

Dataset Anonimyzation for Machine Learning: An ISP Case Study

Lelio Campanile[1], Fabio Forgione[2], Fiammetta Marulli[1], Gianfranco Palmiero[2], and Carlo Sanghez[2(✉)]

[1] Universitá degli Studi della Campania "L. Vanvitelli", Dip. Matematica e Fisica, Caserta, Italy
{lelio.campanile,fiammetta.marulli}@unicampania.it
[2] GAM Engineering srlu, Caserta, Italy
{fforgione,gpalmiero,csanghez}@gameng.it
https://www.gameng.it

Abstract. Internet Service Providers technical support needs personal data to predict potential anomalies. In this paper, we performed a comparative study of forecasting performance using raw data and anonymized data, in order to assess how much performance may vary, when plain personal data are replaced by anonymized personal data.

Keywords: ISP · Cryptography · Customer Premise Equipment · WISP · Logistic regression · Random forest · Attribute suppression · Character masking · Hash · Pseudo-anonymization · Data generalization

1 Introduction

Preservation of privacy in data mining and machine learning has emerged as an absolute prerequisite in several practical scenarios, especially when the processing of sensitive data is outsourced to an external third party. Currently, privacy preservation methods are mainly based on randomization and/or perturbation, secure multiparty computations and cryptographic methods.
In [7] is proposed a strategy that leverages partial homomorphic property of some crypto-systems to train n simple machine learning models on encrypted data. The basic scenario discussed involves three parties, represented by: 1)multiple Data Owners, which provide encrypted training examples; 2) the Algorithm Owner (or Application), processing them to adjust the parameters of its models; 3) a semi-trusted third party, which provides privacy and secure computation services to the Application, when operations that are not supported by the homomorphic cryptosystem occur. More precisely, their focus is on the use of multiple-key crypto-systems, and on the impact of the quantization of real-valued input data, required before encryption.

In [10], the authors design a protocol framework to perform classification tasks in a privacy-preserving manner. To demonstrate the feasibility of their proposal, two protocols, supporting Naive Bayes classification, are implemented and

© Springer Nature Switzerland AG 2021
O. Gervasi et al. (Eds.): ICCSA 2021, LNCS 12950, pp. 589–597, 2021.
https://doi.org/10.1007/978-3-030-86960-1_42

the heavy computational load of conventional fully homomorphic encryption-based privacy-preserving protocols, is overcome by adopting different optimization techniques. This method differs from other existing techniques because it requires no intermediate interactions between the server and the client while executing the protocol, except for the mandatory interaction to obtain the decryption result of the encrypted classification output. As a result of this minimal interaction, the proposed method results in a relatively stable solution. Furthermore, in [8], is presented the *CryptoDL* framework, thus offering new techniques to provide solutions for applying deep neural network-based algorithms to encrypted data.

In the work we propose, we address the problem of evaluating the performance degradation when a classification task is performed over the same data samples but several different machine learning algorithms and various anonymization methodologies are adopted.

2 Related Work

The main field of application of the studies introduced in the previous section, is focused on the personal privacy requirements of each individual, thus preserving as much information as possible and when it applied to a large quantity of data preserving also good performance.

For the purposes of the analysis, there is the need to limit the risks of disclosure to an acceptable level, while retaining the information and thus the usefulness. When using data sets containing sensitive information and data that should not be disclosed, it becomes necessary to use anonymization techniques on the raw data.

In [9], various privacy techniques are compared, evaluating the strengths and weaknesses of the different techniques. An interesting insight that focuses on the possibility of pseudo-anonymizing data and subsequently de-pseudonymizing it, is presentend in [2]; in that case, healthcare patients data are used and the proposed technique enables searches on anonymous aggregated data, then reporting the results to the individual patient.

Several national authorities in different countries have issued interesting guidelines related to privacy and anonymization and pseudo-anonymization of personal data [4], [12], [11].

A more complex approach to pseudo-anonymization of data is offered by the *K-anonimization* technique, which is investigated in [14] and [13].

However, most of the anonymization and pseudo-anonymization techniques result to be inefficient, when evaluated in terms of performance; furthermore, they introduce an overhead to the operations that need to be performed on the data, including machine learning operations. The authors of the work described in [5], face the performance problems and propose a framework that addresses the data privacy preservation in an efficient way. In [6], the authors address, in particular, the topic of the application of machine learning algorithms on anonymized data, carefully concerning with the problems related to the accuracy and performance of the resulting models.

3 Case Study

In this section we describe the case study on which is based our study on performance decay. It consists in *Flyber*[1], a Wireless Internet Service Provider (WISP) based on Teverola, Italy, that offers its customers free technical support to solve problems related to the quality of signal reception. Since the technology is wireless over open bands, the quality of the connection depends mainly on radio interference. To check the health status of the Customer Premise Equipments, Flyber adopts a monitoring system based on two servers: the first one is the Dude[2] network monitor by MikroTik and, the second one, is cnMaestro by Cambium Networks [1]. Since most clients use Cambium CPEs, we focused only on cnMaestro related data. The monitoring system produces a daily report with 25 fields. An extract of a daily report is showed in (Fig. 1).

Polling Timestam	MAC	Device Name	Device Type	Downlink	Uplink MCS	Downlink	Uplink Th	Downlink	Uplink Re	CPUs	Downlink	Uplink SN	Downlink	Uplink Us	Downlink	Uplink RS	Offline Cc	Status	Status Duration
24/12/20 01:00	00:04:56:20	[redacted]	ePMP Force 300-16	DS 6	DS 3		585	36.7		4	7.4	1.71	34	29	5406654	1770722	-56	-65	0	Online	8d 4h 38m
24/12/20 01:00	00:04:56:20	[redacted]	ePMP Force 300-16	DS 8	DS 5	462.3		22.3	4.5		1.37		29	23	4529254	1521332	-61	-69	0	Online	16d 16h 32m
24/12/20 01:00	00:04:56:23	[redacted]	ePMP Force 300-16	DS 8	DS 6	3379.7	111.7		1	3.8	3.72		36	32	27023208	15262559	-53	-60	0	Online	4d 10h 40m
24/12/20 01:00	00:04:56:23	[redacted]	ePMP Force 300-16	0	0	0		0	0		0.00		0	0	0	0	0	0	0	Offline	0d 13h 25m
24/12/20 01:00	00:04:56:23	[redacted]	ePMP Force 300-16	DS 8	DS 7		221	12.2	0.1	2.1	2.50		37	31	1826868	1457965	-54	-61	1	Online	0d 0h 0m
24/12/20 01:00	00:04:56:23	[redacted]	ePMP Force 300-16	DS 8	SS 9	845.4		76.3		0	1.1	2.07	42	34	7873754	3177378	-49	-59	0	Online	1d 11h 38m
24/12/20 01:00	00:04:56:23	[redacted]	ePMP Force 300-16	DS 9	DS 5	1.5		0.2		0	0.4	1.84	41	34	272987	39935	-50	-59	0	Online	1d 11h 38m
24/12/20 01:00	00:04:56:23	[redacted]	ePMP Force 300-16	DS 7	SS 9	9.7		5.3	1.5	0.3	3.39		37	27	324523	113030	-54	-59	0	Online	15d 2h 29m
24/12/20 01:00	00:04:56:23	[redacted]	ePMP Force 300-16	DS 7	SS 2	581.9		26.1		3	1.4	1.93	29	26	6238650	2059942	-61	-66	0	Online	7d 7h 54m
24/12/20 01:00	00:04:56:23	[redacted]	ePMP Force 300-16	DS 8	DS 7	780.4		25.7	1.2	2.1	3.29		36	30	6190973	3773458	-53	-64	1	Online	6d 9h 54m
24/12/20 01:00	00:04:56:23	[redacted]	ePMP Force 300-16	DS 8	DS 4		2127	58.4		0	4.9	1.60	51	35	16826692	4732560	-39	-59	0	Online	1d 11h 38m

Fig. 1. cnMaestro on premises performance 24 h system daily report.

Flyber staff uses a customer relationship management system based on a custom software solution. In Fig. 2 is shown an extract of records.

The first part of the work consists in classifying the records according to two conditions: with and without reporting by the customer. In the first step, we obtained a model able to classify a given condition on the CPE, by using personal data without anonimyzation. Since it was a classic binary classification, we considered the logistic regression and the random forest algorithms. The second step consisted in transforming the data set through the anonymization of personal data and, finally, the last step was to obtain a model able to classify a given condition on the CPE using data from anonimous data set, in order to assess performance changes. The considered time frame was between 21st December 2020 and 23rd March 2021. During this time frame, we collected 68 reports, each including daily data incoming from about 600 antennas and we received 1580 reports to Flyber's customer care. The experiment was therefore conducted on about 40,000 samples. As regards the samples of the reports to customer care, the experiment was conducted by cleaning the data set from administrative issues, or referred to non-Cambium equipment.

[1] www.flyber.it.
[2] https://mikrotik.com/thedude, [Online; accessed 02-May-2021].

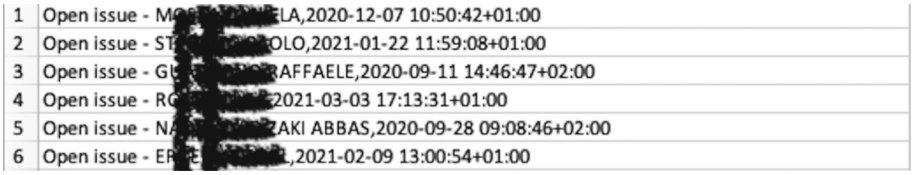

Fig. 2. CRM records with personal data.

4 Discussion and Results

4.1 Raw Dataset

The experiment was conducted considering a data set with 43780 rows × 25 columns. In the following, a description of the columns that include personal data, is provided.

- **MAC** is the MAC address of the CPE.
- **Device Name** is the configured device name of the AP, used for identifying the device in an NMS such as the Cambium Network Services Server (CNSS). It includes personal data.
- **COD. CLIENTE** is the customer ID.
- **COMUNE INSTALLAZIONE** is the ZIP code of customer.

The initial data set was further cleaned by deleting useless, null or dirty records. In fact, sometimes the CPE communicates incorrect data, or the monitoring system adds null values when the CPE does not respond in time. By doing so, the number of samples was reduced to about 17.000.

Classification of Data Set with Personal Data Using the Logistic Regression: the first classification tests were not satisfactory. Indeed, the system was able to correctly classify less than one out of five anomalies.

Classification of Data Set with Personal Data Using the Random Forest: the first classification tests were satisfactory. Indeed, the system was able to correctly classify more than one out of three anomalies. The results are summarized in the Table 1. In the first column of the table we reported the total number of samples, in the second column we reported the percentage of samples with issues, in the third column of the table we reported the percentage of success in identifying the critical samples using logistic regression. Finally, in the last column we reported the success rate we achieved using the random forest on the non-anonymous data set. We could have improved the performance of the algorithms further or changed the algorithm, but the goal here is to measure the performance drop by anonymizing the data. So, we have selected the random forest classification algorithm.

Table 1. Data set with personal data: logistic regression (LR) VS Random forest (RF)

Samples	% Issues	% of success by LR	% of success by RF
17.000	3,2%	20%	37%

4.2 Anonymous Data Set

Our intent was to follow good practices in terms of data anonymization. We knew that we would reduce the original information present in the data set, but the question was how much we would lose in terms of performance. We have chosen to use different approaches to anonymize the data.

Details below.

- **Attribute suppression:** refers to the removal of the personal part from data. We simply deleted the columns with personal data from the data set.
- **Character Masking:** refers to the changing some characters of a personal data record by using a 'x'.
- **Pseudo-anymization by HASH:** refers to the replacement of personal data with anonymous data. We simply used a hash function on strings containing personal data. The new value was used instead of the old one.
- **Generalization:** refers to the converting personal data in an range of values. We simply grouped personal values in order to keep the rough information, but accepting the loss of detail.
- **Generalization and k-anonymity:** refers to the converting data in an range of values as seen before, but in this case we deliberately deleted those samples that had no other k-similar. We did not use K-anonymization due to the size of our data set [3].

Results from Attribute Suppression. We have simply deleted columns that include personal data from the data set. By applying the classification algorithm we obtained a success rate of 31,79%. In fact, during our tests, the system was able to correctly classify 55 out of 173 samples.

Results from Character Masking. We have replaced the last three octets of the MAC with the 'x' character. We have replaced the last four characters of the 'Device name' column with the 'x' character. We have replaced the last two characters of the 'COMMON INSTALLATION' column with the 'x' character. Finally, we have replaced the last three characters of the 'COD. CUSTOMER' with the character 'x'. In this way, we have made the data set completely anonymous. Below is the code block that performs the masking.

```
XY_data_ctc['MAC_MASK']=\
XY_data_ctc['MAC'].str[:-8]+'xx:xx:xx'

XY_data_ctc['Device Name_MASK']=\
```

```
XY_data_ctc['Device Name'].str[:-4]+'xxxx'

XY_data_ctc['COMUNE INSTALLAZIONE_MASK']=\
XY_data_ctc['COMUNE INSTALLAZIONE'].str[:-2]+'xx'
```

By applying the classification algorithm we obtained a success rate of 35,26%. In fact, during our tests, the system was able to correctly classify 61 out of 173 samples. The critical issue here is the time that is lost in transforming personal data via code. In fact, on the ARM Linux system used for the inference tests, about 0,158 s are lost to anonymize the 400 CPEs data. In our case study, samples from 2000 CPEs are collected every second, so considering the time lost to anonymize the data using masking function, we lose 0,785s for each step.

Results from Pseudo-Anonymization. We simply applied a hash function directly to the personal data. Thus, the number of columns and features did not change. We have modified the MAC, Device_Name, COD_CLIENTE, COMUNE_INSTALLAZIONE columns using Python code. The code is shown below.

```
import hashlib

def encrypt_string(hash_string):
    sha_signature = \
        hashlib.sha256(hash_string.encode()).hexdigest()
    return sha_signature

MACRIPTO = []

for row in XY_data_ctc['MAC']:
    temp = (str(row))
    encMessage = encrypt_string(temp)
    MACRIPTO.append(encMessage)

XY_data_ctc['MAC_CRIPTO_HASH'] = MACRIPTO

DeviceNameCRIPTO = []

for row in XY_data_ctc['Device Name']:
    temp = (str(row))
    encMessage = encrypt_string(temp)
    DeviceNameCRIPTO.append(encMessage)

XY_data_ctc['Device Name_HASH'] = DeviceNameCRIPTO

CODCLIENTECRIPTO = []

for row in XY_data_ctc['COD.CLIENTE']:
    temp = (str(row))
    encMessage = encrypt_string(temp)
```

```
    CODCLIENTECRIPTO.append(encMessage)

XY_data_ctc['COD. CLIENTE_HASH'] = CODCLIENTECRIPTO

CCRIPTO = []

for row in XY_data_ctc['COMUNE INSTALLAZIONE']:
    temp = (str(row))
    encMessage = encrypt_string(temp)
    CCRIPTO.append(encMessage)

XY_data_ctc['COMUNE INSTALLAZIONE_HASH'] = CCRIPTO
```

By applying the classification algorithm we obtained a success rate of 37,57%. In fact, during our tests, the system was able to correctly classify 65 out of 173 samples. The critical issue here is the time that is lost in transforming personal data via code. In fact, on the ARM Linux system used for the inference tests, about 0,4778 s are lost to anonymize the 400 CPE data. In our case study, samples from 2000 CPEs are collected every second. Considering the time lost to anonymize the data using a function, we lose too much time at every step. The wasted time is 2,389 s at every step.

Results from Generalization. We have eliminated the last octet present in the 'MAC' column, we have grouped the values present in the 'CUSTOMER CODE' column into groups of 500, we have eliminated the last two digits from the ZIP code present in the 'COMMON INSTALLATION' column and finally we have considered only the first letter present in the values of the 'DEVICE NAME' column. In this way, the number of features has changed. Indeed, the non-anonymous data set included more than 1200 characteristics, while the grouped data set has 311 characteristics.

The code is shown below.

```
XY_data_ctc['MAC_GEN']=XY_data_ctc['MAC'].str[:-2]
XY_data_ctc['Device Name_GEN']=\
XY_data_ctc['Device Name'].str[:1]
XY_data_ctc['COD. CLIENTE_GEN'] =\
XY_data_ctc['COD. CLIENTE'].str[3:]
XY_data_ctc['COD. CLIENTE_GEN']=\
pd.to_numeric(XY_data_ctc['COD. CLIENTE_GEN'])
XY_data_ctc.loc[XY_data_ctc['COD. CLIENTE_GEN'] \
<= 499 ,['COD. CLIENTE_GEN']]   = 10000
XY_data_ctc.loc[XY_data_ctc['COD. CLIENTE_GEN'] \
<= 999 ,['COD. CLIENTE_GEN']]   = 20000
XY_data_ctc.loc[XY_data_ctc['COD. CLIENTE_GEN'] \
<= 1499 ,['COD. CLIENTE_GEN']]  = 30000
XY_data_ctc.loc[XY_data_ctc['COD. CLIENTE_GEN'] \
<= 1999 ,['COD. CLIENTE_GEN']]  = 40000
XY_data_ctc.loc[XY_data_ctc['COD. CLIENTE_GEN'] \
<= 2499 ,['COD. CLIENTE_GEN']]  = 50000
```

```
XY_data_ctc.loc[XY_data_ctc['COD. CLIENTE_GEN'] \
<= 2999 ,['COD. CLIENTE_GEN']]  = 60000
XY_data_ctc['COMUNE INSTALLAZIONE_GEN']=\
XY_data_ctc['COMUNE INSTALLAZIONE'].str[:-2]
```

By applying the classification algorithm we obtained a success rate of 35,26%. In fact, during our tests, the system was able to correctly classify 61 out of 173 samples. The critical issue here is the time that is lost in transforming personal data via code. In fact, on the ARM Linux system used for the inference tests, about 0,17 s are lost to anonymize the 400 CPE data. In our case study, samples from 2000 CPEs are collected every second. Considering the time lost to anonymize the data using this algorithm, we lose too much time at every step. The wasted time is 0,85 s at every step.

Table 2. Attribute suppression (AS) VS Character Masking (CM) VS Pseudo-anonymization (PS) VS Generalization (GE)

Algorithm	Attribute suppression	Character masking	Pseudo-anonymization by HASH	Generalization
Success rate	31,79%	35,26%	37,57%	35,26%
Loss of data?	NO	NO	NO	NO
Time issue?	NO	YES We lose 0,785 s for each step	YES We lose 2,389 s for each step	YES We lose 0,85 s for each step

5 Conclusion and Future Work

In this work, we explored the impact of adopting anonymization and pseudo-anonymization in a real-word application, when a Machine Learning algorithm is performed on anonymized data. Our experiments were focused on computing performance and accuracy, when data anonymization is applied or not. The preliminary results bring evidence of certain differences among the adoption of different techniques, as it is summarized in Table 2. As evidenced from the results provided, the best performance was obtained by applying hashing techniques, but the time spent removing personal data was a big issue. The future direction of this work consists in improving the effectiveness of the anonymization, evenly adopting the k-anonymization technique, and to enhance the computing performance, by distributing the computational load on multiple computational devices.

References

1. Cambium Networks. https://www.cambiumnetworks.com/products/software/cnmaestro-management/. Accessed 02 May 2021
2. Aamot, H., Kohl, C.D., Richter, D., Knaup-Gregori, P.: Pseudonymization of patient identifiers for translational research. BMC Med. Inform. Decis. Mak. **13**(1), 1–15 (2013). https://doi.org/10.1186/1472-6947-13-75
3. Aggarwal, C.C.: On k-anonymity and the curse of dimensionality. VLDB **5**, 901–909 (2005)
4. Article 29 Data Protection Working Party: Opinion 05/2014 on Anonymisation Techniques. Working Party Opinions (April), 1–37 (2014). http://ec.europa.eu/justice/data-protection/index_en.htm%0Aec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf
5. Ghinita, G., Karras, P., Kalnis, P., Mamoulis, N.: Fast data anonymization with low information loss. In: 33rd International Conference on Very Large Data Bases, VLDB 2007 - Conference Proceedings, pp. 758–769 (2007)
6. Goldsteen, A., Ezov, G., Shmelkin, R., Moffie, M., Farkash, A.: Anonymizing machine learning models. arXiv (2020)
7. González-Serrano, F.-J., Amor-Martín, A., Casamayón-Antón, J.: Supervised machine learning using encrypted training data. Int. J. Inf. Secur. **17**(4), 365–377 (2017). https://doi.org/10.1007/s10207-017-0381-1
8. Hesamifard, E., Takabi, H., Ghasemi, M., Wright, R.N.: Privacy-preserving machine learning as a service. In: Proceedings on Privacy Enhancing Technologies, vol. 2018, no. 3, pp. 123–142 (2018)
9. Murthy, S., Abu Bakar, A., Abdul Rahim, F., Ramli, R.: A Comparative study of data anonymization techniques. In: Proceedings - 5th IEEE International Conference on Big Data Security on Cloud, BigDataSecurity 2019, 5th IEEE International Conference on High Performance and Smart Computing, HPSC 2019 and 4th IEEE International Conference on Intelligent Data and Security, IDS 2019, pp. 306–309 (2019). https://doi.org/10.1109/BigDataSecurity-HPSC-IDS.2019.00063
10. Park, H., Kim, P., Kim, H., Park, K.W., Lee, Y.: Efficient machine learning over encrypted data with non-interactive communication. Comput. Stan. Interf. **58**, 87–108 (2018)
11. Rifaut, A.: Office of inspector general, health care compliance association: guidance note: guidance on anonymisation and pseudonymisation. In: 2011 4th International Workshop on Requirements Engineering and Law, RELAW 2011, Proceedings - Held in Conjunction with the 19th International Requirements Engineering Conference (June), pp. 1–54 (2019). https://oig.hhs.gov/compliance/101/files/HCCA-OIG-Resource-Guide.pdf
12. Singapore, P.D.P.C.: Guide to basic data anonymisation techniques. In: Published 25 January 2018. Personal Data Protection Commission Singapore (PDPC) (January), pp. 1–39 (2018). https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Other-Guides/Guide-to-Anonymisation_v1-(250118).pdf
13. Sweeney, L.: k-anonymity: a model for protecting privacy. Int. J. Uncertain. Fuzziness Knowl. Based Syst. **10**(5), 557–570 (2002). https://doi.org/10.1142/S0218488502001648
14. Zhong, S., Yang, Z., Wright, R.N.: Anonymization of customer data, vol. 1, pp. 139–147 (2005)