IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710


# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

## IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 7, July 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 7, July 2024

|DOI: 10.15680/IJIRSET.2024.1307028|

# Redundant Data Removal using Machine Learning

**Sinchana S**

PG Student, Department of MCA, PES Institute of Technology and Management, Shimoga, Karnataka, India

**ABSTRACT**: Redundant data creates significant challenges in data management, resulting in higher storage costs, decreased performance, and lower data quality. This study investigates the usage of machine learning techniques to remove redundant data, resulting in an automated, scalable, and accurate solution. We propose a comprehensive framework for identifying and removing duplicate records from various datasets using advanced algorithms such as clustering, classification, and anomaly detection. Implementing this framework on real-world datasets demonstrate its capacity to reduce data volume while maintaining data integrity and consistency. The outcomes of the experiment indicate significant enhancements inprocessing time and resource utilization when compared to traditional deletion methods. The work highlights the potential of Machine learning will improve data management practices and provides useful future research into automated data cleaning and optimization processes.

**KEYWORDS:** Data deduplication, machine learning, redundancy removal, data cleaning, automated data management.

## I.INTRODUCTION

Data duplication is a major issue in data management and analytics, reducing the effectiveness and precision of data processing systems. Duplicates can be caused by There are several variables, including incorrect data entry, the system migrations, and the combining of datasets from various sources. These duplicates have an impact on decision-making processes because they inflate storage requirements while producing false insights and analysis. Conventional deduplication techniques frequently use manual or rule-based processes, which can be time-consuming and error-prone, particularly if dealing with large datasets.In recent years, machine learning (ML) has shown to be an effective technology for addressing the challenges of data duplication. The algorithms for machine learning are capable of recognizing patterns and similarities in data, This makes them very useful for detecting duplicates That would ultimately be difficult to detect using traditional methods. Deduplication efforts can be enhanced through training these algorithms to detect minute changes and inconsistencies between different data entries, such as misspellings, formatting variants, or acronyms. This study will look into the Machine learning in redundant data removal, looking at different approaches and algorithms that have shown promise in this area. We will evaluate the efficiency of these strategies through extensive studies on a variety datasets, providing insights into their performance, scalability, and usability. The study's findings are expected to help create more efficient handling of data.systems, resulting in better data utilization and decision-making throughout multiple kinds of fields.

## II.RELATED WORK

Record deduplication is an emerging research area in databases and related fields like digital libraries. When data is gathered from multiple sources with different information description styles and metadata standards, inconsistencies can arise, leading to duplicate records. To resolve these inconsistencies, it is crucial to design a deduplication function that effectively combines the evidence available in data repositories to determine if two record entries refer to the same real-world entity.

In the context of bibliographic citations, Lawrence et al. [1], [2] have thoroughly explored this problem. They proposed various matching algorithms based on word matching, phrase matching, edit distance, and subfield extraction, with word and phrase matching producing better results. Initially, users had to manually review the entire dataset for record deduplication, a time-consuming and complex task due to the large number of records. However, later works have suggested various approaches to combine and utilize the evidence extracted from datasets as more strategies for evidence extraction have become available.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 7, July 2024

|DOI: 10.15680/IJIRSET.2024.1307028|

Elmagarmid et al. [3] classified these approaches into two main categories:
1. Ad-Hoc or Domain Knowledge Approaches: These primarily rely on domain knowledge and include techniques using declarative languages.
2. Training-based Approaches: These require supervised or semi-supervised training and include probabilistic and machine learning methods
Newcombe et al. [4] first proposed handling duplicates automatically by treating the record deduplication problem as a probabilistic issue. However, the challenge of combining the evidence persisted. Fellegi and Sunter proposed a method using two boundary values to determine whether two records are duplicates. This method was implemented in Febrl [5], which uses the following boundary values:
1. Positive Identification Boundary: Records are considered duplicates if the similarity value is above this boundary.
2. Negative Identification Boundary: Records are considered non-duplicates if the similarity value is below this boundary. Records with similarity values between the two boundaries are classified as possible matches.
This work is closely related to approaches that use machine learning techniques to achieve record-level similarity from field-level similarity [6], [7]. These approaches use a small portion of the available data for training, ensuring that the training and testing data have similar characteristics so that the results can be applied to new data. In Marlin [6], [7], the extracted evidence is used to train an SVM classifier for this purpose.

## III.SYSTEM MODEL AND ASSUMPTIONS

To remove data duplication using machine learning, a typical system model includes data preparation, feature extraction, and classification algorithms. To clean and standardize the data, pretreatment techniques such as normalization, tokenization, and initial deduplication are performed first. The raw data is then transformed into an analysis-ready format using feature extraction techniques like TF-IDF (Term Frequency-Inverse Document Frequency) for text data and hash functions for numerical data. These features are put into a classification algorithm, such as a Support Vector Machine (SVM), Random Forest, or a deep learning model like a Convolution Neural Network (CNN), which detects and eliminates duplicate items based on patterns and similarities.

The key premise of this approach is that duplicates have considerable similarities in their attributes, allowing the model to successfully distinguish between unique and copied items. This approach requires a large, labeled dataset for training, in which the model learns to identify duplicate records based on predefined criteria. It also implies that the preprocessing and feature extraction methods are robust enough to capture the data's basic features, ensuring accurate duplication identification even with minor alterations. The effectiveness of this system model is dependent on the quality of the input data, the suitability of the feature extraction techniques, and the accuracy of the chosen classification algorithm in detecting duplicate patterns.

### OVERVIEW OF DATA DUPLICATION

Data duplication removal, or deduplication, is an important part of data management that focuses on removing redundant data in order to increase storage efficiency and data quality. Using machine learning for deduplication considerably enhances the process by employing advanced algorithms to reliably and effectively identify and delete duplicates. Machine learning approaches, such as clustering algorithms, neural networks, and support vector machines, study data patterns and similarities to identify duplicates that traditional methods may miss. These models may be trained using labeled data, allowing them to improve over time and adapt to various data kinds and duplicate scenarios. Furthermore, machine learning algorithms can handle big and complicated datasets, providing scalable answers to modern data management challenges. Streamlining the deduplication process provides cleaner datasets, which leads to more reliable data analysis, lower storage costs, and improved overall data integrity.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 7, July 2024

|DOI: 10.15680/IJIRSET.2024.1307028|

## IV. RESULTS

This section summarizes the findings of all studies detailed in the experimental section.

| Learning ζ \ Test ζ | 0 | 1 | 2 |
|---|---|---|---|
| 0 | **1** | 0.764 | 0.649 |
| 1 | 0.999 | 0.994 | 0.941 |
| 2 | 0.999 | **0.998** | 0.984 |
| 3 | 0.997 | 0.997 | **0.989** |
| random(0,4) | **0.998** | 0,998 | 0,9854 |

Table 1 : Accuracy for various ζ values

| Learning ζ \ Test ζ | 3 | mean |
|---|---|---|
| 0 | 0.578 | 0.748 |
| 1 | 0.876 | 0.952 |
| 2 | 0.951 | 0.983 |
| 3 | **0.971** | **0.988** |
| random(0,4) | 0.954 | 0.984 |

Table 2 : Accuracy for various ζ values

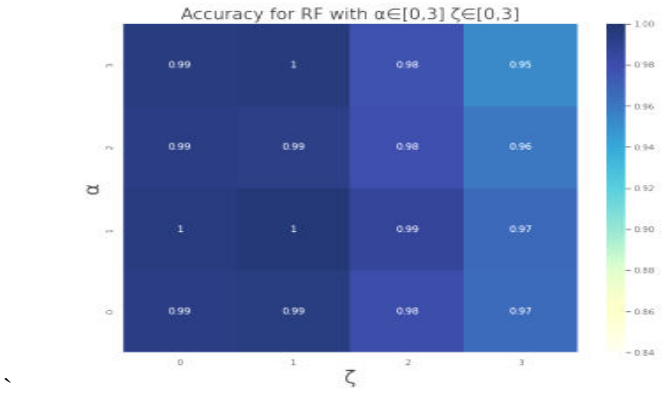


Fig .1: : RF classifier accuracy with α ∈ [0, 3] and ζ ∈ [0, 3]

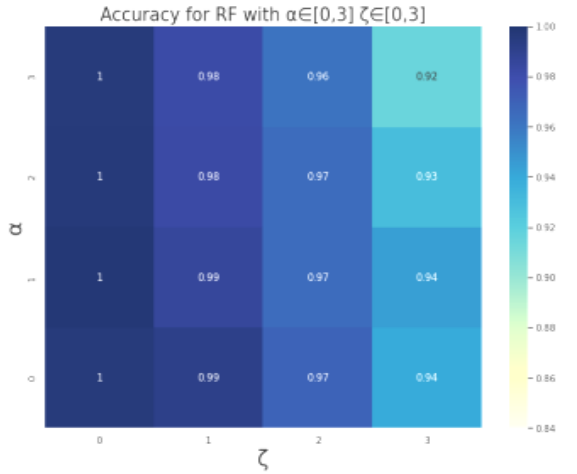


Fig .2 : RF classifier accuracy with α ∈ [0, 3] and ζ ∈ [0, 3]

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 7, July 2024

|DOI: 10.15680/IJIRSET.2024.1307028|

## V.CONCLUSION

This study will look into the use of machine learning in redundant data removal, looking at different approaches and algorithms that have shown promise in this area. We will evaluate the efficiency of these strategies through extensive studies on a variety datasets, providing insights into their performance, scalability, and accessibility. The study's findings are expected to help Create more efficient data management. Systems, resulting in better data utilization and making choices throughout A wide range of fields.

## REFERENCES

1. S. Lawrence, C.L. Giles, and K.D. Bollacker, “Autonomous Citation Matching,” Proc .Third Int’lConf. Autonomous Agents, pp. 392-393, 1999.
2. S. Lawrence, L.Giles and K. Bollacker,“Digital Libraries and Autonomous Citation Indexing, ” Computer, vol. 32, no. 6, pp. 67- 71, June 1999.
3. A.K. Elmagarmid, P.G. Ipeirotis, and V.S. Verykios, “Duplicate Record Detection: A Survey,” IEEE Trans. Knowledge and Data Eng., vol. 19, no. 1, pp. 1-16, Jan. 2007.
4. H. B. Newcombe, J. M. Kennedy, S . Axford, and A. James, “Automatic Linkage of Vital Records,” Science, vol. 130, no. 3381, pp. 954- 959, Oct.1959.
5. “Freely Extensible Biomedical Record Linkage,” http://sourceforge.net/projects/febrl, 2011.
6. M. Bilenko, R. Mooney, W. Cohen, P. Ravikumar, and S. Fienberg, “Adaptive Name Matching in Information Integration,”IEEE Intelligent System, vol. 18, no. 5, pp. 16-23, Sept./Oct. 2003.
7. M. Bilenko and R.J. Mooney, “Adaptive Duplicate Detection Using Learnable String Similarity Measures,” Proc. Ninth ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining, pp. 39- 48, 2003.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

## IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details