Concept Drift Challenges in the Internet of Things Era of Smart Cities: A Preliminary Investigation

Sina Samadzad Etehadi
Ferdowsi University of Mashhad
samadzad@um.ac.ir

Mohammad Hossein Yaghmae
Ferdowsi University of Mashhad
hyaghmae@um.ac.ir

Naserali Noorani
Ferdowsi University of Mashhad
alinoorani517@mail.um.ac.ir

Mahsa Hosseinpour
Ferdowsi University of Mashhad
mahsa.hosseinpour@mail.um.ac.ir

Seyed Amin Hosseini Seno
Ferdowsi University of Mashhad
hosseini@um.ac.ir

Abstract— **In recent years, smart cities and their enabling technologies, such as the Internet of Things (IoT), have grown and expanded significantly. This trend has led to an exponential increase in sensors and smart devices and, as a result, facing the phenomenon of streaming big data continuously being produced with high volume, speed, and variety. This issue has created an unprecedented opportunity for businesses and companies to exploit this data through a new generation of data-driven applications. However, the real-time, dynamic, non-stationary nature and the changing statistical patterns of big data streams in the IoT present data analysis with a challenge called concept drift as the leading cause of the gradual decline in efficiency and unreliability of static machine learning models. One of the research challenges of this field is how to deal with this issue and make real-time decisions in dynamic and complex situations. Although concept drift and IoT have each been independently researched in the literature, the confluence of these subjects has not yet received any attention. Consequently, this paper serves as a beginning step in addressing the problem in this area. After discussing the significance and various notions of the concept drift, this paper will highlight some of the latest research works in this field. To conclude, the paper will present this field's challenges and research opportunities.**

Keywords: Smart City; Internet of Things (IoT); Concept Drift; Big Data, Stream Analytics;

I. INTRODUCTION

Smart cities are complex technical and social infrastructures that include a wide range of users (citizens, operators, executive institutions, and public and private companies) and digital devices (sensors and actuators), which are applied in a vast array of topics such as energy, transportation, healthcare, smart buildings and Industry 4.0 [1]. As shown in Fig. 1, smart cities have various enabling technologies such as the IoT, big data, cloud computing, next-generation networks (5G and 6G), artificial intelligence, and digital twins. However, the ever-increasing expansion of IoT infrastructure as an underlying technology and the heart of the data acquisition process provides a platform for combining and developing other technologies in this field.

In the meantime, the IoT is known as the arm of smart cities' generation, collection, and transfer process. IoT uses numerous

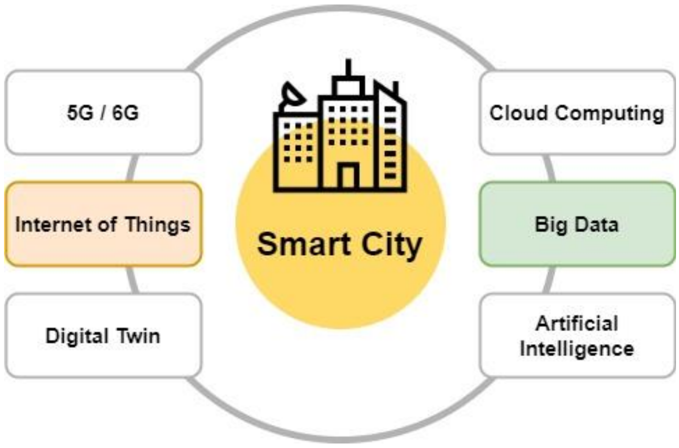

Figure 1. Key enabling technologies in Industry 4.0

sensors, actuators, devices, and complex networks to reach intelligent interconnection, data generation, and sharing [2].

Due to technological advancements, the IoT has found extensive application in industrial automation and intelligence, enhancing efficiency and quality in industrial manufacturing processes [3]. Simultaneously, the growth of the IoT infrastructure has resulted in a proliferation of sensors and devices, leading to a substantial influx of big data streams over time. Accordingly, connected devices are estimated to reach more than 70 billion in 2025 [4]. This behavior shows the exponential growth rate of big data streams in recent years. In addition, one of the potential effects of IoT growth is on the economy, which will reach USD 11 trillion per year in economic productivity by 2025 [5].

With the expansion of the IoT establishment and intelligent sensors as the main arm of the production process, data collection and transmission in the next generation of smart cities make a wide range of condition monitoring data available. Creating such a platform, along with unprecedented developments in the field of data science and especially artificial intelligence techniques and robust theoretical computing platforms of cloud computing, is a unique opportunity to link traditional systems with modern information technology systems to discover the knowledge hidden in data on the way to creating systems [6].

IoT includes a variety of data-related concepts and technologies that interact with each other to create a process. This process starts with the production and collection of data. It will continue until its transmission in the context of communication networks, storage and processing, analysis, decision-making, and feedback to the factory environment. An overview of this data-driven process is depicted in Fig. 2.

This process provides a clear picture of the purpose and operation of a data-driven IoT system, where these large-scale infrastructures continuously generate big data streams to monitor the conditions of urban equipment, such as voltage, current, humidity, and temperature, for various smart city applications. However, this process will only have a significant added value with methodical approaches for the real-time exploitation of these big data streams produced by various and sometimes heterogeneous data sources.

The characteristics of big data are mainly divided into three areas: volume, speed, and variety of data [7]. Meanwhile, speed, which directly refers to real-time data streams produced by sensors, is considered one of the most essential paradigms in this field. Data science refers to statistical techniques and methods for analyzing and extracting knowledge from data. However, analyzing the massive volume of IoT stream data using machine learning models is an open issue[8]. The primary challenges in IoT data analytics stem from limitations in both time and resources, mainly due to the stringent power and cost demands imposed by IoT devices [9].

Conversely, the real-time, dynamic, unusual, and non-stationary characteristics of data streams within the IoT, reflecting the constantly evolving nature of this process, will give rise to unforeseeable behaviors at the urban equipment and device levels. Factors that may trigger inescapable alterations in the statistical distributions of IoT data streams, a phenomenon often termed concept drift [9], were introduced in 1986 by Schlimmer et al. [10]. Concept drift can occur naturally, such as the gradual reduction of the life of urban equipment, or unnaturally, due to factors such as breakdowns or physical or cyber-attacks. With the occurrence of concept drifts and the reduction of statistical similarity between the old data streams under which the model was trained and the new data streams that are caused by the non-static nature of the constantly changing environments of the IoT, the prediction model gradually changes the boundaries of classification and clustering.

Concept drift, in particular, pertains to the gradual and unforeseeable shifts in the statistical and distributional attributes of the target variable that a model aims to predict. Concept drift leads to a disparity between the model's treatment of old and new data, consequently diminishing the efficiency of the predictive model founded on historical data. Concept drift is widely acknowledged as the principal factor that reduces the effectiveness of data-driven information systems. It may fail an IoT system or degrade its performance over time. The complexity of the drift control idea is further heightened by the uncertainty and time-dependence features of IoT data streams [11].

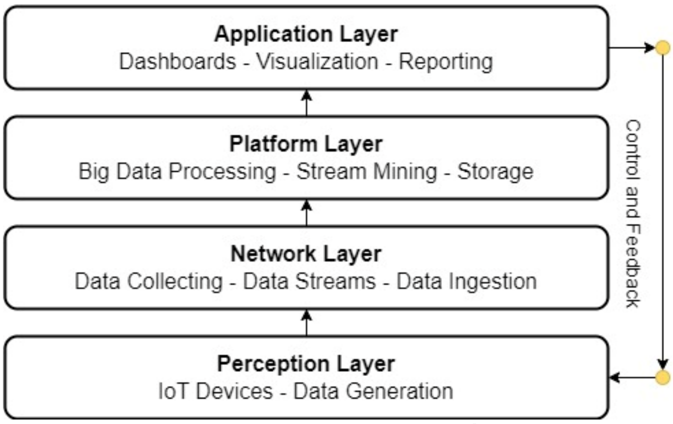

Figure 2. Data management process in the IoT

For conventional offline machine-learning models, concept drift might be an issue. Primarily, updates are required after these models are trained using historical or offline data. Therefore, we need models that can adapt to the expected and unforeseen changes in the dynamic settings of the IoT to prevent concept drift. This is because real-world systems, which are incredibly dynamic environments, present severe difficulties for previously trained machine-learning models [12].

This category of data analysis is implemented across diverse sectors of automated processes, encompassing predictive maintenance and applications like anomaly detection and environmental monitoring. A clear and typical example of this concept is the forecasting of weather conditions, which may change substantially by season. This will directly affect the short-term and long-term behavior of urban equipment. An efficient learning model should be able to track such changes and quickly adapt to them. While some algorithms overreact to noise, mistaking it for concept drift, others may resist noise and slowly adapt to the changes. A learner's robustness to noise and sensitivity to concept drift should be combined [13].

With real-time and online data [14], concept drift is lessened against frequent and continual retraining of the model. To quickly regain their performance, online learning algorithms must be able to detect and respond to concept drift as a mitigation mechanism. Consequently, early concept drift detection becomes an essential talent for online learning algorithms in non-stationary contexts [12]. Recent advancements in the study of learning in non-stationary contexts have benefited from such a necessity. This paper's main objective is to review these recent IoT advancements.

A. Motivation

Early concept drift change detection is crucial for machine learning models to be continuously updated. The significance of this is increased, particularly in the area of time-sensitive IoT applications where early prediction of potential failures and flaws in urban equipment is regarded as a crucial concern. Consequently, it is crucial to identify such behaviors quickly.

In the past decade, there has been a significant surge in research focused on learning with concept drift, leading to the development of numerous algorithms tailored for drift-aware online and adaptive learning. Despite its importance, the concept drift detection and adaptation idea still needs to be thoroughly covered in the IoT-related literature. A reason behind this trend is the evolving technological paradigm known as the IoT and the extensive scope of this field, which is expanding in an unprecedented manner. Given the critical role of real-time applications within the IoT, there is an apparent and pressing demand for a thorough summary of recent research. This summary would establish a unified understanding of definitions, concepts, and terminology among researchers in this domain.

Accordingly, the current study aims to articulate the concepts and provide a road map for individuals interested, researchers, and developers as a first step, considering this and the need to pioneer new territory in this sector.

This work can serve as a concise reference for researchers of real-time data analysis systems employed in the IoT, and it will be helpful in their understanding of the present demands of this field.

B. Contribution

The main contributions of this paper can be summarized as follows:

- To fill the existing research gap, we are taking action to reveal the importance of recognition and adaptation of concept drift in the IoT.
- We will discuss the available strategies and tactics to address this phenomenon after studying the idea of concept drift and outlining its types, causes, and effects.
- We will clarify the research path of this field by reviewing recent and related works.
- We will express the challenges and future research opportunities in this field.

C. Paper Organization

The rest of this paper is organized as follows. In section II, we present a background and definition of necessary concepts emphasizing on concept drift. Section III covers recent contributions to concept drift especially its detection in IoT systems. We mention to some open issues in Section IV. Finally we conclude the paper in Section V.

II. Background

The concepts and terminologies relevant to this topic will be explained in this section to give the researchers a perspective and familiarize them with the generalities. These items are reviewed below.

A. Online Learning

When we talk about the data generation process in the IoT, we face a time series analysis problem. as a sequence of samples or data as random variables listed in time quantity format and represented by scalar samples (x_t) in a vector. This case is shown in the equation (1).

$$X = x_1, x_2, \dots, x_n \tag{1}$$

The types of these data include structured, semi-structured, and unstructured ones. All time series-based applications will incorporate at least one aspect of behavior modeling, prediction, detection, control, and sample generation. The current article focuses on modeling. The data that the model was trained on should be represented in the model.

Based on this, we will have equation (2), which will be the input data and include various features to inject into the model defined based on the function. What we expect from the output of this function is the label corresponding to the input data, which is displayed by the target variable.

$$\hat{y} = f(X) \tag{2}$$

This model can be created in various ways, and assuming the use of artificial neural networks, this hypothetical model is shown in Fig. 3. The main goal in producing the output of the above model is that the diagnosis or prediction made is as close as possible to the actual data label value (y).

For example, consider X, reperesenting a collection of voltage sensor readings obtained from a machine between 10:00 AM and 12:00 AM on February 18th. In this context, a $y =$ "$good$" label indicates that the machine is functioning correctly, while a $y =$ "bad" label signifies that the machine is deteriorating. During the training data used to construct the model, both X (sensor readings) and y (labels indicating machine condition) are known. However, in the case of new examples to which the prediction model is applied, X (sensor readings) is known, but y (the machine's condition) is not known at the time of prediction.

In a general context, we encounter two primary learning approaches: offline learning and online learning. In offline learning, it is essential to have access to the complete training dataset during the model training phase. This model can be utilized for making predictions only after the training is finished. Conversely, online algorithms work sequentially, processing data as it comes in. These algorithms generate and execute a model without initially requiring the entire training dataset.

The model undergoes continuous updates during operation as more training data becomes available [15].

Machine learning models are often built using historical data in the form of input and output value pairs. Once trained, the models can forecast the results for fresh, unforeseen input data. However, most of the data generated by IoT infrastructures are in the form of data streams, and it is almost impractical to place

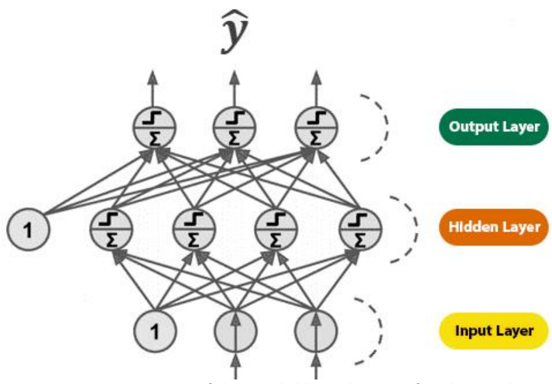

Figure 3. A view of the model based on artificial neural networks

this large volume of streaming data in the main memory and adapt it to a machine learning model pre-built by historical data [15]. Therefore, online learning will be a suitable approach to face this challenge.

With real-time and online data [14], concept drift is lessened against frequent and continual retraining of the model. In order to regain their performance, online learning algorithms must be able to detect and respond to concept drift as a mitigation mechanism. Consequently, early concept drift detection becomes an essential talent for online learning algorithms in non-stationary contexts [12]. In this scenario, machine learning models can undergo incremental training through continuous updates or be retrained using recent batches of data [15].

B. Concept Drift

In dynamic and non-stationary settings, the data distribution has the potential to shift over time, leading to the occurrence of concept drift [10]. The concept drift pertains to alterations in the conditional distribution of the output (i.e., the target variable) based on the input (input features), even though the distribution of the input remains constant [15].

Machine learning has two main components: training and prediction. Accordingly, based on Fig. 4, research in the field of learning under concept drifts will provide two new components in the classification of problems in this field:

- **Concept drift detection:** refers to the occurrence or non-occurrence of concept drift.
- **Concept drift adaptation:** refers to the mutual reaction to concept drift and the necessary measures in adapting the model to it.

According to Fig. 4, which represents the concept drift detection and adaptation framework, the traditional machine learning process, also called offline learning, includes creating a learning model based on historical or batch data. As mentioned, since this model is static and based on past data, it gradually loses its prediction accuracy by observing new data. To overcome this problem, we need to update the model based on streaming data. According to this framework, with the entry of new data and based on the relevant algorithms, we implement the concept drift discovery process; if no drift is discovered, the model is still functional and does not need to be updated. Nevertheless, the model will be matched based on the discovered drift if any concept drift is found.

A. Concept drift defination

According to a period [0, t], a set of data samples are defined as equation (3):

$$S_{0,t} = \{d_0, d_1, \dots d_t\} \tag{3}$$

Where $d_i = (X_i, y_i)$ is an observation or data sample. Also, X_i is the vector of features in the input of the model and y_i is known as the target variable or the label of that data sample. In addition, $S_{0,t}$ follows a certain flat distribution $f_{0,t}(X)$.

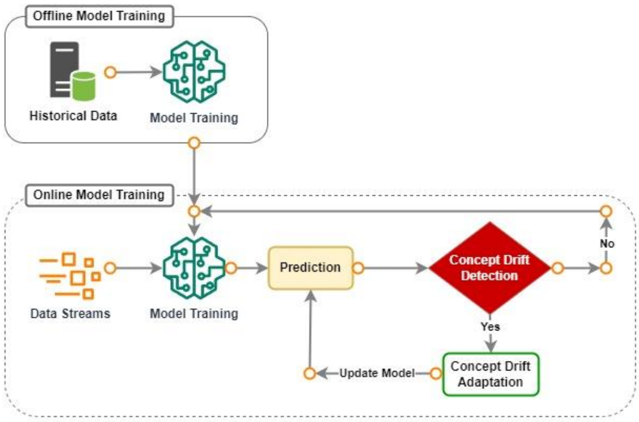

Figure 4. Concept drift detection and adaptation framework

Accordingly, we will witness a concept drift by changing the distribution of data in the $t + 1$ time label shown in equation (4).

$$F_{0,t}(X, y) \neq F_{t+1,\infty}(X, y) \tag{4}$$

Accordingly, The formal definition of concept drift between time point *t* and time point *t+1* can be defined as equation (5):

$$\exists_t : P_t(X, y) \neq P_{t+1}(X, y) \tag{5}$$

Where P_t denotes the joint probability distribution at time t between input variables X and the target variable y, changes in data can be characterized as changes in the components of this relation. Also, P_{t+1} is the joint probability distribution between these two variables at time P_{t+1} . If this equation is established, concept drift has occurred. Concept drift can arise from changes in both $P_t(X)$ and $P_t(y|X)$ [16]. We are keen to explore two consequences of these alterations:

- Whether the modification in the data distribution, represented as *P(y|X)* influences the predictive decision.
- Whether these changes in the data distribution can be observed from the data itself without having knowledge of the true labels, i.e., if *P(X)* undergoes alterations.

From a predictive standpoint, it is necessary to adapt only to changes that have an impact on the prediction decision [15]. Concept drift is also known in the literature with different words such as change in dataset [17] or concept change [18]. However, it is stated in [19] that drift or change of concept is only a subset of change in data set, because change in data set includes change of variable, change of prior probability and change of concept. These definitions alone will be able to determine the scope of each research. However, since concept drift is usually associated with variable change and previous probability, in most of the works titled concept drift and mathematical formulation proposed in equation (4) It is known [20] that will be the basis for our discussion in this paper. Based on this, the drift of the concept at time t can be defined as the change of the joint or joint probability of the features of X (input or independent variables) and y (target or dependent or predictor variable).

B. Causes and sources of concept drift

Various articles may present different causes of concept drift. In most cases [15], the two trends of real and virtual concepts are mentioned. However, the authors in [16] and [20] have classified the causes of concept drift into three areas or source. The causes of concept drift are shown in Fig. 5. It should be noted that the filled forms mean the old data classification and the empty conditions mean the new data classification.

- **Real concept drift:** When the probability of y given X undergoes a change, the probability of X remains unchanged. This scenario directly affects the prediction model and constitutes a genuine concept drift. It not only alters the feature space but also shifts its decision-making boundary. This means that the focus in concept drift will be on $P_t(y,X)$, and $P_t(X)$ will remain unchanged. Notably, the impact of $P_t(y|X)$ will cause a change in the decision boundary and reduce the accuracy of the model. We will mention this drift under the real concept drift. Equation (4) refers to this case:

$$P_t(y,X) \neq P_{t+1}(y,X) \; and \; P_t(X) = P_{t+1}(X) \qquad (4)$$

- **Virtual concept drift**: In cases where the probability of X changes while the probability of y given X remains unchanged. this situation does not constitute a concept drift. This means that the focus in concept drift will be on $P_t(X)$, and $P_t(y,X)$ will remain unchanged. Since the drift of $P_t(X)$ will not affect the decision boundary, it will be known as the virtual concept drift. Equation (5) refers to this case:

$$P_t(X) \neq P_{t+1}(X) \; and \; P_t(y|X) = P_{t+1}(y|X) \qquad (5)$$

- **Hybrid concept drift:** In dynamic and complex environments, it is possible for both real concept drift and virtual concept drift to coexist within the data stream simultaneously. This type of concept drift focuses on both $P_t(X)$ and $P_t(y,X)$ because both kinds of change carry essential information about the learning environment context. Equation (6) refers to this case:

$$P_t(X) \neq P_{t+1}(X) \; and \; P_t(y|X), \neq P_{t+1}(y|X) \qquad (6)$$

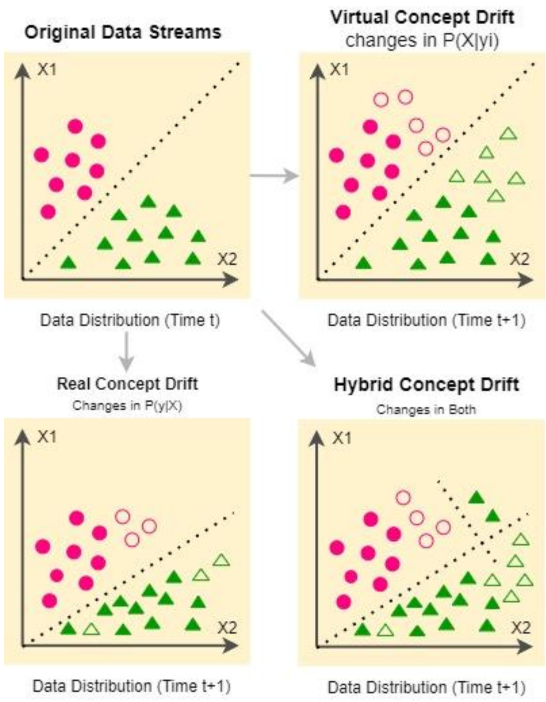

Figure 5. Causes and sources of concept drift

C. *Types of concept drift*

Based on the majority of studies related to this field based on Fig. 6, concept drift has been identified in four levels. According to [15][21][22][16], research in the field of concept drift adaptation focuses on minimizing how the model accuracy drops and achieving the fastest recovery rate during the concept drift process. On the other hand, the research in the field of the mentioned four types of concept drift emphasizes the use of historical concepts, that is, how to find the best historical ideas in the shortest time. A new concept may appear suddenly, gradually, incrementally, or repetitively. A concept drift may occur in the short or long term. Accordingly, an initial thrust may change over time in the form of other intermediate thrusts. An intermediate drift can be a combination of the beginning and end concepts, such as incremental drift, or simply one of the beginning or end concepts, such as gradual drift.

III. RELATED WORK

The purpose of this section is to review previous works with a top-down approach. In this way, firstly, it tried to examine the review paper in the field of concept drift with a general point of view, and then, the topic will be drawn with a gentle slope towards the detection of concept drift in the IoT.

Reviewing the concept drift literature, we find that a small range of works has investigated this challenge. However, as mentioned earlier, This phenomenon was presented for the first time in 1986 as an article [10], and its purpose was to point out the possibility of changing noise data to non-noise over time. Changes that may be caused by changes in latent variables that cannot be directly measured [23]. The work of [15], considered the first and the most comprehensive review article in this field, has investigated various aspects of concept drift in detail, and it is still considered a worthy work for reference. This article, while providing a clear definition of the problem, by providing a proposed framework, investigates in detail the dimensions of concept drift in the four aspects of memory, change detection, learning, and loss estimation. Also, in [24], the issue of learning in non-stationary environments and active and passive approaches to face this phenomenon is comprehensively discussed. Due to the novelty of concept drift and the lack of single and unified terms, the authors in [25] have done a complete review of the definitions to provide an easy understanding of the concept drift issues.

By mentioning an introduction in the field of articles related to concept drift, we will examine the effects of concept drift in the IoT. These articles have mainly focused on cyber security issues, delay issues edge computing, and federated learning.

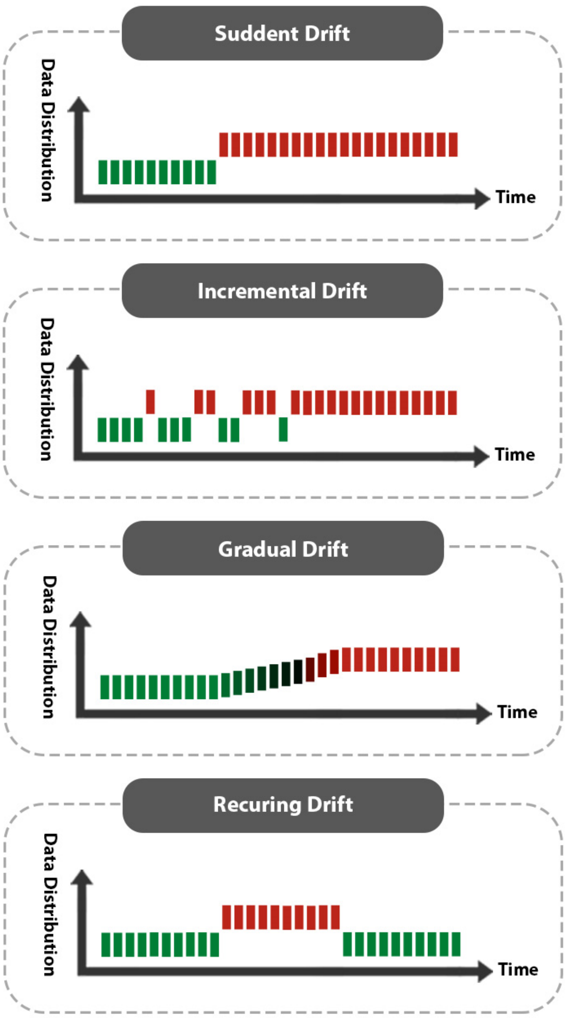

Figure 6: Types of concept drift

Yang et al. in [26], to overcome the gradual destruction of machine learning models and reduce human efforts to improve it, compared to the topic of automatic machine learning (AutoML) in the IoT as a popular field whose purpose is to select, construct, to tune and to update machine learning models to achieve the best performance in specific tasks. Accordingly, this article compared the review of existing methods in model selection, adjustment, and updating ways in AutoML to identify and summarize the optimal solutions for each stage of applying ML algorithms to IoT data analysis. In addition, the authors have conducted a case study in this field.

In [11], Abdullah et al. have considered the issue of uncertainty and time dependence in the applications of IoT and its effect on the detection of concept drift. Accordingly, this work has introduced an unsupervised framework called RADAR to maintain the high performance of the machine learning model over time. According to the claim of the authors of this article, the mentioned framework can identify the changes in the distribution of the changing data streams of the IoT using the temporal discrepancy measure and intensity-aware analyser contexts, which has been neglected in previous works.

Arif et al. in [27], have mentioned the real-time processing of heterogeneous data. The authors state that the most common problem in the dynamic environment of the IoT is performance degradation, which is mainly caused by the virtual concept drift. When statistical qualities of input features change over time and render existing machine learning models obsolete or reduce models' performance and efficiency, the problem of virtual concept drift frequently arises in a dynamic IoT context. As a result, adaptable machine learning models are required. In order to facilitate this way, the main goal of this paper is to develop an adaptive machine learning model for resource-constrained edge computing devices to handle virtual concept drift problems in a dynamic IoT environment. Convolutional neural network (CNN) and real-time transfer learning are used in this model.

In [28], the importance of preventing vulnerabilities related to IoT devices in the home network by Internet service providers has been discussed. Based on this, he has acted on developing a scalable Stentach method for monitoring a wide range of home networks and pointed out its challenges. To test its proposed method, this article collects data from 24 IoT devices from 12 home networks for six weeks to identify temporal and spatial drifts.

In [29], the importance of IoT vulnerabilities is mentioned, and it is stated that machine learning techniques have been developed to detect suspicious activities. The main disadvantage of these methods is that they work in static environments and have not been tested in dynamic environments. In the form of this article, the authors have presented an innovative intrusion detection system based on the IoT, which measures diversity to detect concept drift and, as a result, increase the efficiency of attack detection systems.

Chao et al. in [30] have mentioned the increase in the deployment of the IoT and data transfer to powerful cloud servers and have mentioned federated learning as a solution to keep the processing power close to the data generation sources and, as a result, reduce data transfer in the network. Besides this issue, maintaining the integrity of the learning model is also very important in distributed environments. Based on this, the authors have introduced the FLARE framework to face the two mentioned challenges. This framework, by actively monitoring conceptual trends while trying to reduce the information overhead of the network, ensures the performance and stability of the machine learning model.

In [31], the emphasis of recent works on the importance of distribution and less reference to the issue of privacy is mentioned. Accordingly, according to the claim of the authors in this article, for the first time, a distributed framework based on federated learning has been introduced to detect concept drift called FedStream.

Qiao et al. [32] have focused on how concept drifts affect cyberbot attack detection in IoT scenarios. The mentioned authors have used the Bot-IoT dataset, which includes legitimate and simulated IoT network traffic, to get the results of their proposed method. Using the proposed method, the accuracy of concept drift detection has resulted in a 15% to 26% improvement compared to classification models without reference drift analysis.

Wang et al. in [33] stated that interdependent behaviors among IoT devices may lead to unexpected interactions. Based on this, this article has attempted to provide a data management system based on federated learning in order to analyze the potential vulnerabilities of device interactions in the IoT. This system, called FexIoT, uses graph neural network coding based on federated learning for clustering, the purpose of which is to identify statistical heterogeneities and concept drift.

In [34], in order to overcome the challenge of concept drift in real-time data of IoT, a drift-adaptive ensemble framework called adaptive exponentially weighted average ensemble (AEWAE) is proposed, which includes three stages of pre-processing, model learning, and online ensemblings. The effectiveness of the proposed method through testing on two public IoT datasets shows that the proposed framework surpasses the current methods in achieving accuracy and effectiveness in the analysis of IoT data.

IV. Challenges and Opportunities

In the following, some open issues of this field, which are the result of theoretical studies and empirical observations of the authors of this article in the area of smart cities and factories, will be presented. It should be mentioned that we are currently trying to implement a real-world application of the IoT in the field of renewable energy, and we are impatiently eager to publish the results in the form of several research works for the scientific community and audiences interested in this field. To publish Some of the most critical challenges and opportunities are as follows.

Most of the existing research in the field of matching applications using concept drift focuses on the detection and adaptation of drifts of data streams. One of the most prone areas for research is the prediction of concept drift, which is an attractive research opportunity.

- One of the most critical challenges of the IoT environment is the development of methods related to big data analysis. This issue is important because many of the current traditional methods cannot overcome the vast volume of current streaming data.
- One of the most critical challenges in this field for researchers is the need for real-world data on the IoT. Due to security and privacy considerations, organizations and industries resist releasing their data. The currently published ones either need more ideal dataset features or are artificially generated. On the other hand, many of these limited data sets need specific validity. Considering the importance of using real-world data to value the results, the necessity of setting up data collection systems during the research process has become more evident.
- Considering the existing research gap in developing unique purpose algorithms for IoT environments and the need to use memory-efficient processing algorithms, one of the open research fields is using lightweight collective learning algorithms and selecting the best optimizer to optimize model hyperparameters.

V. Conclusion

Today, the IoT is considered one of the most important pillars of forming smart cities and developing its applications. With the emergence of big data, various artificial intelligence techniques seek to analyze these data to create value. Although anomaly detection methods have been developed for years and work well on historical data, streaming data still needs to improve on the concept drift problem caused by the highly dynamic nature of the IoT environment. A noteworthy point in this context is the need for more attention from many works in the literature to the necessity of providing real-time analysis approaches based on streaming data in the IoT. Many of these articles have presented methods and methods assuming that the environment is static, which is practically impossible in the real world. Considering the importance of this topic and the lack of prominence of the problem of recognition and adaptation of the concept drift in the IoT, the current article intends to fill this research gap and highlight this critical research field and exciting challenges and opportunities in front of it.

References

[1] P. Nesi and G. Pantaleo, “applied sciences IoT-Enabled Smart Cities : A Review of Concepts , Frameworks and Key Technologies,” 2022.

[2] E. Sisinni, A. Saifullah, S. Han, U. Jennehag, and M. Gidlund, “Industrial internet of things: Challenges, opportunities, and directions,” IEEE Trans. Ind. Informatics, vol. 14, no. 11, pp. 4724–4734, Nov. 2018, doi: 10.1109/TII.2018.2852491.

[3] C. Gan, J. Lin, D. W. Huang, Q. Zhu, and L. Tian, “Advanced Persistent Threats and Their Defense Methods in Industrial Internet of Things: A Survey,” Mathematics, vol. 11, no. 14. pp. 1–23, 2023. doi: 10.3390/math11143115.

[4] H. Mouratidis and V. Diamantopoulou, “A Security Analysis Method for Industrial Internet of Things,” IEEE Trans. Ind. Informatics, vol. 14, no. 9, pp. 4093–4100, 2018, doi: 10.1109/TII.2018.2832853.

[5] M. Bauer, L. Sanchez, and J. Song, “Iot-enabled smart cities: Evolution and outlook,” Sensors, vol. 21, no. 13, pp. 1–29, 2021, doi: 10.3390/s21134511.

[6] F. Tao, Q. Qi, A. Liu, and A. Kusiak, “Data-driven smart manufacturing,” J. Manuf. Syst., vol. 48, pp. 157–169, 2018, doi: 10.1016/j.jmsy.2018.01.006.

[7] Q. Qi, Z. Xu, and P. Rani, “Big data analytics challenges to implementing the intelligent Industrial Internet of Things (IIoT) systems in sustainable manufacturing operations,” Technol. Forecast. Soc. Change, vol. 190, no. January, p. 122401, 2023, doi: 10.1016/j.techfore.2023.122401.

[8] L. Yang and A. Shami, “A Lightweight Concept Drift Detection and Adaptation Framework for IoT Data Streams,” IEEE Internet Things Mag., vol. 4, no. 2, pp. 96–101, 2021, doi: 10.1109/iotm.0001.2100012.

[9] A. A. Cook, G. Misirli, and Z. Fan, “Anomaly Detection for IoT Time-Series Data: A Survey,” IEEE Internet Things J., vol. 7, no. 7, pp. 6481–6494, 2020, doi: 10.1109/JIOT.2019.2958185.

[10] J. C. Schlimmer and R. H. Granger, “Incremental learning from noisy data,” Mach. Learn., vol. 1, no. 3, pp. 317–354, 1986, doi: 10.1007/bf00116895.

[11] A. Alsaedi, N. Sohrabi, R. Mahmud, and Z. Tari, “RADAR: Reactive Concept Drift Management with Robust Variational Inference for Evolving IoT Data Streams,” Proc. - Int. Conf. Data Eng., vol. 2023-April, pp. 1995–2007, 2023, doi: 10.1109/ICDE55515.2023.00155.

[12] J. L. Lobo, I. Laña, E. Osaba, and J. Del Ser, “On the Connection between Concept Drift and Uncertainty in Industrial Artificial Intelligence,” vol. 101000162, pp. 2020–2021, 2023, doi: 10.1109/CAI54212.2023.00081.

[13] A. Tsymbal, “The problem of concept drift: definitions and related work,” Comput. Sci. Dep. Trinity Coll. Dublin, 2004.

[14] S. C. H. Hoi, D. Sahoo, J. Lu, and P. Zhao, “Online learning: A comprehensive survey,” Neurocomputing, vol. 459, pp. 249–289, 2021, doi: 10.1016/j.neucom.2021.04.112.

[15] J. Gama, I. Zliobaite, A. Bifet, M. Pechenizkiy, and A. Bouchachia, “A survey on concept drift adaptation,” ACM Comput. Surv., vol. 46, no. 4, Mar. 2014, doi: 10.1145/2523813.

[16] Q. Xiang, L. Zi, X. Cong, and Y. Wang, “Concept Drift Adaptation Methods under the Deep Learning Framework: A Literature Review,” Appl. Sci., vol. 13, no. 11, 2023, doi: 10.3390/app13116515.

[17] S. Amos, “When Training and Test Sets Are Different: Characterizing Learning Transfer,” Dataset Shift Mach. Learn., pp. 2–28, 2013, doi: 10.7551/mitpress/9780262170055.003.0001.

[18] K. Namitha and G. Santhosh Kumar, “Learning in the presence of concept recurrence in data stream clustering,” J. Big Data, vol. 7, no. 1, pp. 1–28, 2020, doi: 10.1186/s40537-020-00354-1.

[19] J. G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N. V. Chawla, and F. Herrera, “A unifying view on dataset shift in classification,” Pattern Recognit., vol. 45, no. 1, pp. 521–530, 2012, doi: 10.1016/j.patcog.2011.06.019.

[20] J. Lu, A. Liu, F. Dong, F. Gu, J. Gama, and G. Zhang, “Learning under Concept Drift: A Review,” IEEE Trans. Knowl. Data Eng., vol. 31, no. 12, pp. 2346–2363, 2019, doi: 10.1109/TKDE.2018.2876857.

[21] Nidhi, V. Mangat, V. Gupta, and R. Vig, “Methods to investigate concept drift in big data streams,” Knowl. Comput. Its Appl. Knowl. Manip. Process. Tech. Vol. 1, pp. 51–74, 2018, doi: 10.1007/978-981-10-6680-1_3.

[22] M. Lima, M. Neto, T. S. Filho, and R. A. Roberta, “Learning Under Concept Drift for Regression - A Systematic Literature Review,” IEEE Access, vol. 10, pp. 45410–45429, 2022, doi: 10.1109/ACCESS.2022.3169785.

[23] A. Liu, Y. Song, G. Zhang, and J. Lu, “Regional concept drift detection and density synchronized drift adaptation,” IJCAI Int. Jt. Conf. Artif. Intell., vol. 0, pp. 2280–2286, 2017, doi: 10.24963/ijcai.2017/317.

[24] G. Ditzler, M. Roveri, C. Alippi, and R. Polikar, “Learning in Nonstationary Environments: A Survey,” IEEE Comput. Intell. Mag., vol. 10, no. 4, pp. 12–25, 2015, doi: 10.1109/MCI.2015.2471196.

[25] I. Khamassi, M. Sayed-Mouchaweh, M. Hammami, and K. Ghédira, “Discussion and review on evolving data streams and concept drift adapting,” Evol. Syst., vol. 9, no. 1, pp. 1–23, 2018, doi: 10.1007/s12530-016-9168-2.

[26] L. Yang and A. Shami, “IoT data analytics in dynamic environments: From an automated machine learning perspective,” Eng. Appl. Artif. Intell., vol. 116, no. September, p. 105366, 2022, doi: 10.1016/j.engappai.2022.105366.

[27] M. Arif and D. G. Perera, “Edge Computing-based Adaptive Machine Learning Model for Dynamic IoT Environment,” 2023 IEEE Int. Symp. Circuits Syst., pp. 1–5, 2023, doi: 10.1109/ISCAS46773.2023.10181740.

[28] A. Pashamokhtari, N. Okui, M. Nakahara, A. Kubota, G. Batista, and H. H. Gharakheili, “Quantifying and Managing Impacts of Concept Drifts on IoT Traffic Inference in Residential ISP Networks.” 2023.

[29] O. A. Mahdi, A. Alazab, S. Bevinakoppa, N. Ali, and A. Khraisat, “Enhancing IoT Intrusion Detection System Performance with the Diversity Measure as a Novel Drift Detection Method,” 2023 9th Int. Conf. Inf. Technol. Trends, ITT 2023, pp. 50–54, 2023, doi: 10.1109/ITT59889.2023.10184268.

[30] T. Chow, U. Raza, I. Mavromatis, and A. Khan, “FLARE: Detection and Mitigation of Concept Drift for Federated Learning based IoT Deployments,” 2023, doi: 10.1109/IWCMC58020.2023.10182870.

[31] C. B. Mawuli et al., “FedStream: Prototype-Based Federated Learning on Distributed Concept-Drifting Data Streams,” IEEE Trans. Syst. Man, Cybern. Syst., vol. PP, pp. 1–13, 2023, doi: 10.1109/TSMC.2023.3293462.

[32] H. Qiao, B. Novikov, and J. O. Blech, “Concept Drift Analysis by Dynamic Residual Projection for Effectively Detecting Botnet Cyber-Attacks in IoT Scenarios,” IEEE Trans. Ind. Informatics, vol. 18, no. 6, pp. 3692–3701, 2022, doi: 10.1109/TII.2021.3108464.

[33] G. Wang, H. Guo, A. Li, X. Liu, and Q. Yan, “Federated IoT Interaction Vulnerability Analysis,” in 2023 IEEE 39th International Conference on Data Engineering (ICDE), 2023, pp. 1517–1530. doi: 10.1109/ICDE55515.2023.00120.

[34] Y. Wu, L. Liu, Y. Yu, G. Chen, and J. Hu, “An Adaptive Ensemble Framework for Addressing Concept Drift in IoT Data Streams.” 2023.