IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 11, Special Issue 2, March 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

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

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

| International Conference on Recent Innovations in Engineering and Technology (ICRIET'23)|

| Sharadchandra Pawar College of Engineering, Pune, India |

|| Volume 11, Special Issue 2, March 2023 ||

Data Privacy Protection Based on Microaggregation Using Machine Learning Techniques

Donapati Srikanth Chaitanya, Dr.G.Madhavi

Research Scholar, Chaitanya deemed to be University, Hanamkonda, Warangal, India

Assistant Professor, Chaitanya deemed to be University, Hanamkonda, Warangal, India

ABSTRACT: In the era of big data, the availability of massive amounts of information make privacy protection more necessary than ever. The preservation of privacy of sensitive attributes in healthcare system is securely and efficiently achieved by using Microaggregation as well as Machine Learning techniques. Microaggregation is a technique for disclosure limitation aimed at protecting the privacy of data subjects in micro data releases. It has been used as an alternative to generalization and suppression to generate k-anonymous data sets, where the identity of each subject is hidden within a group of k subjects. This paper presents Data privacy protection based on Microaggregation using Machine learning techniques. We apply this new concept in the context of mobile health and we show that a distributed architecture consisting of patients and several intermediate entities can apply it to protect the privacy of patients, whose data are released to third parties for secondary use. After recalling some fundamental concepts of statistical disclosure control and microaggregation, we detail the distributed architecture that allows the private gathering, storage, and sharing of biomedical data. We quantify utility accordingly as the accuracy and privacy of machine learning models from microaggregated data, evaluated over original test data.

KEYWORDS: Microaggregation, Machine learning, Privacy, Large-scale databases, mobile health.

I. INTRODUCTION

Due to the recent advances in information and communication technologies the gathering, storage and sharing of data are becoming simpler and faster than ever. The myriad of benefits these technologies can bring to private companies, public institutions and, in general, our society are innumerable. Healthcare, transportation, banking and marketing are just a few fields in which big-data analytics is leading a profound transformation of the traditional models. In this regard, one of the most significant breakthroughs is the use of mobile devices (e.g., cell phones) to monitor patients [1].

Mobile communications are experiencing a tremendous development that leads to new ways of providing healthcare services, the so-called mobile health (m-health) [2]. m-Health affects the way we understand healthcare services in three main aspects: (a) m-Health simplifies the access to classical and new services. (b) m-Health is patient-oriented. (c) m-Health is personalized. Patients receive customized services that fit their specific needs.

Most of the data today is published is in the form of microdata. Microdata are database tables whose records carry data concerning individual subjects. A microdata set is a file with n records and every record contain m variables also called as attributes of an individual of whom the information is collected. To protect this microdata from individual identification, Statistical Disclosure Control (SDC) methods are applied before the data is disseminated or published for analysis [3]. SDC method seeks to alter the original microdata so that the statistical analysis of the original data and the published data are almost similar i.e., the information loss is less, and the disclosure risk of individual identification is low. Microaggregation is one of SDC methods used for protecting microdata sets. Microaggregation works into steps. Step I: Partition the dataset S into clusters of k groups consisting of at least k-records. Step II: Publish a value (commonly called as centroid which could be median, mode variance) over each group, replacing the original values [4].

In some cases, the owners of the cloud or data centers need to publish the data. Therefore, how to make the best use of the data in the risk of personal information leakage has become a popular research topic. The most common method of

data privacy protection is the data anonymization, which has two main problems: (1) The availability of information after clustering will be reduced, and it cannot be flexibly adjusted. (2) Most methods are static. When the data is released multiple times, it will cause personal privacy leakage. To solve the problems, two contributions are introduced. The first one is to propose a new method based on micro-aggregation using Machine Learning techniques (classification and clustering). In this way, the data availability and the privacy protection can be adjusted flexibly by considering the concepts of distance. The second contribution is to propose a dynamic update mechanism that guarantees that the individual privacy is not compromised after the data has been subjected to multiple releases, and minimizes the loss of information.

II. LITERATURE SURVEY

J. Zhang, G. Cormode, C.M. Procopiuc, D. Srivastava, and X. Xiao, et al. [5] considered the data utility of differential privacy and constructed a Bayesian network to the data prior to adding noise; in addition, they theoretically analyzed the privacy and provided a utility guarantee. G. Cormode, E. Shen, X. Gong, T. Yu, C. M. Procopiuc, and D. Srivastava, [6] proposes publishing synthetic microdata generated from differentially private models applied on original data. For that, machine learning techniques are integrated to improve utility.

G. Cormode, C.M. Procopiuc, E. Shen, D. Srivastava, and T. Yu, et al. [7] analyzed in detail the connection, advantages and disadvantages of using microaggregation to realize differential privacy. In short, the methods for improving the utility of differentially private data publishing differ in terms of their limitations. Compared with conventional differential privacy methods, the microaggregation methods can improve the data utility and do not impose excessive restrictions on the query type. J. Soria-Comas, J. Domingo-Ferrer, D.S. Anchez, and S. Martinez, et al. [8] proposed the synergy between differential privacy and k-anonymity when publishing anonymous data. However, their method was not sufficient for considering the utility of numerical attributes.

J. Soria-Comas and J. Domingo-Ferrer. et. al. [9] introduces the probabilistic k-anonymity property, which relaxes the indistinguishability requirement of k-anonymity and only requires that the probability of re-identification be the same as in k-anonymity. Two computational heuristics to achieve probabilistic k-anonymity based on data swapping are proposed: MDAV microaggregation on the quasi-identifiers plus swapping, and individual ranking microaggregation on individual confidential attributes plus swapping. We report experimental results, where we compare the utility of original, k-anonymous and probabilistically k-anonymous data.

N. Li, T. Li, and S. Venkatasubramanian. et. al. [10] proposes a new notion of privacy called "closeness." We first present the base model t-closeness, which requires that the distribution of a sensitive attribute in any equivalence class is close to the distribution of the attribute in the overall table (i.e., the distance between the two distributions should be no more than a threshold t). We then propose a more flexible privacy model called (n,t)-closeness that offers higher utility. We describe our desiderata for designing a distance measure between two probability distributions and present two distance measures. We discuss the rationale for using closeness as a privacy measure and illustrate its advantages through examples and experiments.

III. DATA PRIVACY PROTECTION

The schematic of Data privacy protection based on Microaggregation using Machine learning techniques is represented in below Fig. 1. The proposed architecture considers main actors, namely data owners (mobile devices, patients), healthcare centers (HC), research centers and a centralized storage and aggregation server (SAS). These actors/entities interact so as to guarantee the private collection and sharing of data.

Patients have mobile devices with communication capabilities able to collect data and encrypt them by using a public key cryptosystem. The data collected from patients have the same attributes for all patients participating in a given study/trial. Note that this is quite common because clinical studies tend to be highly parametrized and strict with the treated variables. We refer to these collected data as,

$$D = (d_{u1}, d_{u2}, \dots \dots d_{ui} \dots \dots d_{un}) \dots (1)$$

Where represents the data of user/patient. After collecting the data, the mobile device of each patient encrypts the gathered data with the public key of the healthcare center to which he/she is assigned and generates a message with the following information:

• User ID:
• Healthcare Center ID:
• Encrypted data

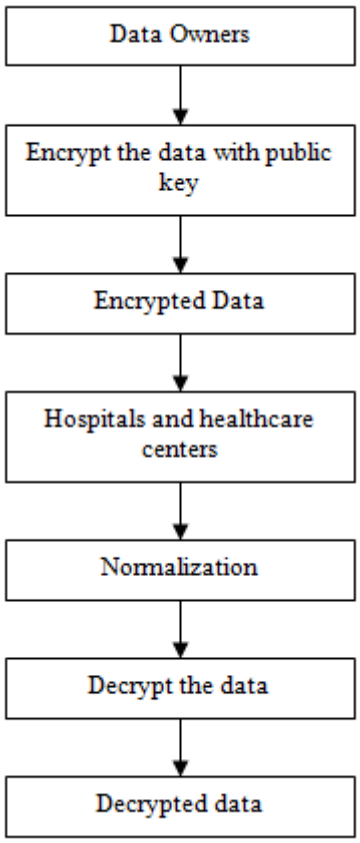

Fig. 1: ARCHITECTURE OF DESCRIBED DATA PRIVACY PROTECTION

After encrypting the data and generating the message described above, the mobile device sends to the storage and aggregation server (SAS). Once the SAS receives the patients' data, it checks the healthcare center ID of each message and forwards it to right. Note that the SAS cannot access to the raw data of the patients because it is encrypted with the public key of HCi. Thanks to the use of their private key, each healthcare center can decrypt and access the raw biomedical data of the patients for which they are responsible.

By doing so, doctors can analyze the complete data and decide on the proper procedures and protocols to apply without any information loss (in this regard we assume that healthcare centers are trusted). After receiving and decrypting the data from the patients participating in a given trial/study, each healthcare center microaggregates all the data (by using a microaggregation algorithm such as MDAV) with a given security parameter HC(k) and sends the resulting microaggregated data set back to the SAS. To this end, previously we follow the common practice of normalizing each column of the data to have zero mean and unit variance. With the microaggregated versions of each (training) data set, we then construct a classification model over each of those versions using Weka and 10-fold cross validation.

When the SAS receives the microaggregated data sets from all healthcare centers, it merges them all and microaggregates them again by using again a microaggregation algorithm such as MDAV, with a given security parameter SAS(k). Finally, we evaluate the accuracy of the resulting classification models over the non-anonymized test subset, reproducing the application scenario where a database user would use the classification model to classify their original samples of data.

IV. RESULT ANALYSIS

The proposed architecture guarantees the private exchange of data between patients and healthcare centers through the SAS thanks to the use of public key cryptography. In this regard, even if the SAS is not a trusted party the architecture remains private. Due to privacy issues, we have not used biomedical data belonging to real patients. However, with the aim to provide consistent evidence of the usefulness of our proposal, we have created a synthetic but realistic biomedical data set, based on real data. Our data set consists of three quasi-identifiers (i.e., age, weight and height) and five outcome attributes: maximum blood pressure (systolic), minimum blood pressure (diastolic), heartbeat rate, blood glucose level and blood oxygen saturation.

Compare the performance in terms of information loss and disclosure risk of microaggregation model based on MDAV (Maximum Distance to Average Vector) with and without machine learning models. The aim of this comparison is to determine which of these methods is the best candidate to be implemented later as the main component of the microaggregation algorithm of our architecture.

Information loss (IL) is defined as the ratio between the Sum of Square Errors (SSE) and the Sum of Square errors Total (SST).

Disclosure risk (DR) is a standard measure based on record linkage. Considering that an attacker has a given subset of quasi-identifiers, the probability of re-identification success is measured.

Table 1: PERCENTAGE OF IL AND DR FOR SEVERAL VALUES OF K

Method	Security parameter (k)	Information loss (IL)	Disclosure risk (DR)
MDAV microaggregation algorithm with machine learning (MDAV-with ML)	2	22	95
	4	31	79
	6	38	72
	8	45	65
	10	55	53
MDAV microaggregation algorithm without machine learning (MDAV-without ML)	2	25	90
	4	34	75
	6	56	62
	8	75	49
	10	84	33

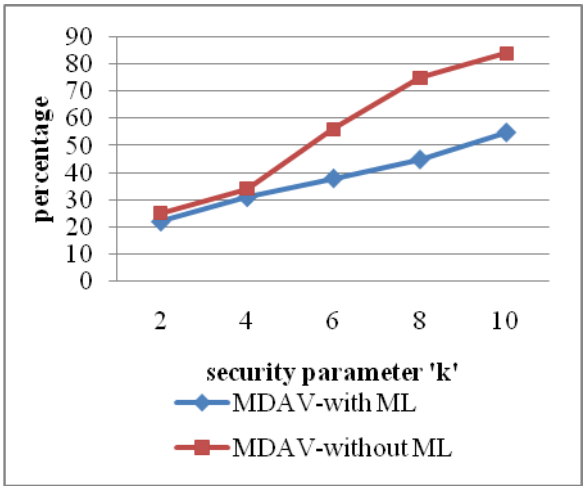

Fig. 2: COMPARATIVE ANALYSIS IN TERMS OF INFORMATION LOSS (IL)

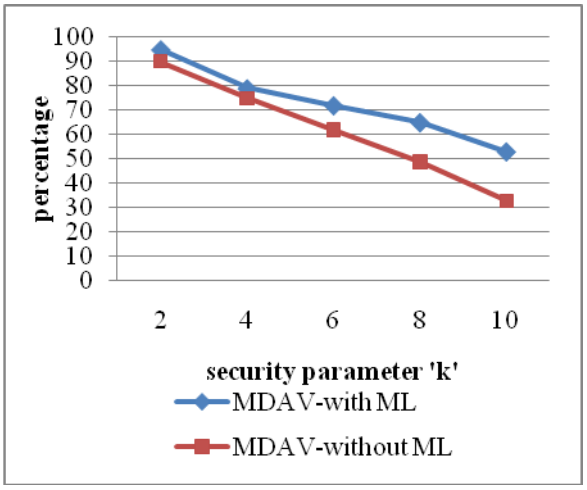

Fig. 2: COMPARATIVE ANALYSIS IN TERMS OF DISCLOSURE RISK (DR)

As it can be observed in Fig. 2 and Fig. 3, the IL and DR evolve inversely. The lower the IL the higher the DR. MDAV microaggregation algorithm with machine learning performs better than the MDAV microaggregation algorithm without machine learning in terms of IL with a reasonable DR. below Fig. 4 shows the achieved Accuracy and Privacy performance of MDAV-with ML and MDAV-without ML.

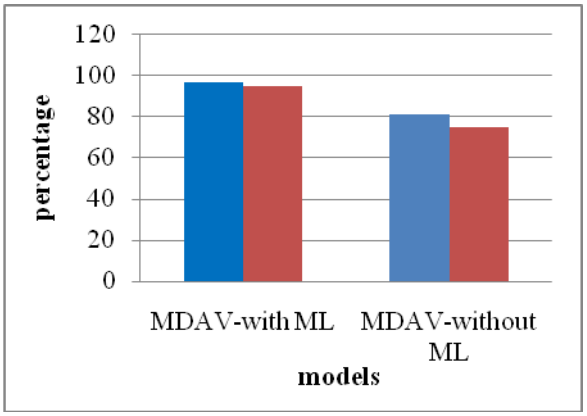

Fig. 4: ACCURACY AND PRIVACY ANALYSIS

MDAV microaggregation algorithm with machine learning model performs better than the MDAV microaggregation algorithm without machine learning model.

V. CONCLUSION

In this paper, Data privacy protection based on Microaggregation using Machine learning techniques is described. The preservation of privacy of sensitive attributes in healthcare system is securely and efficiently achieved by using Microaggregation as well as Machine Learning techniques. The proposed architecture guarantees the private exchange of data between patients. Patients have mobile devices with communication capabilities able to collect data and encrypt them by using a public key cryptosystem. When the SAS receives the microaggregated data sets from all healthcare centers, it merges them all and microaggregates them again by using again a microaggregation algorithm such as MDAV, with a given security parameter SAS(k). Compare the performance in terms of information loss and disclosure risk of microaggregation model based on MDAV (Maximum Distance to Average Vector) with and without machine learning models. From results, MDAV microaggregation algorithm with machine learning model performs better than the MDAV microaggregation algorithm without machine learning model.

REFERENCES

[1] Zeki Saeed Tawfik, Alaa Hussein Al-Hamami, Mustafa Tareq Abd, "Comparison of Data Mining Techniques in Healthcare Data", 2022 International Conference for Natural and Applied Sciences (ICNAS), Year: 2022

[2] Haibin Yang, "Application and Development of Mobile Communication Technology", 2021 International Wireless Communications and Mobile Computing (IWCMC), Year: 2021

[3] Md Enamul Kabir, Abdun Naser Mahmood, Hua Wang, Abdul K. Mustafa, "Microaggregation Sorting Framework for K-Anonymity Statistical Disclosure Control in Cloud Computing", IEEE Transactions on Cloud Computing, Volume: 8, Issue: 2, Year: 2020

[4] Veena Gadad, Sowmyarani C.N., "Understanding Microaggregation- A technique of Statistical Disclosure Control for Privacy Preserving and Data Publishing in Inter-Cloud", 2018 Second International Conference on Advances in Electronics, Computers and Communications (ICAECC), Year: 2018

[5] J. Zhang, G. Cormode, C.M. Procopiuc, D. Srivastava, and X. Xiao, "PrivBayes: Private data release via Bayesian networks," ACM Trans. Database Syst., vol. 42, no. 4, pp. 1-41, Oct. 2017.

[6] G. Cormode, E. Shen, X. Gong, T. Yu, C. M. Procopiuc, and D. Srivastava, "UMicS: From anonymized data to usable microdata," in Proc. ACM Int. Conf. Inform., Knowl. Mgmt. (CIKM), San Francisco, CA, Oct. 2013, pp. 2255–2260

[7] G. Cormode, C.M. Procopiuc, E. Shen, D. Srivastava, and T. Yu, "Empirical privacy and empirical utility of anonymized data," in Proc. 2013 IEEE 29th International Conference on Data Engineering Workshops Brisbane, QLD, Australia, 2013, pp. 77-82.

[8] J. Soria-Comas, J. Domingo-Ferrer, D.S. Anchez, and S. Martinez, "Improving the utility of differentially private data releases via k-anonymity," in 2013 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, Melbourne, VIC, Australia, 2013, pp. 372-379.

[9] J. Soria-Comas and J. Domingo-Ferrer. "Probabilistic k-anonymity through microaggregation and data swapping", In Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2012), pp. 1–8, IEEE, 2012

[10] N. Li, T. Li, and S. Venkatasubramanian. "Closeness: a new privacy measure for data publishing", IEEE Transactions on Knowledge and Data Engineering, 22(7):943– 956, 2010.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.379

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details