IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Big Personal Data Anonymization Using Deep Learning Techniques

Sharing personal identifiable data while maintaining privacy and anonymization

[1]Jacob Mziya, [2]Dr. Glorindal Selvam

[1]MSC Student, [2]HOD & PG-Coordinator

[1]Masters of Computer Science,

[1]DMI – St. Eugene University, Lusaka, Zambia

Abstract: Sharing of big personal data is usually possible and allowed in various data sharing agreements between organizations/ institutions because the data is anonymized, without containing any identifiable data. Anonymization is done to enforce privacy and to maintain confidentiality. Anonymized data however is much efficient for aggregate statistics and analytics, what if we want to be able to analyze the data at an individual level but maintain the privacy of the individual? In this paper we aim to achieve sharing of both anonymized and identifiable data while maintaining privacy. We do this using the K-Anonymity algorithm and a custom mapping of the anonymized data to the identifiable data. We consider extracting identifiable data into a separate table out of the compatible destination database. Extracting all the quasi-identifiable data and combine them. Then store the concatenated quasi-identifiers in the same separate table as the identifiers.

Index Terms: Big Personal Data, Data Privacy, Privacy Preserving, Data Anonymization, K-Anonymity, Differential Privacy

I. INTRODUCTION

Data Privacy assists a person to opt-in to sharing their sensitive data and be in position to choose what type of data about them gets shared. However, even if a person allows their specific identifiable data to be shared with the digital platform, they are providing sharing rights to, some policies will still not allow for this data to be shared between the platform collecting the data and other parties who they have a sharing agreement with. Sectors that are popularly known for being strict with data sharing policies include the health, and finance sectors.

For years, big data has been collected from various platforms and sources. Little of this data has been able to be shared with personal identifiable data inclusive in the shared data. This is mainly because data is mostly shared for analysis. Identifiable data is necessary to make effective decisions but does not really matter for aggregate statistics. Data is currently being shared for aggregate statistics in most cases.

Sharing of data for marketing purposes, recommendations, and follow-ups in health-related activities, however, requires personal data to be made available. Let us reference a case in the health sector. An example case in Africa, where HIV is spreading, and innovations are being created to minimize the spread. These innovations are proving to be effective, but a challenge remains that for new HIV cases, there is a need to track the person who has been identified with a new infection. Tracking that person would mean being able to trace back to other people they have been in contact with to be able to find the possible person who was infected in the first place.

Data is always collected with identifiable data, only at time of shared is this data de-identified. In most cases using the k-anonymity algorithm as this study develops on top of that. The data is identified by sharing to maintain data privacy. This theoretic study combines the k-anonymity algorithm and various quasi-identifiers to be able to share identifiable data without it necessarily allowing identification of the person but providing a way of making even better decisions and follow-ups were needed.

Sharing of personal identifiable data is by law not allowed, and if any party prefers to share such data while using the below theory to achieve anonymization and privacy, at least this must be done if there is a clear legal basis to do so or if the data subjected to sharing has given their clear consent. If, however, there is a valid reason for sharing personal identifiable data that can be justified, then it is also allowed to do so. An example remains in a case for HIV activities in African regions where the spread of HIV is high. One party can collect person identifiable data for an HIV patient, and another can be shared this data to allow them to conduct a contact tracing exercise to determine how many other people might have been potentially infected by the subject patient. This data however, using the proposed methods, will not be able to be traced back to the patient themselves.

Abbreviations and Acronyms

Abbreviation /Acronym	Definition /Description
HIV	Human Immunodeficiency Virus
HTS	HIV Testing Services
STI	Sexually Transmitted Infection

II. ADOPTED RELATED WORK

2.1 On K-Anonymity

[6] K-Anonymization Approach for Privacy Preserving IN Data Mining by Vimalkumar B. Vaghela
Focuses on both aspect privacy preserving as well as information loss because information loss is main challenge with k-anonymization, so, for that, 2-level anonymization approach introduced for in case of k=8 and 4. K=4 has less information loss as compared to k=8 and higher data utility but by comparing privacy parameter of both this case, k=8 has higher privacy than k=4 by this 2-level anonymization approach which gives benefits of both first privacy parameter from k=8 and second less information loss form k=4.

2.2 On Quasi-Identifiers

[3] Learning quasi-identifiers for privacy-preserving exchanges by Sol, C., et.al
The challenging and pervasive issue associated with information exchange is inferential disclosure. It occurs in the following three situations.

- The exchanged data correlate with publicly available information
- The exchanged data comprise patterns like those in a sharing partner's data
- The shared data's attributes are interdependent.

III. RESEARCH METHODOLOGY, APPROACH, AND FORMULAS

The methodology of the study was based on a few conversations with relative key stakeholders in the field of health sciences and health informatics to understand the challenges of not having access to identified data.

3.1 Interviews (One-on-One conversations)

Conversations with key stakeholders in the health sciences and health informatics were beneficial to the scope of identifying why identified data needs to be shared and possibly how it can be shared.

The conversations were conducted on a one-on-one basis with a sample size of 10. Out of these, 3 were medical practitioners; an HTS coordinator, an HIV coordinator, and an STI coordinator. 7 of these were individuals involved with project management and decision makers for HIV prevention activities.

3.2 Data Mining

As this study is towards big personal data, data mining is essential. Related data from various sources is to be mined, collected, transformed to match the destination dataset for analysis, and then stored. On this data is where all the deep learning techniques are going to be used to combine the anonymized data and include the identified data to enable data sharing.

The algorithm created is tested against the data that has anonymized combined with identified data, to identify the rate of data that can be easily identified and exposing the person's privacy.

3.3 K-Anonymization Approach

This approach is the most preferred for data privacy and anonymization. It provides better time complexity compared to most approaches, but it also poses a higher risk of data loss [6]. The approach in this study is adopted as it is to maintain minimal rate of information loss. The only modification done to the approach is to enable the capability of sharing identifiable data.

3.4 Theoretical Framework

This section elaborates the combination of quasi-identifiers to achieve anonymity on the personal identifiable data. Combine quasi-identifiers with k-anonymization and you will achieve data sharing for big personal data while preserving data-privacy. In a separate data table, you will have identifiable data linked to the quasi-identifiers to achieve this form of data sharing.

The pseudocode below states how we can achieve this.

Step 1: Collect data from multiple sources
Step 2: Transform the data as it is into a compatible dataset for our platform
Step 3: Create a separate table to hold all identifiable data
Step 4: Map the identifiable data table with quasi-identifiers combined attributes

We adopt the k-anonymization algorithm *[6]* and modify it to combine quasi-identifiers and map them to the identifiers, while maintaining the sensitive datum in its entirety.

Algorithm:

Input: Dataset D which has r tuples
Output: y = {σ1, σ2, σ3, …, σp}
// D is the compatible database
// r is the number of tuples in the dataset
// y is a subdivision of r tuples
// σ is a cluster
// k is number of tuples on anonymization

Begin

1. Recognize the attributes. Identifiers *[7]* (I), quasi-identifiers *[7]* (QI), sensitive attributes *[7]* (SA), non-sensitive attributes *[7]* (NSA)
2. Extract quasi-identifiers and concatenate them to form one attribute, a concatenated quasi-identifier (CQI)
3. Extract the I and store it in a separate tuple together with the CQI
4. Sort all tuples by their QI
5. Map the sorted tuples to a relationship in the table containing I and CQI through/ on CQI
6. Recognize the amount of equivalence classes and groups
7. Create an equal/unequal grouping of QI and SA to generate the sub-database
8. Make a divider of all tuples into k tuples group
9. Select a tuple ri arbitrary from the initial block of k tuples
10. Likewise select next tuples rj from the block of k tuples
11. Do generalization and suppression
12. Do further clustering only that cluster which has more information loss K=k/2
13. Do generalization and suppression
14. Compute info loss

$$IL(y) = \Sigma_{i=1}^{\rho} IL(\sigma i)$$

15. Transfer the tuples in a group with lowest information loss
16. Search additional element in a group that surpass the k size
17. Add further element in a group whose info loss is lowest

End

The algorithm steps are taken from the k-anonymization approach for data privacy in data mining, a customization of the equal/unequal group of QA and SA. Steps 2, 3, and 5 are customized by this study to allow identified data not to be removed from the k-anonymity, altering the algorithm to achieve big personal data sharing with identified data. These 3 steps extract the identifiable data instead of eliminating them and store them in a separate table or a sub-database alongside a combination of the quasi-identifiers to allow them to be referenced later by the anonymized data. By doing so, k-anonymization achieves data anonymization that is needed to preserve data privacy and the custom steps achieve what is needed to share big personal identifiable data. The algorithm combination achieves this study's goal of achieving big personal data anonymization using deep learning techniques to share personal identifiable data while maintaining privacy and anonymization. Steps 12 and 13, were added to the equal/ unequal group of QA and SA to minimize information loss.

We recognize and identify the attributes as Identifiers, Quasi-Identifiers, Sensitive Attributes, Non-Sensitive Attributes (step 1). Any dataset containing information about individual subjects will contain records (rows) with attributes (columns) that fall into at least one of those categories *[7]*. Next, instead of eliminating the identifiers, we extract quasi-identifiers, combine them into one attribute, a concatenated quasi-identifiers attribute, then store it in a separate database with the identifiers (step 2 and 3). Then we find out the number of groups and clusters such that $\sigma=\frac{r}{k}$, where r is the number of tuples in a database and k is the anonymization factor (step 6). Using systematic clustering algorithm in demand to generate the clusters *[8]*, we select a tuple from the first cluster *[6]* (step 8, 9, 10). Conduct a generalization and suppression that does further clustering for only the clusters that have information loss. Repeat the generalization and suppression on all clusters having information loss (step 11, 12, 13, 14). Throughout the clustering process if a cluster exceeds the k-size, extra records should be added in the cluster whose information loss is lowest (step 15, 16, 17).

Quasi-identifiers and identifiers combination can be used to de-identify the data for identifiable data to be able to re-identify the data without infringing data privacy and breaking anonymization. A more alike granulation is used in a strategy similar to that used in k-anonymity to de-identify private information systems *[3]*.

IV. FINDINGS AND ENHANCEMENTS

4.1 Results of Information Loss through k-anonymity in the algorithm

Vimalkumar run the proposed approach on various k values up to 100 [6]. We replicated the experimental on a sample case of up to 50. Information loss with the proposed approach is minimal. This gives us the confidence to incorporate this method in the inclusion of the identifiable data.

Figure 4.1: Information Loss against k

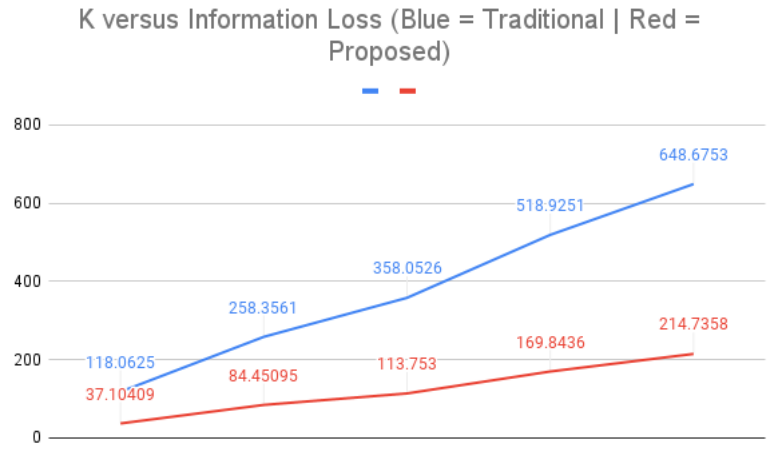

Table 4.1: Information Loss through k-anonymity *[6]*

K	Traditional approach Information loss	Proposed approach Information loss
10	**118.0625**	**37.10409**
20	**258.3561**	**84.45095**
30	**358.0526**	**113.753**
40	**518.9251**	**169.8436**
50	**648.6753**	**214.7358**

Table 4.1 shows a comparison of the information loss through k-anonymity in the algorithm between the traditional approach and the proposed approach. The results states are as referenced from equal/ unequal group of QA and SA with customization for generalization and suppression *[6]* but not with the inclusion quasi-identifiers and identifiers as the goal of the study was not to alter the algorithm's way of minimizing information loss but to maintain it and achieve inclusion of personal identifiable data.

4.2 Identifiers and Concatenated Quasi-Identifiers in the k-anonymity algorithm

In the k-anonymity algorithm, we modified the approach to include steps 2, 3, and 5. These steps included separate data schema design. The figure below shows the extracted and mapped data attributes for identifiers and concatenated quasi-identifiers.

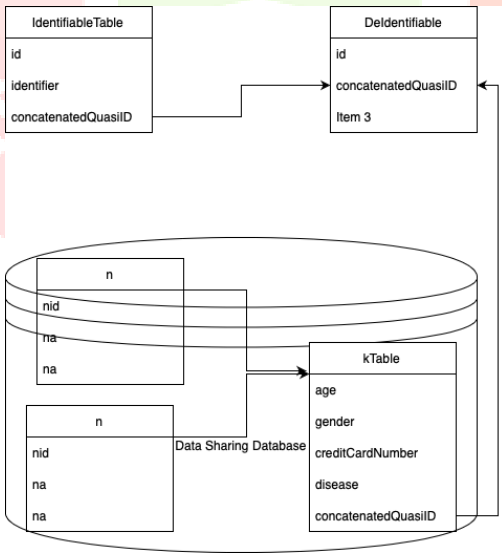

Figure 4.2: Custom dataset for the extracted identifiers and quasi-identifiers, and their mapping to tha compatible transformed dataset

4.3 Conclusion

Through this study, we achieve sharing of big personal data with identified data while achieving data privacy and anonymization. We ably modify the k-anonymization method to include identified data and map them to the de-identified data through quasi-identified data. We created an algorithm that has seen a minimal ratio of information loss.

4.4 Future Enhancements

Create a model that can combine the quasi-identifiers and auto map them to their respective identifiers. The model could be to rapidly assess the rate of information loss in created clusters.

Allow policies to enable data sharing with personal identifiable data, however, on condition that there is minimal information loss, and that data re-identification is not harmful.

ACKNOWLEDGMENT

Thanks to DMI – St. Eugene University for the support rendered towards this study.

REFERENCES

[1] Rasim, M. Algguliyev, Ramiz, M., Aligguliyev and Fargana J. Abdullayeva. 2019.Privacy-preserving deep learning algorithm for big personal data analysis. Journal of Industrial Information Integration, 15: 1–14.

[2] Ajmeera Kiran, N. Shirisha. 2022. K-Anonymization approach for privacy preservation using data perturbation techniques in data mining.

[3] Sol, C. & Njilla, L. & Kwiat, K. & Kamhoua, C. (2020). Learning quasi-identifiers for privacy-preserving exchanges: a rough set theory approach. Granular Computing, 5 (1).

[4] Sweeney, L. 2022. K-anonymity: a model for protecting privacy. International Journal for Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5): 557–570.

[5] Anastasiia Girka, Vagan Terziyan, Mariia Gavriushenko, Andrii Gontarenko. 2021. Anonymization as homeomorphic data space transformation for privacy-preserving deep learning. Procedia Computer Science, 180: 867-876.

[6] Vimalkumar B. Vaghela. 2020. K-Anonymization Approach for Privacy Preserving in Data Mining. International, Journal of Science & Technology Research, 9(01): 1-6.

[7] Angiuli, Olivia Marie. 2015. The effect of quasi-identifier characteristics on statistical bias introduced by k-anonymization. Bachelor's thesis, Harvard College.

[8] M. E. Kabir, H. Wang and E. Bertino. 2011. Efficient systematic clustering method for k-anonymization. Acta Informatica. 48: 51-66.

[9] Kato Mivule. 2017. Data Swapping for Private Information Sharing of Web Search Logs. Procedia Computer Science. 114: 149-158.

[10] Ouazzani, Z.E., & Bakkali, H.E. 2018. A new technique ensuring privacy in big data: K-anonymity without prior value of the threshold k. Procedia Computer Science. 127: 52-59.