**Quality-Driven Co-Optimization of Data Cleaning and Clustering: Framework and Mechanistic Evidence**

**Abstract**

# **Real-world data often contains multiple types of errors, such as missing values and anomalies. Such fluctuations in data quality expand the search space of unsupervised AutoML and compromise the stability of the results. This paper proposes a core hypothesis: there is a predictable structure between data quality and clustering features. As long as sufficient data quality and algorithm features are captured, high-potential cleaning and clustering combinations can be screened out in advance. Based on this hypothesis, we developed three strategies: in the offline phase, feature enhancement is used to improve candidate hit rates; in the search phase, pruning is used to compress the space to O(k log n); and in the online phase, clustering statistics are monitored and parameter drift is dynamically corrected. The experimental section follows a process of “compatibility → mechanism → gain” and verifies that the proposed scheme improves runtime speed by more than 8 times compared to full search on datasets of different scales, while the average loss in clustering quality does not exceed 2%. The results demonstrate that the proposed collaborative optimization framework can achieve both speed and robustness in large-scale scenarios involving dirty data.**

# **1 Introduction**

In data-intensive scenarios, such as precision medicine, financial risk management, and industrial IoT, unsupervised clustering has become a crucial tool for discovering potential patterns and supporting informed decision-making. However, clustering algorithms are susceptible to data distribution: a single spelling error, an unlabeled anomaly, or a missing interval can distort the distance structure and density estimates, thereby disrupting cluster partitioning, convergence paths, and even downstream analysis chains. Unlike supervised learning, which relies on labels to filter out noise, clustering is almost entirely dependent on the quality of the raw data. An intuitive solution is to leverage AutoML to chain together several data cleaning operators with various clustering algorithms, then let an optimizer automatically select the best combination in the parameter space. However, as the number of cleaning operators, clustering models, and hyperparameter dimensions grows exponentially, optimizers often waste computational resources on a large number of doomed-to-fail solutions, thereby undermining the value of automation itself.

Over the past decade, data cleaning and clustering algorithms have each made significant breakthroughs: for multi-source heterogeneous data, data cleaning methods such as missing value imputation, anomaly detection, and fault-tolerant matching have emerged in rapid succession; while K-Means, DBSCAN, hierarchical clustering, and even deep embedding clustering have demonstrated good adaptability in complex data structures. However, **the two technical approaches have developed almost independently**: data cleaning research has focused on local error correction, with little consideration of the cascading effects of corrections on unsupervised tasks; clustering research has assumed that the input is “sufficiently clean,” and has primarily focused on similarity measurement and optimization strategies. As data volume and data quality issues grow in tandem, this disconnect leads to threefold amplified complexity: data cleaning alters distance distributions, causing existing hyperparameters to become mismatched; data quality varies greatly across tables and batches, making it challenging to apply a single strategy; and the exponential number of candidate combinations causes system overhead to skyrocket, which cannot be offset linearly by increasing computing power. **The industry has attempted to utilize AutoML to link cleaning and clustering**; however, in the absence of prior knowledge, this approach still falls into the trap of “blind trial and error,” with input and output often being disproportionate.

Given this practical challenge, this paper proposes a core hypothesis supported by empirical evidence: **the impact of data quality on clustering performance has a learnable structure**. In other words, as long as the characteristics of data contamination and algorithm features are described in sufficient detail, the quality of the cleaning-clustering pipeline can be predicted before time-consuming evaluation. This hypothesis leads to two direct implications: first, construct a mapping Φ that maps data quality feature vectors to the priority of candidate pipelines; second, if Φ is a reliable predictor, we can boldly prune the pipeline at the beginning of the search, retaining only the top-ranked candidates and concentrating computational resources on the most promising few solutions.

Based on this hypothesis, we developed a three-stage framework that strikes a balance between speed, robustness, and interpretability. First, in the **feature enhancement stage**, we introduce higher-order interaction features and operators in addition to traditional statistics, and use historical experimental data to train a gradient boosting ranking model to implement Φ and generate a ranking of candidate pipelines. Next, in the **confidence upper bound pruning stage**, we calculate the upper bound of the cluster score that each candidate pipeline can achieve. If this upper bound is lower than the current optimal value, the candidate is immediately eliminated. Theoretical proof shows that the upper bound of the retained candidate size is O(k log n), where k is the number of truly high-quality solutions and n is the size of the original search space. Finally, in the runtime phase, we introduce **dynamic tuning**, which monitors the contour coefficient slope and Davies–Bouldin index in real time. Once distribution drift is detected, key hyperparameters, such as k and ε, are automatically fine-tuned to restore robustness without requiring a restart of the search.

**The main contributions of this paper are as follows:**

(1) We systematically propose and validate the hypothesis that there is a structural correlation between the granularity of data cleaning and the effectiveness of clustering. Based on experiments with eight types of cleaners, six types of clusterers, and 60 public datasets, we reveal four layers of patterns, including monotonic segmentation, threshold inflection points, and interactions, providing a quantitative basis for the development of automated algorithms.

(2) We propose a confidence upper bound-driven search pruning algorithm, design a feature-enhanced multi-label predictor, and combine it with a confidence upper bound strategy to prune the search tree for unsupervised AutoML. Theoretically, the upper bound of the search space is reduced to O(k\*logn), and experimentally, we achieve 6–12 × acceleration and accuracy loss < 2% on a million-row dataset.

(3) We propose a runtime parameter drift monitoring and adaptive rollback strategy that automatically adjusts key hyperparameters, such as k, ε, and γ, when a decline in SSE slope or drift in DB index is detected. On 1 million lines of real and simulated data, we control the clustering stability fluctuation within ±1% while maintaining an overall acceleration of 8×.

The structure of this paper is as follows: Section 2 reviews recent advances in the cross-domain of data cleaning, clustering, and AutoML; Section 3 formally defines the data quality vector, objective function, and evaluation metrics; Section 4 details the proposed three-stage framework and provides complexity and theoretical analysis; Section 5 introduces the experimental setup, including the dataset, data corruption injection strategy, and complete experimental design; Section 6 presents experimental results along these three main lines and conducts in-depth discussions; Section 7 summarizes the findings and provides conclusions and future directions; the appendix provides additional figures and open-source code links.

# **2 Related Work**

This chapter reviews relevant research from three main perspectives: data cleaning and quality management, robustness of clustering algorithms in dirty data environments, and the latest developments in unsupervised AutoML. It summarizes existing limitations and lays the foundation for the problem definition and methods in the next chapter.

Data cleaning aims to detect and repair various defects such as missing values, outliers, duplicates, and format errors, and is a prerequisite for improving the usability of analysis. Early methods focused on statistical imputation (mean, mode) or rule-based anomaly detection. Subsequently, techniques such as probabilistic graph models, active learning, and neural networks significantly improved the ability to identify complex errors. For domain-specific data, customized cleaning frameworks combining knowledge graphs or external constraints have also emerged. However, unsupervised tasks with no labels lack “clean controls,” and excessive cleaning can easily delete a small number of critical anomalies, while conservative cleaning amplifies noise interference. Existing work has focused mainly on the “data quality” dimension. It has not directly linked cleaning decisions to downstream clustering performance, which is one of the motivations for this paper to develop a joint “cleaning-clustering-search” model further.

Among the three mainstream clustering algorithms, centroid-based K-Means and its improved methods are efficient on large-scale data but sensitive to outliers and geometric structures; density-based DBSCAN and OPTICS can identify clusters of arbitrary shapes and are more robust to low-density noise, but they are highly dependent on hyperparameters ε and minPts; bottom-up hierarchical clustering is suitable for capturing multi-scale structures but has high computational complexity. To address dirty data, researchers have proposed local repair strategies such as weighted distance and robust centroid updating, or introduced local density re-estimation into DBSCAN to mitigate misclassification. However, these methods mainly optimize either “cleaning” or “clustering” separately, lacking a holistic perspective. The impact of different cleaning operations on cluster shapes, parameter sensitivity, and the iterative paths within algorithms remains poorly quantified.

AutoML frameworks (Auto-sklearn, TPOT, etc.) have enabled the automatic search of “model–feature–hyperparameter” combinations in supervised tasks. Representative research in the unsupervised direction has primarily focused on the automatic recommendation of clustering algorithms and hyperparameters, but rarely includes data cleaning or utilizes PCA as a noise reduction method. Recent work has attempted to incorporate cleaning operators into the search space; however, most remain at the level of empirical rules and have not yet provided a unified, end-to-end, closed-loop modeling approach.

To address these shortcomings, this paper will utilize data and algorithm feature prediction to identify the optimal combination and reduce the search space, while quantitatively describing the impact of cleaning accuracy on the clustering process and results. This will lead to the proposal of a unified end-to-end optimization framework for “cleaning × clustering × hyperparameter tuning.”

**3 Problem Definition**