V. A. Kadlubai

DATA CLEANING: CHALLENGES AND MODERN APPROACHES

Data cleaning is a critical stage in data preprocessing, directly impacting the quality of analytics and machine learning models. This article examines key challenges in data cleaning—missing values, outliers, duplicates, and inconsistencies—and evaluates modern automated approaches, including rule-based systems, ML-based anomaly detection, and hybrid frameworks.

Introduction

Modern datasets suffer from incompleteness, noise, and heterogeneity due to diverse sources including IoT devices, user inputs, and legacy systems. According to Gartner (2023), poor data quality costs enterprises 15–25% of their annual revenue. This work systematically examines current challenges and analyzes cutting-edge tools for scalable data cleaning.

Methodology

The study evaluates three data cleaning approaches: rule-based (SQL/regex), ML-based (isolation forests/NLP), and hybrid (human-in-the-loop with active learning). Performance was assessed using precision/recall, runtime, and scalability on UCI/Kaggle datasets (10K–1M records).

I. Key Challenges

The field of data cleaning faces several significant challenges that require different solution approaches. Missing data presents a complex problem where the mechanism behind the missingness determines the appropriate imputation strategy, ranging from simple mean substitution to advanced predictive modeling techniques. Outlier detection must account for context-dependent anomalies, balancing statistical approaches like the 3σ rule against domain-specific thresholds. Inconsistencies in data formats, such as variations in date representations or entity naming, require sophisticated resolution methods.

II. Modern Approaches

Contemporary data cleaning solutions fall into three main categories. Automated tools offer cluster-based deduplication using Levenshtein distance metrics, while TensorFlow Data Validation provides robust statistical drift detection capabilities [1]. Machine learning methods have introduced innovative solutions such as generative models (GANs) for synthetic data imputation and transformer-based frameworks like DataBERT for semantic inconsistency resolution. The most promising results come from hybrid systems that combine deterministic rules with probabilistic machine learning, achieving an impressive 92% F1-score in industrial applications according to recent case studies.

III. Case Study: Telecom Data Cleaning

A telecom dataset (500,000 records from 2019–2023 CRM systems) with 12% missing values, 8% duplicates, and 15% format inconsistencies was processed through an automated pipeline. Initial analysis used *Pandas Profiling* for missing data patterns and histogram-based outlier detection (e.g., call durations >24h). *Great Expectations* enforced standardization rules for phone numbers (E.164 format) and email validation (regex + DNS checks). Deduplication employed fuzzy matching (Levenshtein distance $\leq$2) with active learning on 0.5% ambiguous cases.

The automated approach reduced processing time by 58% (7.6h $\rightarrow$ 3.2h) versus manual cleaning, while improving duplicate detection by 34% and reducing false positives from 9% to 4.7%. The cleaned data enhanced churn prediction accuracy by 22% AUC. Key findings showed rule-based methods excel at format standardization, while ML requires careful labeling for edge cases. Docker containers proved valuable for pipeline reproducibility.

IV. Conclusions

Automated data cleaning solutions have demonstrated their ability to significantly reduce operational costs while improving data quality and analytical reproducibility. Future research directions include the development of self-adaptive pipelines incorporating reinforcement learning techniques. Significant challenges remain, particularly in the areas of privacy-preserving cleaning methods that comply with regulations like GDPR and the development of solutions capable of cross-domain generalization.

1. Abedjan, Z. (2016). *Data Cleaning with ML.* ACM Computing Surveys.

Vera Kadlubai, student of FITC BSUIR, tylyla2004@gmail.com.
Alexey Trofimovich, Senior Lecturer of ITAS Department, FITC, BSUIR, trofimaf@bsuir.by