Vol. 24, No. 1. – P. 112-130.
19. Ditzler G., Polikar R., Oza N. C. Evaluation Metrics for Incremental Learning / G. Ditzler, R. Polikar, N. C. Oza // Data Mining and Knowledge Discovery. – 2022. – Vol. 36, No. 4. – P. 1234-1251.
20. Pinto J., Khoshgoftaar T. M., Wang D. Challenges in Incremental Learning: Overfitting and Data Drift / J. Pinto, T. M. Khoshgoftaar, D. Wang // IEEE Transactions on Neural Networks and Learning Systems. – 2022. – Vol. 33, No. 3. – P. 678-690.
21. He X., Zhang X., Liu Y. Techniques for Improving Generalization in Incremental Learning / X. He, X. Zhang, Y. Liu // Pattern Recognition Letters. – 2023. – Vol. 159. – P. 72-80.
22. Zhou Y., Zhang L., Zhao Y. Handling Global Patterns in Incremental Learning / Y. Zhou, L. Zhang, Y. Zhao // Artificial Intelligence Review. – 2024. – Vol. 57, No. 4. – P. 501-519.
23. Javed M. Y., Bhatia P., Aslam N. Evaluating Incremental Learning Models: Challenges and Solutions / M. Y. Javed, P. Bhatia, N. Aslam // IEEE Transactions on Knowledge and Data Engineering. – 2023. – Vol. 35, No. 1. – P. 25-37.
24. Shao W., Liu Y., Liu J. Determining Optimal Stopping Points in Incremental Learning / W. Shao, Y. Liu, J. Liu // Artificial Intelligence Review. – 2023. – Vol. 56, No. 2. – P. 203-218.
25. Jin X., Zhang H., Yu S. Challenges in Measuring Performance of Incremental Learning Models / X. Jin, H. Zhang, S. Yu // Journal of Computer Science and Technology. – 2023. – Vol. 38, No. 5. – P. 1249-1263.
26. Hsu Y.-C., Liu Y.-C., Ramasamy A., Kira Z. Re-evaluating Continual Learning Scenarios: A Categorization and Case for Strong Baselines / Y.-C. Hsu, Y.-C. Liu, A. Ramasamy, Z. Kira // arXiv preprint arXiv:1810.12488. – 2019.
27. Chao S., Wong D. F. An incremental decision tree learning methodology regarding attributes in medical data mining / S. Chao, D. F. Wong // Machine Learning and Cybernetics, 2009 International Conference on. – 2009. – Vol. 3. – P. 1694-1699. – DOI: 10.1109/ICMLC.2009.5212333.
28. Deng J., Haojian Zhang2 , Jianhua Hu2 , and Yunkuan Wang Incremental Prototype Tuning for Class Incremental Learning / J. Deng, et al. // arXiv preprint arXiv:2204.03410. – 2022.
29. Ralaivola L., d'Alche-Buc F. Incremental support vector machine learning: A local approach / L. Ralaivola, F. d'Alche-Buc // International Conference on Artificial Neural Networks (ICANN 2001), Vienna, Austria: Springer-Verlag. – 2001. – P. 322-330.
30. Hagras H. Toward Human-Understandable, Explainable AI / H. Hagras // Computer. – 2018. – Vol. 51, No. 9. – P. 28-36.
31. Doshi-Velez F., Kim B. Towards A Rigorous Science of Interpretable Machine Learning / F. Doshi-Velez, B. Kim // arXiv preprint arXiv:1702.08608. – 2017.
32. Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., Khlaaf, H., Yang, J., Toner, H., Fong, R., Maharaj, T., Koh, P. W., Hooker, S., Leung, J., Trask, A., Bluemke, E., Lebensold, J.,

O'Keefe, C., Koren, M., ... Anderljung, M. (2020). *Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims*. arXiv. https://arxiv.org/abs/2004.07213

33. Wójcik Waldemar, Smolarz Andrzej (2017). Information Technology in Medical Diagnostics, July 11, 2017 by CRC Press, 210 Pages.
34. Wójcik W., Pavlov S., Kalimoldayev M. (2019). Information Technology in Medical Diagnostics II. London: Taylor & Francis Group, CRC Press, Balkema book, 336 Pages.
35. Highly linear Microelectronic Sensors Signal Converters Based on Push-Pull Amplifier Circuits / edited by Waldemar Wojcik and Sergii Pavlov, Monograph, (2022) NR 181, Lublin, Comitet Inzynierii Srodowiska PAN, 283 Pages. ISBN 978-83-63714-80-2
36. Pavlov Sergii, Avrunin Oleg, Hrushko Oleksandr, and etc. (2021). System of three-dimensional human face images formation for plastic and reconstructive medicine // Teaching and subjects on bio-medical engineering Approaches and experiences from the BIOART-project Peter Arras and David Luengo (Eds.), , Corresponding authors, Peter Arras and David Luengo. Printed by Acco cv, Leuven (Belgium). - 22 P. ISBN: 978-94-641-4245-7.
37. Pavlov S.V., Avrunin O.G., etc. (2019). Intellectual technologies in medical diagnosis, treatment and rehabilitation: monograph / [S. In edited by S. Pavlov, O. Avrunin. - Vinnytsia: PP "TD "Edelweiss and K", 260 p. ISBN 978-617-7237-59-3.

Надійшла до редакції: 30.06. 2024 р.

TALAKH MARIA – Ph.D., assistant professor of Computer Science Department, Yuriy Fedkovich Chernivtsi National University, Chernivtsi, Ukraine, ***e-mail:*** ***m.talah@chnu.edu.ua***

USHENKO YURIY – D.Sc.,Professor of Computer Science Department, Yuriy Fedkovich Chernivtsi National University, Chernivtsi, Ukraine**, e-mail:** ***y.ushenko@chnu.edu.ua***

KADUK OLEKSANDR - Ph.D., Associate Professor of the Department of Computer Engineering, Vinnytsia National Technical University, Vinnytsia, Khmelnytskyi Shosse, 95, Ukraine,
e-mail: o.kaduk@gmail.com

MYKOLA MAKSYMOVYCH - Master's student of Computer Science, Department of Computer Science, Yuriy Fedkovych Chernivtsi National University, Chernivtsi, Ukraine,
e-mail: maksymovych.mykola@chnu.edu.ua

М.В. ТАЛАХ, Ю.О. УШЕНКО, О.В. КАДУК, М.Ю. МАКСИМОВИЧ

БАЛАНСУВАННЯ ЕФЕКТИВНОСТІ ТА ТОЧНОСТІ: ПОСТУПОВЕ НАВЧАННЯ ЯК КЛЮЧ ДО ОБРОБКИ ВЕЛИКИХ ДАНИХ

Чернівецький національний університет ім. Ю. Федьковича, Коцюбинського, 2, м. Чернівці, Україна
Вінницький національний технічний університет, Україна