




Automated Data Quality

Improvement System Using Self-Learning Anomaly Detection

Made by: Suhani(12320321), Sunny Kumar(12312528), Faraz
Ahmad Khan(12313142)






Introduction

It is an AI-powered system designed to automatically detect and fix poor-quality data using machine learning. It focuses on two main tasks:


It will use self-learning anomaly detection and dynamic data augmentation to improve the quality of the data. Autoencoders, synthetic data generation, and the like are utilized in machine learning to continuously update data quality without much human involvement. This would be a much needed improvement of data reliability for domains such as finance, healthcare, and IoT.





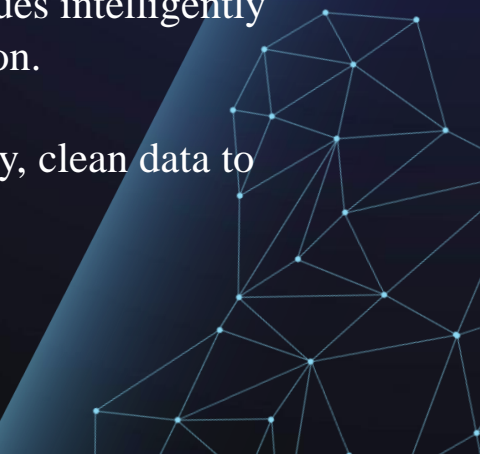
Problem Statements



- Data quality issues affect analytics, machine learning, and decision-making.
 - Challenges include:
 - Missing/incomplete data
 - Noisy/redundant entries
 - Manual cleaning limitations
 - Inability to handle dynamic, real-time data
- 



Objectives


- **Automate End-to-End Data Quality Management:** Eliminate the need for manual intervention in data cleaning, anomaly detection, and augmentation.
 - **Enable Real-Time Anomaly Detection and Correction:** Identify and fix inconsistencies, outliers, and missing data on-the-fly as data flows in.
 - **Improve Dataset Completeness and Consistency:** Fill missing values intelligently using machine learning-based data imputation and synthetic generation.
 - **Enhance Machine Learning Model Accuracy:** Provide high-quality, clean data to ensure better model training and prediction outcomes.
- 





Self-Learning Methods

Self-learning anomaly detection methods automatically identify anomalous patterns in data without relying heavily on labeled datasets. These methods adapt to new data and changing conditions over time, making them particularly suitable for dynamic systems. Unlike traditional techniques, which use fixed rules, self-learning models utilize machine learning algorithms to continuously learn from incoming data, thereby improving anomaly detection in various applications including cybersecurity, fraud detection, and predictive maintenance.

- Detects anomalies without labeled datasets
 - Adapts to data patterns over time
 - Supports use cases in finance, healthcare, IoT, cybersecurity
 - Includes Graph-based, RL-based, and Time-Series detection methods
- 

Types of Detection

There are several types of self-learning anomaly detection methods:

- **Unsupervised methods** identify anomalies by clustering and density-based algorithms without prior labeled data.
- **Semi-supervised methods** are trained on normal data, detecting anomalies based on learned patterns.
- **Reinforcement learning approaches** adapt over time using feedback, ideal for environments requiring dynamic detection like finance.
- **Graph-based methods** analyze relationships within data structures to find anomalies.
- **Time-series detection** focuses on sequential data to identify deviations in trends and patterns over time.

Applications

Self-learning anomaly detection has diverse applications across numerous industries:

Healthcare: It can help identify abnormal patient data, thus enhancing diagnosis accuracy.

Finance: Uses techniques that are crucial for detecting fraud patterns in transactions.


IoT sector: Anomaly detection can monitor sensor data to preemptively identify faulty devices or potential breaches. By continuously learning from data, these systems ensure enhanced operational efficiency and reliability.





Dynamic Data Augmentation

Dynamic data augmentation involves real-time modifications of datasets to improve model performance. Techniques often include introducing noise, altering data distributions, or synthesizing new data points based on existing ones. By continuously updating the dataset with augmented variations, models are trained to be more robust and generalize better to unseen data. This adaptability is critical in rapidly changing environments, leveraging AI's capabilities to create high-quality training datasets.

- Generates real-time synthetic data
 - Uses techniques like geometric transforms, noise injection, text augmentation
 - Helps build robust, diverse datasets
 - Enables better generalization and model performance
- 





Integration with AI

Implementing data augmentation in AI systems enhances model training by providing diverse data scenarios. AI algorithms, particularly deep learning models, benefit from exposure to varied data characteristics, allowing them to learn more generalized features effectively. Moreover, integrating augmentation techniques with machine learning pipelines leads to improved accuracy and reduces overfitting, ultimately ensuring that the model performs well in real-world applications across varied datasets.





Conclusions

In summary, automated systems that leverage self-learning anomaly detection and dynamic data augmentation play a fundamental role in enhancing data quality. These methodologies not only identify and correct anomalies effectively but also enrich datasets, leading to better decision-making in important sectors such as healthcare, finance, and IoT. Continuous advancements in these areas promise improved efficiencies and insights in data-driven applications.





Thank You!