IJARIIT INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 8, Issue 3 - V8I3-1399)
Available online at: https://www.ijariit.com

Sensitivity Analysis using k-anonymization

Krishna Yanmantram
krishnayanmantram@gmail.com
Vellore Institute of Technology, Vellore, Tamil Nadu

T. M. Vishnu Mukundan
tm.vishnu.m@gmail.com
Vellore Institute of Technology, Vellore, Tamil Nadu

Jama Surya Teja
jama.surya2019@vitstudent.ac.in
Vellore Institute of Technology, Vellore, Tamil Nadu

ABSTRACT

It's necessary to keep sensitivedata and private data safe. When financial information, healthcare information, and other sensitive consumer or user data are mishandled,they can be destructive. Due to a lack of access control over personal information, individuals may be susceptible to fraud and identity theft. This paper provides an overview of the ideas of data privacy, re- identification risk, and dataset utility, as well as the correlations between the three. On the adult dataset from the UCI Machine Learning Repository, this study presents a sensitivity analysis of the k-anonymization algorithm. ARX, anopen-source anonymization tool, was used to show this.

Keywords*—Data Anonymization, ARX, K- Anonymity, Re-Identification Risk*

1. INTRODUCTION

Sensitivity evaluation is an economic version that determines how goal variables are affected primarily based totally on modifications indifferent variables referred to as enter variables.

This version is likewise known as what-if or simulation evaluation. It is a manner to expect thefinal results of a choice given a positive variety ofvariables.

k-Anonymity is a technique for offering privateness safety via means of making sure that facts can't be traced to an individual. In a k- nameless dataset, any figuring out data takes place in at least k tuples. To acquire ultimate and realistic k-anonymity, recently, many one-of-a- kind types of algorithms with diverse assumptions and regulations were proposed with one-of-a-kind metrics to determine quality.

Hence by using ARX we will be performing the following operations and then go about comparing input and output data, perform analysis on the empirical data observed and check the efficiency as per classification and then check for incidental expense which will lead us to best security levels.

2. RELATED WORK

Researchers and data mining firms can use research data to develop new and improved methods for identifying trends and patterns. Many studies have been done on how to share data while maintaining privacy. Utility and security are difficult to define clearly because they are very dependent on the application that will be used with the anonymized data. A datasetis said to be k-anonymous if each record cannot be distinguished from at least k-1 other records [5]. Most anonymization methods use K-anonymity as their foundation.

[2] studies various techniques of anonymization from a variety of perspectives, including generalisation, micro aggregation, and data characteristics. [4] discusses four different techniques to achieve k-anonymity for data anonymization - generalisation and suppression, clustering methods, graph methods and set representation methods. [3] aims to categorize the sensitive attributes to high and low sensitivevalues. [1] demonstrates a k-anonymization sensitivity analysis using generalisation and suppression, altering the value of k. The work given in [6] analyses and concludes secure and useful data anonymization and re-identification strategies through competitions. A re-identification is a process that attempts to identify a record subject from an anonymized record basedon some features of the original record [6].

Each potential adversary will be judged on their purpose, resources, and purpose to harm, as well as their prior knowledge of the target and knowledge of the target's participation in the trial classifies the adversaries into three primarymodels based on their background knowledge and knowledge about the target's participation - prosecutor, journalist and marketer. The work in shows that two de-identification strategies,k- anonymization and adding a 'fuzzy factor,' effectively lowered the chance of re-identification of patients in a dataset of 5 million patient records [7].

3. PROPOSED WORK

By using ARX we will be performing the following operations which include analysing the data quality, comparing input and output data, studying empirical data, checking performance grouped by classification and later checking for contingency which thus results in the end of our analysis of sensitivity of the given data set.

The chosen dataset by us is the adult dataset which is based on an individual's annual income which is determined by multiple reasons such as education, age, gender, occupation etc. Thus, we will try our best to reduce the sensitivity of the data after performing anonymization and visualizing thesame.

The k-value was taken as input and the relative risk of re-identification and the absolute differencebetween descriptive statistics acquired from the original data set and the k-anonymized data sets are the outputs in the sensitivity analysis. The value of k was varied from 2 to 5. Age, work class,final weight income, education, occupation, relationship, gender, capital-gain, capital-loss, hours-per-week, native-country and income were classified as quasi-identifiers. Educational number,marital status and race were classified as insensitive attributes.

4. RESULT/DISCUSSIONS

According to the table below, theattributes were transformed.

Variable	Type of hierarchy
Age	Interval
Work class	Ordering
Fnlwgt (final weightincome)	Interval
Education	Ordering
Education number	Interval
Marital	Ordering
Occupation	Ordering
Relationship	Ordering
Race	Generalised to race as the data was less for some races
Gender	Generalised to person
Capital-gain	Identifier
Capital-loss	Identifier
Hours-per-week	Interval
Native country	Ordering
Income	Ordering

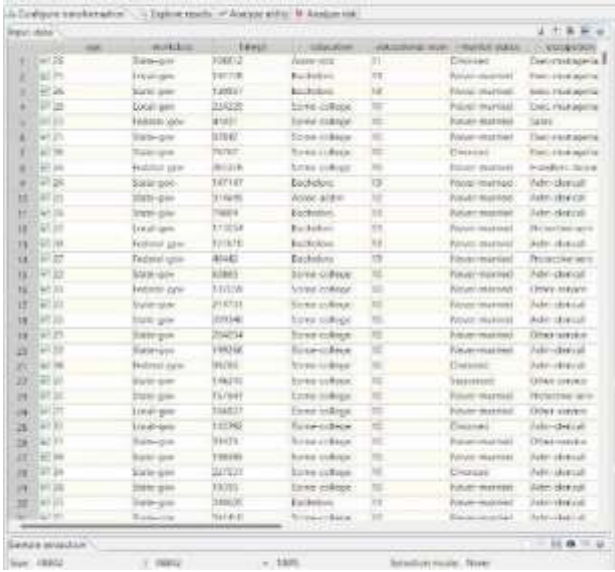


Figure 1 - Configuring data

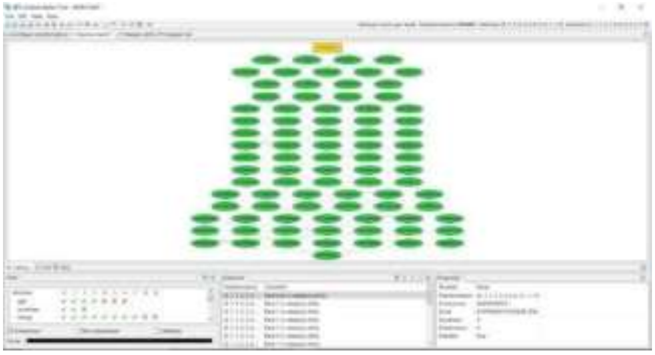


Figure 2 – Domain Generalization Hierarchy

This is the graphical representation of the subset of solution space also known as the domaingeneralization hierarchy. Here we consider all the records with different generalization levels and form combinations, picking the one that satisfies k anonymity.

Here we have set our k value to 5. Each node here represents one transformation. These nodes are labelled with the generalization levels which are applied to the quasi-identifiers. The node in yellow represents the global optimum with respectto utility. In this project, we have left up to three generalization levels for age, one for work-class etc. even if we tend to increase or decrease it, the global optimum will always remain the same.

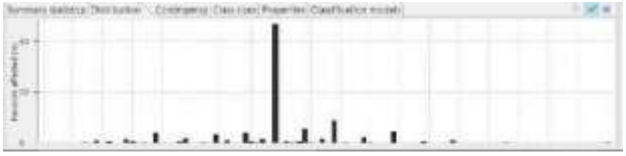


Figure 3 – Distribution of age before anonymization

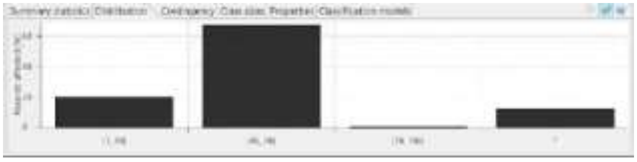


Figure 4 – Distribution of age after anonymization

The purpose of the utility view is to allow users to compare transformed data representationto the original data and to assess their suitability.Left side we have the original data and on the right side we have the transformed data.

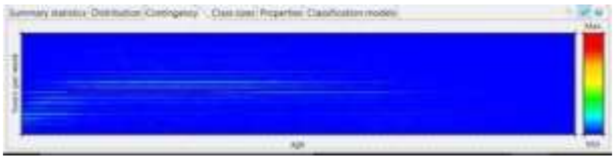


Figure 5 – Contingency of age before anonymization

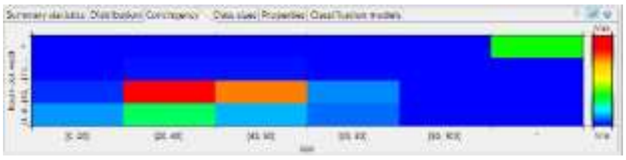


Figure 6 – Contingency of age after anonymization

Statistical information includes frequency distributions and contingency tables. Each information is displayed as raw tabular data andas graphs. After choosing the age attribute, herewe see the frequency distribution of the attribute 'age' and how the distribution is changed on the right-hand side by applying generalization.

Metadata about the input dataset includes basic properties like tuples and an overview of the attribute classifications. Metadata about the transformed dataset includes no. of outliers that have been removed from the dataset and the no. of equivalence classes as well as minimum and maximum class size.

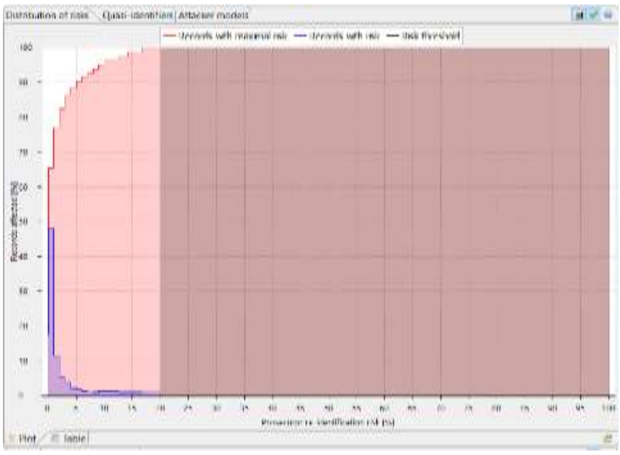


Figure 7 – Histogram of distribution of re-identification risk

The distribution of re-identification risks among the dataset's records can be viewed in theform histogram and table. The 20 percent threshold can be observed from the above figure.

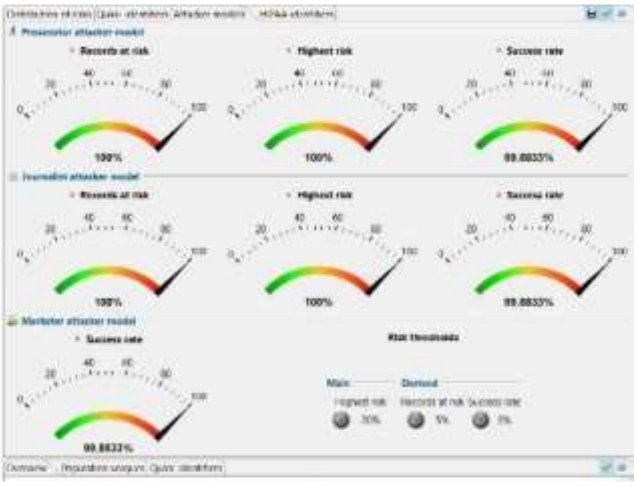


Figure 8 – Re-identification risk before anonymization

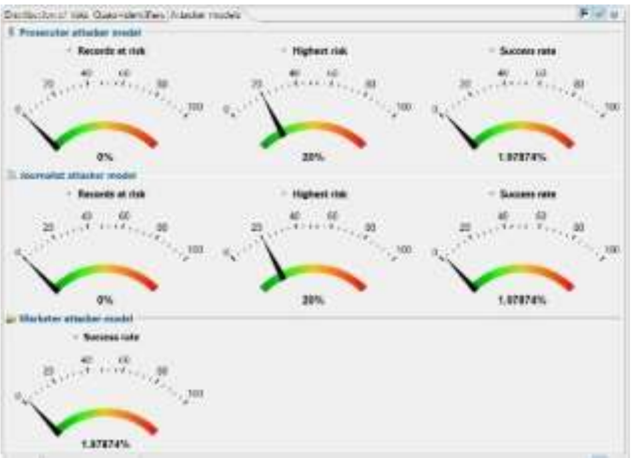
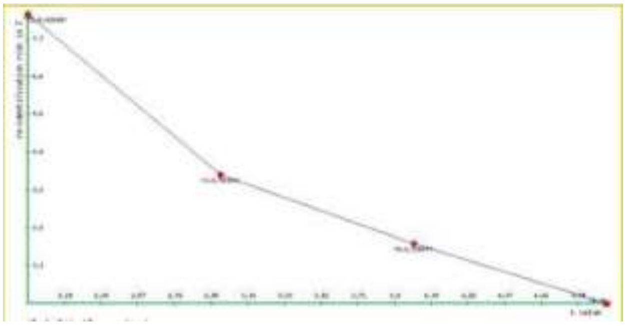


Figure 9 – Re-identification risk after anonymization

The prosecutor model's attacker targets a specific person, and that the attacker is aware thatthat person's data is in the dataset. The journalist model's attacker targets a specific person, though it's doubtful that they were aware of the user's participation in the dataset beforehand. The attacker in the marketing model doesn't target a single person; instead, they try to re-identify a huge group of individuals. Hence, an attack is only successful if a higher proportion of the data can be re-identified. Protecting the data against prosecutor or journalist risk ensures that the data is protected from marketer risk [7]. The followingtable shows the average prosecutor risk of re- identification. The average risk is displayed in thek=1 column before k-anonymity is applied.

The graph below illustrates how prosecutor riskchanges as the k value increases. X-axis has k- values, Y-axis has the re-identification risk in percentage.



5. CONCLUSION

The adult data set was set through the above parameters and the following parameters which include analysing the data quality, comparing input and output data, studying empirical data, checking performance grouped by classification and later checking for contingency were studied and the same was also visually observed. This has now made the data set have a very negligible security breach post anonymizing.

6. REFERENCES

[1] W. Santos, G. Sousa, P. Prata and M. E. Ferrão, "Data Anonymization: K-anonymity Sensitivity Analysis," 2020 15th Iberian Conference on Information Systems and Technologies (CISTI), 2020, pp. 1-6, doi: 10.23919/CISTI49556.2020.9141044.

[2] X. Ren and J. Yang, "Research on Privacy Protection Based on K-Anonymity," 2010 International Conference on Biomedical Engineering and Computer Science, 2010, pp. 1-5, doi: 10.1109/ICBECS.2010.5462427.

[3] Widodo, E. K. Budiardjo, W. C. Wibowo and H. T. Y. Achsan, "An Approach for Distributing Sensitive Values in k-Anonymity," 2019 International Workshop on Big Data and Information Security (IWBIS), 2019, pp. 109-114, doi: 10.1109/IWBIS.2019.8935849.

[4] V. Sharma, "Methods for privacy protection using k-anonymity," 2014 International Conference on Reliability Optimization and Information Technology (ICROIT), 2014, pp. 149-152, doi: 10.1109/ICROIT.2014.6798301.

[5] L. Sweeney. “k-anonymity: a model for protecting privacy,” International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, Vol. 10 (5), pp 557-570, 2002.

[6] H. Kikuchi, T. Yamaguchi, K. Hamada, Y. Yamaoka, H. Oguri and J. Sakuma, "Ice and Fire: Quantifying the Risk of Re-identification and Utility in Data Anonymization," 2016 IEEE 30th International Conference on Advanced Information Networking and Applications (AINA), 2016, pp. 1035-1042, doi: 10.1109/AINA.2016.151.

[7] L. Kniola, “Plausible Adversaries in Re- Identification Risk Assessment” PhUSE Annual Conference, 2017.

[8] Ursin G, Sen S, Mottu JM, Nygård M. Protecting Privacy in Large Datasets-First We Assess the Risk; Then We Fuzzy the Data. Cancer Epidemiol Biomarkers Prev. 2017 Aug 1;26(8):1219-1224. doi: 10.1158/1055-9965.EPI-17-0172.

Epub 2017 Jul 28. PMID: 28754793.

[9] El Emam K, Dankar FK. Protecting privacy using k-anonymity. J Am Med Inform Assoc. 2008 Sep-Oct;15(5):627-37. doi: 10.1197/jamia.M2716. Epub 2008 Jun 25. PMID: 18579830; PMCID: PMC25280