QIDLearningLib: A Library for Quasi-Identifier Detection and Evaluation

Sancho Amaral Simões
University of Coimbra, Portugal
uc2019217590@student.uc.pt, sanchosimoes@student.dei.uc.pt

Editor: Editor Name

Abstract

QIDLearningLib, developed as part of the Master's in Informatics Engineering - Intelligent Systems program at the University of Coimbra, particularly in the thesis titled "Automated Data Privacy Protection using Deep Learning and Causality Techniques", is a specialized library designed primarily to provide metrics to automate the identification and evaluation of quasi-identifiers. These quasi-identifiers, such as demographic information, play a critical role in potential privacy breaches and linkage attacks.

Utilizing a combination of machine learning, statistical, and causality techniques, QIDLearningLib offers metrics for assessing data privacy, utility, and the performance of quasi-identification recognition algorithms. Despite being in its initial version, the library stands as a significant contribution to the open-source community, specifically addressing the needs of researchers focused on data privacy and quasi-identifier analysis.

Keywords: Quasi-identifiers, Data Privacy, Machine Learning, Open Source, Library, Evaluation Metrics.

1 Introduction

In the contemporary landscape of ever-growing digital interactions, safeguarding individual privacy has become an overarching concern. The proliferation of data, often containing sensitive information, necessitates robust measures to prevent unauthorized access and the risk of re-identification. A pivotal aspect of this challenge lies in the thorough examination and understanding of quasi-identifiers—attributes that, when combined, can uniquely identify an individual. Addressing this complex issue, QIDLearningLib stands as a sophisticated library designed to revolutionize the automated identification and evaluation of quasi-identifiers within datasets.

2 Motivation

The motivation driving the development of QIDLearningLib is deeply rooted in the evolving dynamics of data privacy. As datasets continue to grow in complexity and volume, traditional methods of safeguarding sensitive information prove increasingly inadequate. Quasi-identifiers, encompassing facets such as demographics, pose a subtle yet potent threat by serving as potential conduits for the re-identification of individuals. This library is motivated by the imperative to offer a comprehensive solution that goes beyond conventional approaches.

QIDLearningLib is not merely a response to an existing gap; it is a proactive initiative to equip researchers, developers, and educators with a powerful toolset that can redefine the landscape of data privacy. By seamlessly integrating advanced machine learning techniques and causality analysis, the library seeks to empower users with the means to automate the detection of quasi-identifiers and, consequently, fortify the protective barriers around sensitive information.

3 Key Objectives

QIDLearningLib sets forth several key objectives to address the multifaceted challenges in the realm of data privacy:

- **Automation:** QIDLearningLib aims to provide a ground basis for the automated identification of quasi-identifiers within diverse datasets. By leveraging statistical, causal, and machine techniques algorithms, the library endeavors to streamline and enhance this crucial process.
- **Evaluation Empowerment:** The library is driven by the commitment to equip users with a comprehensive set of evaluation metrics. These metrics not only gauge the effectiveness of quasi-identifier recognition algorithms but also contribute to a deeper understanding of the privacy landscape within a dataset.
- **Holistic Data Privacy:** QIDLearningLib aspires to offer tools for evaluating the overall level of data privacy in a given dataset. By adopting a holistic approach, the library addresses the intricate interplay of quasi-identifiers and their potential impact on privacy.
- **Educational Utility:** Beyond its practical applications, QIDLearningLib is envisioned as an educational resource. The library can serve as a valuable tool in academic settings, facilitating the exploration of machine learning and data privacy concepts.

In essence, the motivation propelling QIDLearningLib transcends the immediate challenges of data privacy—it seeks to establish a paradigm shift in how quasi-identifiers are identified, evaluated, and integrated into broader machine learning architectures.

4 Features

QIDLearningLib encompasses a diverse set of features designed to facilitate robust quasi-identifier analysis:

- **Causality Metrics:** QIDLearningLib integrates causality metrics tailored to quasi-identifiers, providing insights into the causal relationships associated with sensitive attributes.
- **Data Utility Metrics:** The library incorporates metrics specifically crafted to assess the utility of data concerning quasi-identifiers, ensuring a nuanced understanding of information preservation.

- **Data Privacy Metrics:** QIDLearningLib offers tools for evaluating data privacy concerning quasi-identifiers, enabling users to gauge and mitigate potential privacy risks.
- **Performance Metrics:** Evaluate the effectiveness of the QID recognition system by comparing results with ground truth through comprehensive performance metrics.
- **Quasi Identifier-Specific Metrics:** QIDLearningLib introduces metrics focused on quasi-identifiers, with an emphasis on distinction and separation to enhance the granularity of analysis.
- **Metric Encapsulation:** The library provides an object-oriented approach, encapsulating causality, data utility, data privacy, and performance metrics. This allows users to inspect relevant statistics and gain a comprehensive understanding of the data.
- **Test Scripts:** QIDLearningLib includes test scripts that facilitate the evaluation of the aforementioned metrics. Users can apply these scripts to imported datasets or generate simple dummy datasets using library primitives for thorough testing and validation.

5 Use Cases

The primary use case for QIDLearningLib revolves around its integration as context information in a machine learning architecture dedicated to the classification of quasi-identifiers within a given dataset, which is the goal of the research related to the development of the mentioned library.

Additionally, researchers and practitioners in the field of data privacy can benefit from QIDLearningLib in the following ways:

- **Research:** QIDLearningLib serves as a valuable resource for researchers conducting studies on data privacy and quasi-identifier analysis.
- **Development:** The library provides a solid foundation for developers working on projects related to data privacy and security.
- **Education:** QIDLearningLib can be used as an educational tool in courses covering machine learning and data privacy.

6 Future Work

The continuous development of QIDLearningLib is dedicated to further augmenting its functionality and addressing emerging challenges in data privacy. The forthcoming updates will include:

6.1 New Metric Development

Exploration into the development of novel metrics will be a primary focus, specifically within the existing context of causality, performance, and privacy evaluation. These metrics aim to

provide a more nuanced and comprehensive understanding of quasi-identifiers, reinforcing the library's analytical capabilities.

6.2 Primitives for Metric Combination

To enhance user flexibility, QIDLearningLib will introduce primitives that facilitate the seamless combination of metrics. This feature will enable users to create customized evaluation criteria tailored to their specific requirements and dataset particularities.

6.3 Graphication Method Advancements

Continuous refinement of graphication methods will be pursued to improve the visualization of metrics. This includes optimizing existing graphing techniques and introducing new methods to enhance the interpretability of metrics.

6.4 Graphical Interface Development

The development of a user-friendly graphical interface is planned to simplify the process of testing various metrics on imported datasets. This interface will provide an intuitive platform for users to interact with QIDLearningLib, streamlining the evaluation and analysis of quasi-identifiers.

6.5 Metric Selection Mechanism

QIDLearningLib will incorporate mechanisms to assist users in selecting the most suitable metrics for their specific datasets and validation methods. This involves implementing advanced techniques such as Bayesian optimization or genetic algorithms, and optimizing metric combinations for better results.

6.6 Integration with ML Libraries

Efforts will be made to ensure seamless integration with existing machine learning libraries, such as scikit-learn for pipeline compatibility and Keras for custom neural network layers. This integration aims to enhance the versatility of QIDLearningLib within broader machine learning workflows.

6.7 Explainability Enhancements

One of the key areas of improvement will be the integration of explainability features within QIDLearningLib. This entails developing mechanisms that provide insights into the decision-making process of quasi-identifier recognition algorithms. Users will benefit from a clearer understanding of how and why certain quasi-identifiers are identified, promoting transparency and trust in the evaluation process.

6.8 Scalability Optimization

Scalability is a crucial consideration, especially when dealing with large datasets. Future work will focus on optimizing QIDLearningLib for scalability, ensuring efficient perfor-

mance even with extensive datasets. This improvement aims to broaden the applicability of the library across diverse use cases, accommodating varying data sizes without compromising computational efficiency.

6.9 Dynamic Metric Adaptation

To enhance adaptability to different data scenarios, QIDLearningLib will explore the concept of dynamic metric adaptation. The library will intelligently adjust the selection and weighting of metrics based on the characteristics of the dataset, providing more tailored and effective evaluations. This adaptive approach ensures that the library remains robust and applicable across a spectrum of data types and structures.

6.10 Continuous Integration

The implementation of continuous integration practices will be integrated into the development workflow of QIDLearningLib. This ensures that changes and updates to the library are systematically tested and validated, maintaining the integrity of the codebase.

6.11 Unit Testing

A comprehensive unit testing framework will be established for QIDLearningLib. This will involve creating test suites to validate individual units of code, ensuring that each component functions as intended. Unit testing is crucial for identifying and resolving bugs or issues early in the development process.

7 Conclusion

QIDLearningLib, with its ongoing development roadmap, stands as a promising initiative in advancing the capabilities of data privacy tools. By automating quasi-identifier detection and evaluation, the library contributes significantly to the overarching mission of securing sensitive information in our interconnected world.

Acknowledgments and Disclosure of Funding

The author acknowledges the support and guidance of Professor Pedro Manuel Henriques da Cunha Abreu and Professor João Paulo da Silva Machado Garcia Vilela.

Appendix: Installation Guide

For detailed instructions on installing and using QIDLearningLib, please refer to the documentation available on the official GitHub repository: `https://github.com/smartlord7/QIDLearningLib.git`.

To install the software, the repository can be cloned or downloaded as a ZIP file. The Python package of the library is located at `QIDLearningLib/src/QIDLearningLib`, and the documentation is available at `QIDLearningLib/src/QIDLearningLib/doc/QIDLearningLib/`. It is highly recommended to consult the documentation before using the library.

Once the repository is obtained, the user can move the package to their project directory and start using it like any other Python package.

Example Commands:

```
git clone https://github.com/smartlord7/QIDLearningLib.git
cd QIDLearningLib/src/QIDLearningLib
```

After obtaining the library, the user can integrate it into their Python project seamlessly, leveraging the functionalities provided by QIDLearningLib.