International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
Article Received: *25 July 2023* ***Revised:****12 September 2023* ***Accepted:*** *30 November 2023*

Advanced Techniques for Mitigating Model Drift and Enhancing the Robustness of AI/ML Models

Rajiv Avacharmal
AI/ML Risk Lead, AI/ML expert
University of Connecticut, USA
rajiv.avacharmal@gmail.com

Abstract: The implementation of AI/ML models in mission-critical applications requires rigorous drift mitigation as well as model robustness enhancement. This paper studies the state-of-the-art approaches to these problems which involve the inclusion of continuous monitoring and drift detection, data management strategies, model retraining and updating, adversarial training for robustness and stability, enhancing model interpretability and governance frameworks, and robust deployment strategies and frameworks. Continuous process control monitored through the use of statistical process control and concept drift detection algorithms enables the early detection of performance deterioration. Implementing data quality assurance, feature engineering, and augmentation processes ensures that training data exists in its representative form. The incremental learning, transfer learning and online learning used for model retraining will help with adapting to new data distributions. Adversarial training that includes gradient-based attacks and generative adversarial networks enhances resistance against changes through perturbations. Employing the above methods the organizations would be able to prevent machine drift, strengthen robustness, and secure the robust performance of machine intelligence.

Keywords: *AI, Machine Learning, Model Drift, Robustness, Monitoring, Drift Detection, Data Management, Model Retraining, and Model Updating.*

Introduction

AI and ML models have seen the technology quickly infiltrating vital services in various sectors emphasizing the need to guarantee their accuracy and dependability. Even though the models can deal with model drift which is the issue of model performance degradation where the data distribution or the environment has changed, their capacity to deal with such problems is limited. The shutting out of the model drift and the retention of AI/ML models' robustness is a burning issue involving supplied techniques and approaches.

Continuous Monitoring and Drift Detection

Detecting the drift and continuous monitoring are vital steps toward obtaining robust AI/ML models without the model drift risks. Instead of waiting for an incident to occur, these approaches allow individuals to detect early the decline of performance or the deviation of data distribution to take necessary and timely actions for recovery [1]. An efficient way of continuous monitoring is in the building of the SPC methods. SPC techniques, including control charts and cumulative sum (CUSUM) charts, are implemented to monitor model performance metrics, i.e. correctness, precision, and recall, among others [2]. Using control limits derived from history data or predefined thresholds, the charts can demonstrate deviation from expected and signify that the model is drifting [3]. Concept drift is probably the most dangerous problem for machine learning classifiers. Concept drift detection algorithms are one of the major tools for detecting changes in the underlying data distribution. Incoming data flow is examined to search for noticeable feature space twists, target variable distribution drifts, or feature-target relationships [4]. The EDM charts, DDM, and EWMA are the popular change drift detection methods [5].

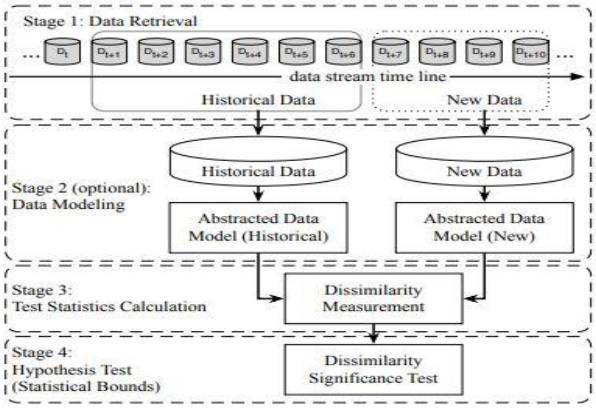

Figure 1: Data Drift Detection

(Source: https://www.analyticsvidhya.com/blog/2021/10/mlops-and-the-importance-of-data-drift-detection/)

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
Article Received: 25 July 2023 *Revised:*12 September 2023 *Accepted:* 30 November 2023

Online learning can also be quite useful in these processes of continuous monitoring and shift detection as happened. Such ways endow models to train continually on data streams in the changing data structure, not requiring full re-training otherwise [6]. Techniques such as online gradient descent, online Bayesian updating, and adaptive learning rate algorithms that can algorithmically update the model's parameters incrementally, on-the-fly adaptation to environment change is possible [7]. In addition to the utilization of algorithms, the ideal approach to a monitoring system should be based on precise data collection and logging mechanisms. It covers metadata information retrieval from the input data as well as labels, in which the previously detected errors can be analysed easily and identify the root cause [8]. Moreover, monitoring systems should be designed to enable alerts and notifications so that when drifts are detected the identified stakeholders and actions concerning them can be taken. Some of the measures may include retraining models, data auditing, or implementing emergency strategies to maintain the system's functionality and performance if necessary.

Companies should have set up clear executing measures, baseline figures as a reference, and thresholds for correct models to help firm executing bar and drift detection. Model rating methods and testing procedures should have been done regularly. For instance, holdout datasets, cross validation, and stilted collections can be employed to delineate model lustiness for the aim of the find of drift scenarios. With the help of full successive monitoring and using drift contactable methods, the model drift can be detected if it occurs. These techniques, moreover, as well as ensured the convincingness and blandness of AI/ML model in its appendage environment.

Data Management Strategies

The strategical executing of alive data direction is a key broker in minimizing model drifts and reinforcing the resilience of the AI/ML system. The methods and strategies forming the range of approaches are those which check the accuracy, relevance and the change of the data used for training and evaluating models [9]. Data type is one of the leading factors of data direction strategies.

The strategies imply setting up mechanisms and tools to trace such mistakes through different error types such as missing values, outliers, and data inconsistencies. For example, data can be cleansed, imputed and normalized in order to make it machine learning-ready [10]. They are in charge of data integrity and data quality and also make sure that the data is preserved for model training and application. The contribution of feature engineering to that is more as it generates features from the provided data itself. This is a process that involves: feature selection, feature extraction, and feature transformation which in fact make the model stronger in extracting the sense of meanings and relationships inside the data [11]. In other words, competent feature engineering can be the key to lower the negative impact of the noisy data on a model accuracy. Providing methods for updating data is crucial when data is limited or biased in order to solve model drift. These strategies are the end goal that is accomplished by the generation of synthetic data samples which are obtained by applying some transformations, e.g. flipping, rotating, and adding noise, to the current data [12]. GANs together with deep learning-based generative models can be used to generate synthetic high-quality training datasets without limits in diversity and generalizability [13]. Continuous monitoring and validation are necessary for identifying any swerves in data flows as well as adjusting them. This process can be simplified by utilizing powerful data validation pipelines that ensure data integrity, mutation, and adherence to all the specific requirements which are to be defined previously [14].

Methods like declining tests, and drifting detection algorithms which are related to domain-specific rules can be used to spot drifting in data. This causes various activities, for example, retraining the model, redoing some mending or data gain. Engaging with provenance and lineage management are the key factors, which enable transparency and responsibility throughout the data management course. Such techniques often include diligently recording the changes in data sources and elements along with the procedure followed.

Figure 2: Data Management Strategies

(Source: https://www.divectors.com/blog/data-management-strategy/)

The recorded data always has the potential of being reproduced, which is a highly desirable feature. Version-driven systems, metadata management tools and provenance tracking can be used to aid in data versioning and allow the tracking of any changes made in the data and the model performance to the initial causes. Also, the framework and policies should be fabricated to protect the use of data which will be ethical and responsible. Such frameworks should include provisions covering data protection and security issues, and compliance with the relevant rules, while data

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
Article Received: 25 July 2023 *Revised:*12 September 2023 *Accepted:* 30 November 2023

quality, accessibility, and data transparency tracking, should also be fostered. Through developing appropriate data management strategies, it will be possible to guarantee high-quality, diversity and relevance of data which will avoid the risk of a model being a victim of drift and promote the robustness and reliability of the AI/ML model.

Model Retraining and Updating

At the time of model drift detection, the model's retraining or updating becomes necessary to retain its performance and reliability. For the purpose of preventing of model drifting and maintaining its relevance during the period of rising uncertainty, consider implementing appropriate model retraining and updating techniques [15]. Partial training approaches allow models to serve various purposes by means of adapting themselves to the new data distributions or conditions transfer with no need to retrain completely. These techniques modify the model parameters not only step by step but also consistent with prior knowledge obtained then training. Algorithms like SGD with adaptive learning rates (such as Adam) can be used for gradual updating of parameter values as the data streams come in [16]. Transfer learning is the other useful technique employed for removing the effect of obsolescence from the model as well. The transfer of knowledge from pre-trained models to a new model performing this role of training a new model which is quicker and much better than the first is also known as this approach [17]. Transfer learning leads to huge savings of data and computational resources because the model takes advantage of already learned representations and weights.

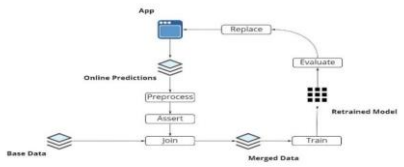

Figure 3: Model Retraining

(Source: https://towardsdatascience.com/monitoring-and-retraining-your-machine-learning-models-f385b596b285)

These models can be used to average the predictions of a discrete number of individual models to minimize overfitting and also to facilitate model updating. Ensemble techniques generate the conclusion by combining the decisions of several predictive models, including those trained on various data subsets or employing various algorithms, to increase accuracy and stability [18]. With every new data or model drift, individual models are re-trained or updated while the ensemble continues to work. Each of these models that are a part of the ensemble ensures that uninterrupted services are not interrupted. The key to this image is the productive approach to model updating strategies and deployment pipes which provide for continuous retraining and updating [19]. The model retraining workflows that are automated and conjoined with the continuous integration and deployment pipelines would aid in updating and deploying the models in a continual and automated process.

Strategies like model containerization technology and orchestration frameworks will simplify the process of seamless deployment and scaling out the new weights in production environments. Last, but no less important, it becomes necessary to provide some precise standards and boundaries to introduce model retraining or renewing this model. The criteria for controlling model updates should be founded on either alert algorithms that work with specified performance metrics, or rules in the case of a specific domain, which will guarantee that applicable changes will be introduced only in the right time frame. Through the implementation of appropriate model retraining and updating approaches, organizations can keep the models' performance and trustworthiness at the same time. As a result, they can diminish the impact of model drift and ensure the continued development of their models in dynamically evolving environments.

Adversarial Training and Robustness

Adversarial training techniques are specifically enforced to keep the models of AI/ML models robust against adversarial attacks and input variations. These approaches consist of educating the models to be able to distinguish and react against these attacks through an advanced type of training known as adversarial learning [20]. The use of gradient-based attacks appears to be another common adversarial training approach. These assaults use the partials of the model loss function accounted for the input data to generate guaranteed inaccurate predictions by some means. Adding the adversarial examples to the training dataset helps in skillfully dealing with the fallout of such attacks [21]. The model becomes more robust to future attacks as a result. Generative adversarial networks (GANs) can do well for adversarial learning as well.

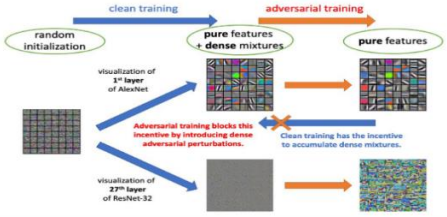

Figure 4: Adversarial Training

(Source: https://www.microsoft.com)

A generative model is trained to generate a variety of adversarial examples that can successfully bypass the target model, while a discriminator network is trained to distinguish between the real and the adversarial examples. In which, the target model is taught to recognize and defend against a variety of adversarial attacks via the guidance-strengthening process [22]. Apart from the mentioned methods, defensive distillation is also an effective technique for improving the robustness of the system. This approach applies training of a secondary model to emulate the behaviour of the primary model while being more robust to manipulations by adversarial attacks [23]. By replacing the main objective function with the refined one, the secondary model is trained to generate smoother outputs less unstable to the attacks of adversaries. The use of adversarial training can simultaneously be coupled with other methods of data augmentation and regulation to yield even better model defenses.

By acting as data augmentation, other adversarial examples can then be created to make the training data more diverse as well as broad. Regularization techniques, including dropout and weight decays, can be employed to avoid overfitting and improve the model's generalization power which in turn keeps the adversarial attacks at bay It is a fact that the adversarial training should be organized in a systematic and controlled manner making sure that the mitigation techniques used do not violate any additional assumptions or stipulations of the problem area. Furthermore, close monitoring and assessment of the adversarial training process should be made to check that the model's performance for the clean, non-adversarial data is unharmed. Using an adversarial training approach, companies can make their AI/ML models robust, fighting adversary attacks and input perturbation, and ensuring credibility and reliability of models where they are critical.

Model Interpretability and Governance

Providing model interpretability and developing reliable regulatory frameworks as keys to model drift mitigation and responsible AI/ML model deployment has emanated. The models with interpretability help to comprehend the model behaviour and decision-making mechanisms and are useful for performance monitoring and model drifts. Approaches like feature importance, partial dependency plots, and Local Interpretable Model-Agnostic Explanations (LIME) will assist in the interpretation of the causality of the model's predictions, therefore, it will be easier to identify any biases/drift in the decision-making procedure.

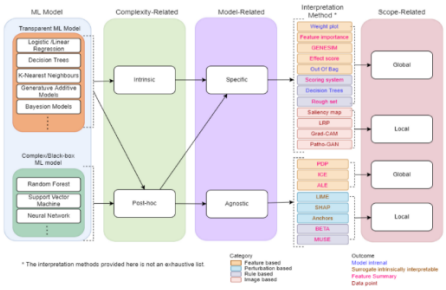

Figure 5: Interpreting ML Models

(Source: https://www.mdpi.com/2073-8994/13/12/2439)

An effective governance structure should therefore contain ethics, compliance and risk management practices among others to ensure the responsible use and operationalization of AI/ML models. These frameworks should set out clear guidelines and protocols for model accuracy progress, and update processes that face evolvement with regulations, laws, and industrial standards. Furthermore, the philosophies of governance should include topics about data privacy, security and fairness, making the transparent and responsible development process. Transparent and constant monitoring using audits and impact assessments to detect and alleviate risks or unintended consequences arising from AI/ML model applications should be exercised.

Deployment Strategies and Frameworks

Highly sophisticated deployment strategies and frameworks are the key factors for the minimization of model drift and better frameworks of AI/ML models. Ensembling goes further as it makes individual models negligible and harnesses more robust results. Containerizing and one-model orchestration platforms, such as Docker and Kubernetes, become a basis for convenience, scaling, and upgrading AI/ML models. Developed on the principles of the federated learning architecture, decentralized model training and deployment algorithms allow models to learn from their surroundings and adapt to different environmental conditions. Transfer learning methodology, which involves the use of edge computing architecture where the models are deployed and executed on the edge devices or gateways near to data source can also help to address this problem by processing the data locally and updating models in real time.

The creation of durable release channels for CI/CD ensures that the procedure of models' updating and fielding is adequately covered up. Automation in testing, validation, and monitoring can be added to the upstream pipeline which will help to perform an overall and excellent evaluation of the system, before implementing that system. Performance monitoring and drifting mechanisms are also essential parts of deployment frameworks. These comprise the side of

telemetry, monitoring, and logging that allows collection of runtime data and measurement of the model performance for continuous monitoring, post-deployment analysis and model refinement.

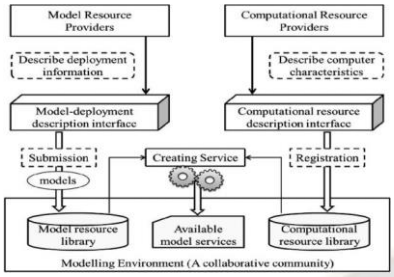

Figure 6: Model Deployment Strategies

(Source: https://www.researchgate.net/publication)

Conclusion

Reducing drift events and improving the Resilience of AI/ML Models is a multi-dimensional problem that requires many advanced methodologies and the use of various tactics. To achieve this purpose, various measures such as the continual monitoring of drifting and the implementation of effective data management are proposed as well as the updating and retraining of models. The other measures include resisting adversarial attacks and the development of interpretable AI. Additionally, proper governance framework and complete deployment strategies and guidelines are also used as an important means of ensuring the reliability and trust of AI/ML in key applications.

References:

[1] Zliobaite, I. (2010). Learning under concept drift: an overview. arXiv preprint arXiv:1010.4784.

[2] Gama, J., Medas, P., Castillo, G., & Rodrigues, P. (2004, May). Learning with drift detection. In Brazilian symposium on artificial intelligence (pp. 286-295). Springer, Berlin, Heidelberg.

[3] Klinkenberg, R., & Renz, I. (1998, April). Adaptive information filtering: Learning in the presence of concept drifts. In Learning for Text Categorization (pp. 33-40).

[4] Tsymbal, A. (2004). The problem of concept drift: definitions and related work. Computer Science Department, Trinity College Dublin, 106(2), 58.

[5] Baena-García, M., del Campo-Ávila, J., Fidalgo, R., Bifet, A., Gavaldà, R., & Morales-Bueno, R. (2006, September). Early drift detection method. In Fourth international workshop on knowledge discovery from data streams (Vol. 6, pp. 77-86).

[6] Losing, V., Hammer, B., & Wersing, H. (2018). Incremental on-line learning: A review and comparison of state of the art algorithms. Neurocomputing, 275, 1261-1274.

[7] Gepperth, A., & Hammer, B. (2016). Incremental learning algorithms and applications. In European symposium on artificial neural networks (ESANN) (pp. 357-368).

[8] Žliobaitė, I., Pechenizkiy, M., & Gama, J. (2016). An overview of concept drift applications. Big data analysis: new algorithms for a new society, 16, 91-114.

[9] Khamassi, I., Sayed-Mouchaweh, M., Hammami, M., & Ghédira, K. (2018). Discussion and review on evolving data streams and concept drift adapting. Evolving systems, 9(1), 1-23.

[10] García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J. M., & Herrera, F. (2016). Big data preprocessing: methods and prospects. Big Data Analytics, 1(1), 9.

[11] Khaire, U. M., & Dhanalakshmi, R. (2019). Stability of feature selection algorithm: A review. Journal of King Saud University-Computer and Information Sciences.

[12] Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. Journal of big data, 6(1), 1-48.

[13] Wang, Y., Gonzalez-Garcia, A., Berga, D., Herranz, L., Khan, F. S., & van de Weijer, J. (2019). Minegan: effective knowledge transfer from GANs to target domains with few images. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 9332-9341).

[14] Polyzotis, N., Roy, S., Whang, S. E., & Zinkevich, M. (2017). Data management challenges in production machine learning. In Proceedings of the 2017 ACM international conference on management of data (pp. 1723-1726).

[15] Ditzler, G., & Polikar, R. (2013). Incremental learning of concept drift from streaming imbalanced data. IEEE transactions on knowledge and data engineering, 25(10), 2283-2301.

[16] Gepperth, A., & Hammer, B. (2016). Incremental learning algorithms and applications. In European symposium on artificial neural networks (ESANN) (pp. 357-368).

[17] Sayed-Mouchaweh, M., & Lughofer, E. (2014). Prologue: Aspects of incremental learning. In Incremental Learning in Nonstationary Environments (pp. 1-11). Springer, Cham.

[18] Minku, L. L. (2019). Transfer learning in non-stationary environments. In Springer Handbook of Computational Intelligence (pp. 1203-1217). Springer, Berlin, Heidelberg.

[19] Bose, A. J. C., Aarabi, A., & Raimondo, J. (2021). Machine Learning Operations (MLOps): Overview, Definition, and Architecture. arXiv preprint arXiv:2205.02302.

[20] Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., & Jordan, M. (2019). Theoretically principled trade-off between robustness and accuracy. In International Conference on Machine Learning (pp. 7472-7482). PMLR.

[21] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018, July). Towards deep learning models resistant to adversarial attacks. In International Conference on Learning Representations.

[22] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.

[23] Papernot, N., McDaniel, P., Wu, X., Jha, S., & Swami, A. (2016, May). Distillation as a defense to adversarial perturbations against deep neural networks. In 2016 IEEE symposium on security and privacy (SP) (pp. 582-597).