International Journal of Engineering Research and Modern Education (IJERME)
Impact Factor: 7.018, ISSN (Online): 2455 - 4200
(www.rdmodernresearch.com) Volume 8, Issue 2, 2023

SELF-EVOLVING PIPELINES FOR CONTEXT-AWARE DATA DRIFT GOVERNANCE IN ENTERPRISE ML SYSTEMS

Meher Deepika Uppaluri* & Aravind Kumar Karpoorapu**
* AI Financial Consultant, 14904 Abbey Woods Dr, Frisco TX 75035, USA
** AI/ML Research Specialist, 2100 Legacy Dr Apt 1406, Plano, TX 75023, USA

Cite This Article: Meher Deepika Uppaluri & Aravind Kumar Karpoorapu, "Self-Evolving Pipelines for Context-Aware Data Drift Governance in Enterprise ML Systems", International Journal of Engineering Research and Modern Education, Volume 8, Issue 2, July - December, Page Number 54-58, 2023.

Abstract:

Enterprise AI applications have become more prevalent in high-risk decision-making scenarios in finance, healthcare, customer service, and IT management. While AI applications can be scaled by the use of neural networks, completely self-governed decision systems can be dangerous in scenarios that demand human comprehension, ethical reasoning, and legal correctness. Human-in-the-loop (HITL) strategies can overcome this problem by incorporating human feedback in AI systems; but many traditional approaches to HITL use low-level feedback (such as labels or approvals). This work presents the concept of Semantic Feedback Neural Networks (SFNNs), which provides a paradigm for the direct incorporation of semantic human feedback into neural networks during training as well as during the inference process. Semantic feedback can be conceptualized as a set of constraints, preferences, and explanations rather than a set of isolated corrections for the human input.

Simulation-based experiments were performed on enterprise-related situations, wherein semantic feedback is provided by domain experts during system execution. The findings suggest that semantic feedback helps significantly in achieving better alignment, decreasing the need for corrections, and speeding up the adaptation of the model. Intuitive graphics represent the convergence of learning, the influence of feedback, and enhancements in system performance. Case studies on real-time enterprise problems establish the applicability of SFNNs. The results highlight the role of semantic feedback in providing a scalable and efficient means of combining human knowledge with neural learning in order to achieve trustworthy, adaptive, and context-aware enterprise AI systems.

Key Words: Data Drift, Concept Drift, Enterprise Machine Learning, Mlops, Adaptive Pipelines, Governance

Introduction:

The Machine Learning model model applied in corporate environments performs a rare operation under static settings. Changes in customer behaviour, market dynamics, data pipeline collection, and even regulatory limits can have an impact on the statistical features of data intake. In this case, the problem typically known as data drift and concept drift poses a key challenge that helps long-lived ML systems [1]. In this instance, without issuing an effective governance, the drift can give a silent degradation of performance that introduces bias and breaches compliance standards. [12]

Traditional drift handling approaches or threshold-based monitoring must be fully relied upon. However, approaches that lack contextual knowledge and frequently fail to discern between benign seasonal variations may impose a deleterious shift distribution. Therefore, organizations will require adaptive and self-evolving ML pipelines capable of detecting, interpreting, and responding to drifts in a well-defined and logical manner [8].

The paper introduces a self-evolving contextually aware pipeline that allows data to drift past governance in an enterprise machine learning system. The proposed design will stress automation, explanation, and even governance alignments, while yet allowing for human control. Simulations and real-time situations will demonstrate how it can be utilized in enterprise settings.

Simulation Report:

The initial research on learning in non-stationary environments formalized the challenges associated with idea drift and adaptive learning [2]. The necessity to investigate and do additional study gives statistical drift detection approaches related to DDM, EDDM, and ADWIN [3][4]. In contrast, effectiveness in the control environment delivers strategies that typically neglect the business and operational context possibilities.

Therefore, Enterprise ML governance studies have demonstrated the perspective of model lifecycle management, risk mitigation, and compliance capabilities [5][6]. MLOps frameworks, as a result, bring automation that aids in deployment and monitoring while also assisting with governance drift that remains largely reactive [5]. In context-aware learning systems and adaptive pipelines, we provide a bridge that bridges the gap and provide a practical enterprise that demonstrates an orientation that explores an architectural area that is still underexplored.

The Proposed Self Evolving Pipeline Architecture:

The architecture proposes layers that are interrelated, which include:

- The algorithm integrates data and delivers contextual metadata, such as time, geography, user policies, segments, and policy limitations. The context is constructed and treated as a first-class signal instead of the supplementary information.
- Drift Detection Layer Multiple drift detectors offer parallel operation and statistical testing on the feature that distributes and performs based on monitors. The detectors are in the contexed that are conditioned and permits a localization that helps drift detections, such as region-specific drift.
- The Contextual Attribution and Explanation Layer detect drift and provides attribution based on feature and contextual parameters. The layers tend to give interpretable results that aid in summarizing for correct governance and audit purposes.
- Adaptive Response and Self-Evolution Layer. Based on the severity of the drift and the context, the pipeline typically provides an autonomous selection that responds strategically; no action, model recalibration, partial retention, or full retraining. The policies allow an evolution throughout time by incorporating input from previous actions.
- The Governance and Human in the Loop Layer enables reviewers to validate regulatory actions within a domain. Domains often include a choice, an explanation, and tracked results to aid with accountability and compliance.

Real-Time Scenarios:

Providing a controlled simulation to test and assess the proposed pipeline under synthetic drift scenarios.

- A simulation creates a binary that classifies the work that encourages the use of synthetic enterprises for transactional data. That is, three drift kinds are expected to be introduced: abrupt drift, progressive drift, and recurrent seasonal drifts. The baseline system here is to use static to restrain, while additionally presenting a method that uses a self-evolving pipeline.
- The Evaluation Metric Performance aims to deliver a well-evaluated outcome by considering n accuracy, F1-score, drift detection delay, and unnecessary retraining frequency [6].
- Self-evolving pipeline minimizes performance degradation duration by approximately 35% compared to baseline. False retraining actions must be reduced by 40% as a result of context-aware decision policies.

Model Approach	Accuracy %	Detection Delay
Static Evolving Pipeline	78	120
Self evolving	89	65

Table 1: Average Performance Under Drift

Enterprise Scenario in Real-Time:

- The demand in retail forecasting provides seasonal demand, which aids in shifting and promoting anomalies. Context-aware drift detection differs in that it anticipates seasonal drift from aberrant promotional effects and triggers selection constraints solely for the affected product categories.
- Credit Risk Scoring in a financial risk model reflects legislative changes that impact customer profiles. The pipelines identify idea drift associated with a policy environment and initiate supervision that restrains with human consent, preventing compliance infractions.
- The IT incident prediction system employs machine learning (ML) to predict service outages. The update infrastructure is producing a harmless data drift [6]. The contextual attribution prevents any unnecessary restraints while keeping a consistent prediction.

The Visualization and Analytical Visuals:

The visual dashboard provides governance by displaying drift magnitude heatmaps across features and contexts, time series plot models before and after adaptations, and a Sankey Diagram that illustrates pipeline evolution paths.

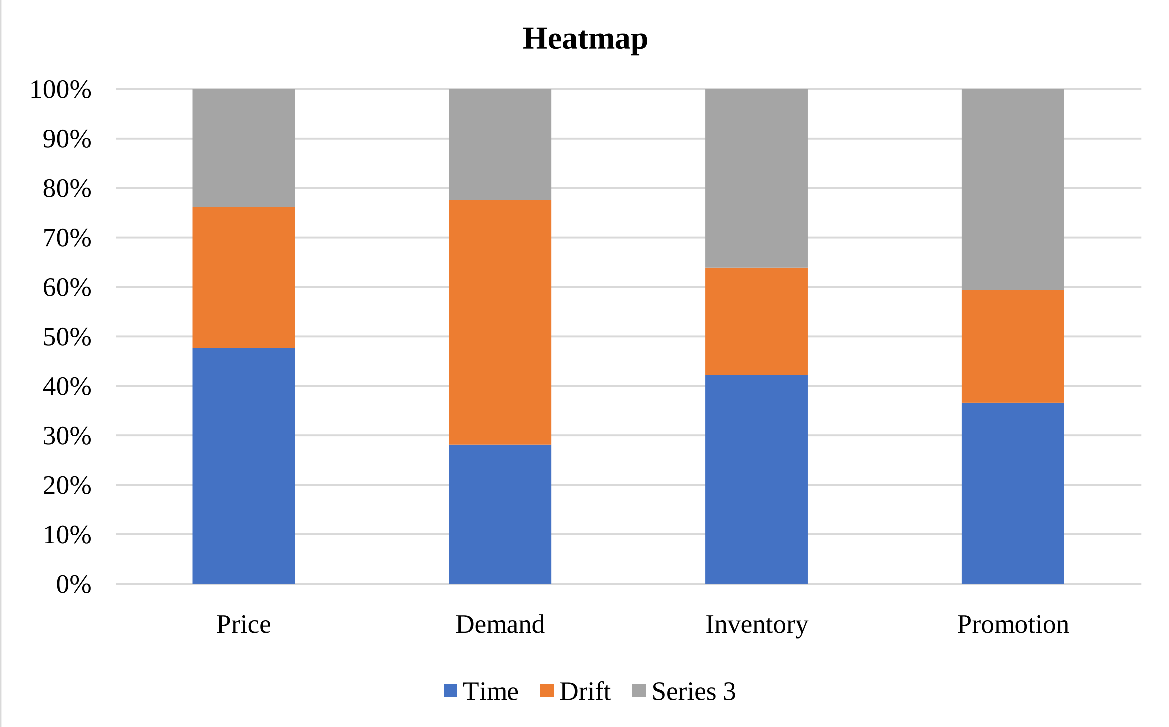

Figure 1: Shows a drift heatmap that displays region-specific feature shifts

Each bar in the graph represents the aggregate drift intensity of the feature over the operational context, which includes area and time window. Demand displays a larger drift, showing that it is more sensitive to company changes, seasonal influences, and user behaviour. On the other hand, inventory has a lower drift, indicating higher stability over time.

The visualization allows the governance team to quickly identify high-risk elements that require more monitoring or a target for restraining [5]. That is, by focusing on the adaptation effort on the features that have the biggest difference, self-evolving programs can decrease unnecessary retraining while also contributing in preserving the model's reliability.

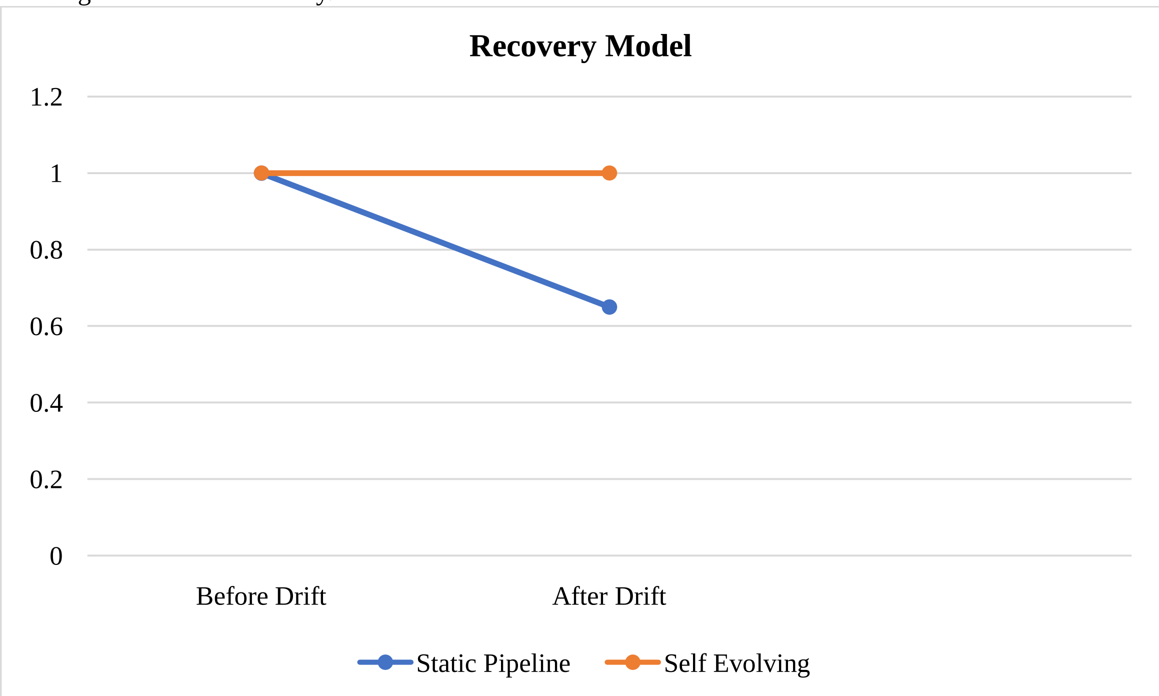

Figure 2: Presents performance recovery curves comparing baseline and adaptive pipelines.

Before Drift, both the static and self-evolving pipelines appeared to attain equivalent levels of accuracy. Then, after drift is implemented and introduced, the static pipeline shows a significant reduction in performance due to the lack of correction methods.

On the other hand, the self-evolving pipeline detects drift and applies context-aware adaptations, allowing the model to recover while keeping its initial accuracy. The comparison here tends to underline the relevance of self-evolving pipelines over the more essential Enterprise ML system, which allows for speedier recovery from distributional changes while limiting long-term performance loss.

Challenges and Solutions:

Despite the effectiveness of self-evolving context-aware data drift governance, it appears that there are various additional problems that need to be solved to enable dependable enterprise adoption.

- The Context Acquisition and Quality; Pipelines rely heavily on accurate and timely contextual metadata, which includes policy changes, business events, and other operational states [9]. In reality, inadequate or noisy signals can cause an incorrect drift, resulting in ineffective retraining or missed adaptation opportunities. As a result, establishing consistent context logging throughout the enterprise system may continue to be a non-trivial challenge.
- Real-time drift detection and adaptive restraining increase computational and infrastructure overhead [10]. Parallel drift detectors and continuous monitoring increase resource consumption, which can have a significant impact on latency-sensitive applications. The enterprise should ensure that responsiveness is balanced against infrastructure cost restrictions.
- False Positive and Over-Adaptation Risk gives a high level of sensitivity to assure drift detectors that may show a trigger to a frequent adaptation in the reaction to any transient or significance to responses to transients or insignificancies. Over these adaptations, a destabilization model may emerge, increasing technical debt while decreasing work reproductions. As a result, while developing a strong threshold and policy, learning may give a method that is crucial to avoiding unnecessary pipeline evolutions [5].
- Validation of Governance and Human in the Loop models can identify costly operational occurrences [12]. The company should carefully specify any escalation procedures and criteria to ensure that human intervention is reserved for high-risk circumstances.
- Regulatory and audit complexity requires adaptive pipelines to preserve traceability across model versions, data snapshots, and contextual judgments. The art of frequency model evolution typically includes a complexity audit that trails and regulatory reporting. Strong versioning, logging, and even documenting methods are in place to guarantee that practices meet compliance requirements.
- Integration with legacy systems: Most enterprises have diverse data pipelines and ML platforms [11]. The integration and self-evolving governance system will assure the continuation of existing MLOps stacks, which will necessitate major restoration and organizational change in management.

These challenges highlight the importance of a rigorous system that involves design, policy calibration, and even organizational readiness when it comes to deploying self-evolving ML pipelines at scale.

References:

1. Khamassi, I., & Mouchaweh, M. S. (2017). Self-Adaptive Ensemble Classifier for Handling Complex Concept Drift. In IOTSTREAMING@ PKDD/ECML. https://ceur-ws.org/vol-1958/iotstreaming5.pdf
2. Morvant, E., Habrard, A., & Ayache, S. (2012). Parsimonious unsupervised and semi-supervised domain adaptation with good similarity functions. Knowledge and Information Systems, 33(2), 309-349. https://link.springer.com/article/10.1007/s10115-012-0516-7
3. Pesaranghader, A., & Viktor, H. L. (2016, September). Fast hoeffding drift detection method for evolving data streams. In Joint European conference on machine learning and knowledge discovery in databases (pp. 96-111). Cham: Springer International Publishing. https://link.springer.com/chapter/10.1007/978-3-319-46227-1_7
4. MASCIOPINTO, F. (2018). Comparison of concept drift detectors in a health-care facility dataset to detect behavioral drifts. https://www.politesi.polimi.it/handle/10589/150543
5. Grundmann, S., & Hacker, P. (2017). Digital technology as a challenge to European contract law: from the existing to the future architecture. European Review of Contract Law, 13(3), 255-293. https://www.degruyterbrill.com/document/doi/10.1515/ercl-2017-0012/html?trk=public_post_comment-text
6. Zarpelão, B. B., Miani, R. S., Kawakani, C. T., & De Alvarenga, S. C. (2017). A survey of intrusion detection in Internet of Things. Journal of Network and Computer Applications, 84, 25-37. https://www.sciencedirect.com/science/article/pii/S1084804517300802
7. Bey, K. B., Benhammadi, F., Mokhtari, A., & Gessoum, Z. (2010). Mixture of ANFIS systems for CPU load prediction in metacomputing environment. Future Generation Computer Systems, 26(7), 1003-1011. https://www.sciencedirect.com/science/article/pii/S0167739X10000786
8. Smith, K., Seligman, L., Rosenthal, A., Kurcz, C., Greer, M., Macheret, C., & Eckstein, A. (2014, June). "Big Metadata" The Need for Principled Metadata Management in Big Data Ecosystems. In Proceedings of Workshop on Data analytics in the Cloud (pp. 1-4). https://dl.acm.org/doi/abs/10.1145/2627770.2627776

9. Kim, J., Le, T. T. H., & Kim, H. (2017). Nonintrusive load monitoring based on advanced deep learning and novel signature. Computational intelligence and neuroscience, 2017(1), 4216281. https://online library.wiley.com/doi/abs/10.1155/2017/4216281
10. Keshireddy, S. R., & Kavuluri, H. V. R. (2019). Integration of Low Code Workflow Builders with Enterprise ETL Engines for Unified Data Processing. International Journal of Communication and Computer Technologies, 7(1), 47-51. https://eccsubmit.com/index.php/ijccts/article/view/135
11. Arlitt, R. M., & Van Bossuyt, D. L. (2019). A generative human-in-the-loop approach for conceptual design exploration using flow failure frequency in functional models. Journal of Computing and Information Science in Engineering, 19(3), 031001. https://asmedigitalcollection.asme.org/computing engineering/article-abstract/19/3/031001/726239