An Approach to Trade-off Privacy and Classification Accuracy in Machine Learning Processes

Loredana Caruccio[1], Domenico Desiato[2], Giuseppe Polese[1], Genoveffa Tortora[1] and Nicola Zannone[3]

[1] *Department of Computer Science, University of Salerno, via Giovanni Paolo II n.132, 84084 Fisciano (SA), Italy*
[2] *Department of Computer Science, University of Bari Aldo Moro, via Edoardo Orabona n.4, 70125 Bari (BA), Italy*
[3] *Eindhoven University of Technology, Eindhoven, Netherlands*

Abstract

Machine learning techniques applied to large and distributed data archives might result in the disclosure of sensitive information. Data often contain sensitive identifiable information, and even if these are protected, the excessive processing capabilities of current machine learning techniques might facilitate the identification of individuals. This discussion paper presents a decision-support framework for data anonymization. The latter relies on a novel approach that exploits data correlations, expressed in terms of relaxed functional dependencies (RFDs), to identify data anonymization strategies for providing suitable trade-offs between privacy and data utility. It also permits to generate anonymization strategies leveraging multiple data correlations simultaneously to increase the utility of anonymized datasets. In addition, our framework provides support in the selection of the anonymization strategies by enabling an understanding of the trade-offs between privacy and data utility offered by the obtained strategies. Experiments on real-life datasets show that our approach achieves promising results in data utility while guaranteeing the desired privacy level. Additionally, it allows data owners to select anonymization strategies balancing their privacy and data utility requirements.

Keywords

Privacy preserving machine learning, k-anonymity, Relaxed functional dependencies, Generalization strategies

1. Introduction

The increasing amounts of data available together with the advances in information technology have brought several benefits and opened new opportunities for the industry, individuals, and society. In particular, Big Data analytics has enabled the development of increasingly sophisticated applications ranging from personalized medicine and e-commerce to crowd management and fraud detection [1]. However, these applications have also introduced new privacy and ethical challenges [2]. Big Data typically holds large amounts of personally identifiable information (e.g., criminal records, shopping habits, credit and medical history, and driving records), which can enable mass surveillance and profiling programs and raise several privacy issues [3].

SEBD 2023: 31st Symposium on Advanced Database System, July 02–05, 2023, Galzignano Terme, Padua, Italy

[†] These authors contributed equally.

lcaruccio@unisa.it (L. Caruccio); domenico.desiato@uniba.it (D. Desiato); gpolese@unisa.it (G. Polese); tortora@unisa.it (G. Tortora); n.zannone@tue.nl (N. Zannone)
0000-0002-2418-1606 (L. Caruccio); 0000-0002-6327-459X (D. Desiato); 0000-0002-8496-2658 (G. Polese); 0000-0003-4765-8371 (G. Tortora); 0000-0002-9081-5996 (N. Zannone)

To prevent these issues arising, data protection and privacy frameworks usually define strict requirements on the collection and processing of personally identifiable information. For instance, the General Data Protection Regulation (GDPR)[1] requires organizations to collect, process, and share personal data only for legitimate and lawful purposes, and to periodically identify privacy risks that can affect the data subjects.

Employing all the measures and procedures for the protection of personally identifiable information, as required by data protection regulations and, especially, by the GDPR, can be expensive for organizations. Thus, many organizations need to ensure that the personal data they collect for data analytics are sufficiently anonymized to reduce the associated compliance burdens [4, 5].[2] To this end, they often eliminate any unique identifier for each user when collecting personal data. However, this in itself may not solve the problem, since removing unique identifiers might not be sufficient to guarantee data anonymity [6]. In fact, anonymized data could be de-anonymized through cross-referencing with data gathered from other sources [7]. Moreover, the application of machine learning techniques to anonymized data might still lead to the disclosure of sensitive and confidential information about data subjects, thanks to the power of current predictive models. On the other hand, we might still want to enable machine learning and data analytics processes to extract useful knowledge and insights from data while avoiding the disclosure of sensitive information. Thus, the challenge is to devise anonymization techniques that do not allow re-identification of individuals by using machine learning techniques on anonymized data [6].

Related work. Several techniques relying on cryptography, randomization, and perturbation, have been proposed to anonymize data in data sharing and analytics contexts [8, 9]. In our proposal, we focus on anonymization techniques based on generalization. The latter consists of replacing attribute values with more generalized ones to make the records in a dataset indistinguishable from each other [10, 11]. When generalizing data to protect the privacy of individual records, some information is lost, which can impact the data utility for further analysis [7, 6]. Current solutions usually suggest approaches to meet anonymity requirements while minimizing loss of information or trade-off between privacy and data utility requirements [12, 13, 14]. However, generalization strategies often fail to consider correlations in the data, resulting in excessive penalties to data utility.

Contribution. In this discussion paper, we describe the data anonymization framework proposed in [15]. In detail, such framework exploits (multiple) data correlations, represented as relaxed functional dependencies (RFDs) [16], in order to define generalization strategies that guarantee the required level of privacy and supports the entity responsible for the anonymization of the data (e.g., the data owner) in balancing privacy and data utility requirements.

The remainder of the paper is organized as follows. Section 2 presents the problem statement and Section 3 describes the proposed approach. Section 4 presents experimental results, whereas Section 5 concludes the paper and provides directions for future work.

[1] GDPR - Final version URL: http://data.consilium.1125europa.eu/doc/document/ST-5419-2016-INIT/en/pdf.

[2] Notice that the principles of the GDPR do not apply to anonymized information, i.e., information from which the data subject is no longer identifiable.

	age	workclass	fnlwgt	education	maritial-status	occupation	relationship	sex	capital-gain	classes
t_1	39	State-gov	77516	Bachelors	Never-married	Adm-clerical	Not-in-family	Male	2174	>50K
t_2	50	Self-emp-not-inc	83311	Bachelors	Married-civ-spouse	Exec-managerial	Husband	Male	0	>50K
t_3	38	Private	215646	HS-grad	Divorced	Handlers-cleaners	Not-in-family	Male	0	<=50K
t_4	53	Private	234721	11th	Married-civ-spouse	Handlers-cleaners	Husband	Male	0	<=50K
t_5	37	Private	159449	Bachelors	Married-civ-spouse	Prof-specialty	Wife	Female	0	>50K
t_6	37	Private	284582	Masters	Married-civ-spouse	Exec-managerial	Wife	Female	0	<=50K
t_7	49	Private	160187	9th	Married-spouse-absent	Other-service	Not-in-family	Female	0	>50K
t_8	52	Self-emp-not-inc	209642	HS-grad	Married-civ-spouse	Exec-managerial	Husband	Male	0	<=50K
t_9	38	Private	45781	Masters	Never-married	Prof-specialty	Not-in-family	Female	14084	>50K
t_{10}	49	Private	159449	Bachelors	Married-civ-spouse	Exec-managerial	Husband	Male	5178	>50K

Table 1
A sample dataset containing users' information.

2. Problem statement

Classification models capture correlations between the attributes of individuals and a class value, and are often used to predict the class value for any unseen new observation. Classification models are built from a training dataset, which might contain sensitive information. This information could be inferred from the classification model by exploiting the correlations encoded in the model [17]. To this end, training data are usually anonymized by removing identifiable information before the classifier is trained. However, data can still be re-identified using quasi-identifiers [6].

Example 1. Let us consider the sample dataset in Table 1, which is extracted from the Adult dataset[3]. Each tuple describes an individual, where `age`, `workclass`, `fnlwgt`, `education`, `maritial-status`, `occupation`, `relationship`, `sex`, and `capital gain` are attributes characterizing her, whereas attribute `classes` indicates whether her annual income is greater or lower than $50K$. From this sample dataset it is possible to narrow down tuple t_1 to a specific individual by looking, for instance, at the `age` attribute, as this is the only tuple for which `age` is equal to 39.

This simple example shows that only removing identifiable information from a dataset might not be sufficient to guarantee anonymization. Anonymized data can be re-identified by linking the data by means of other data sources [18]. An anonymization model largely used for this purpose is k-anonymity [19], which requires that at least k individuals in the dataset share the same set of attribute values.

In this work, we propose a novel anonymization technique that uses generalization and k-anonymity validation to anonymize a dataset while minimizing the loss of data utility. To this end, we exploit data correlations in the dataset, expressed in terms of relaxed functional dependencies (RFDS), as a guideline to define suitable generalization strategies.

3. A decision-support framework for data anonymization

This section presents a decision-support framework for data anonymization we propose in [15]. In detail, given an input dataset and a taxonomy of its quasi-identifiers, we first extract generalization rules expressed in terms of RFDS and use them to determine which attributes should be generalized and at which level. To assess the quality of a generalization rule, we first apply it to the input dataset to replace attribute values with more general ones, and then compute the anonymity level and the data utility for the resulting generalized dataset. In a second step, we extend the coverage of the RFDS that satisfy a given level of anonymity by joining generalization rules to increase data utility. The data anonymization and utility provided by the obtained extended RFDS are then

[3] https://www.openml.org/d/179

assessed as in the previous step. The obtained generalization rules provide data owners with a view of which generalization rules can be used to anonymize their datasets and their effects in terms of data utility and anonymization.

3.1. Generalization rule extraction

The first phase of our approach aims to extract generalization rules in terms of RFDS and to determine the level of anonymity and data utility they achieve when applied on a dataset. RFDS are extracted from the input dataset, along with the generalization levels (defined with respect to the given attribute taxonomies), by integrating the semantics of roll-up dependencies [20].

Definition 1. (Roll-up dependency). Let G be a genschema of a relation schema R and $X, Y \subseteq attr(R)$, a roll-up dependency (RUD) $X_{\Phi_1} \rightarrow Y_{\Phi_2}$ is valid on an instance r of R, if and only if for each tuple pair (t_1, t_2) of r, if $\Pi_X(t_1)$ and $\Pi_X(t_2)$ are α-equivalent, then also $\Pi_Y(t_1)$ and $\Pi_Y(t_2)$ must be α-equivalent, where t_1, t_2 are said to be α-equivalent iff they become equal after rolling up their attribute values at most as many levels as the ones specified by α.

During RFD extraction, we only consider RFDS having the classification attribute (i.e., attribute `classes` in the example dataset of Table 1) on the right-hand side, with generalization level equal to 0. This is because we are interested in the generation of anonymized datasets that can be used to train a classification model. Accordingly, our focus is on correlations involving the classification attribute and preserving its original values.

Example 2. Suppose that the following RFD is extracted from the dataset of Table 1: $\texttt{age}_{\leq 3}$, $\texttt{fnlwgt}_{\leq 2} \rightarrow \texttt{classes}_{\leq 0}$. The right-hand side of the RFD contains the classification attribute `classes`, whereas the left-hand side contains the subset of attributes `age` and `fnlwgt` to be generalized. The generalization level is defined by the values after the tag "$\leq$".

Moreover, data owners are left with the task to determine which generalization rules should be used for the anonymization of their datasets. This can be a complex task, as a large number of RFDS can be potentially extracted from the dataset itself [21, 4], and not all of them might satisfy the desired level of anonymity. In addition, RFDS usually capture basic correlations in the data, involving a limited number of attributes and, thus, limiting the data utility that can be achieved from their application. Increasing the number of attributes on the left-hand side of an RFD will make it possible to involve more attributes in the anonymization of the dataset, and thus, increase its data utility [7]. However, the use of more attributes could reduce the level of anonymity guaranteed by the generalization rules. Therefore, the data utility can be improved only where, and to the extent that, the minimum level of anonymity required by the data owner is satisfied.

3.2. Generalization rule selection and improvement

This phase of the approach aims to generate a set of candidate generalization rules from the RFDS derived in the previous phase of the approach (cf. Section 3.1), which satisfy at least a given level of anonymity and, at the same time, limit the data utility loss due to the anonymization process.

RFDS may not guarantee a level of anonymity that is acceptable for the data owner. In particular, the data owner might define minimum anonymization requirements for a dataset to be shared with other parties. According to the k-anonymity model, we model these requirements as a user-defined threshold t, indicating the minimum anonymity level that the dataset should satisfy in order to be considered for sharing. We use the threshold t to determine whether an RFD provides a sufficient level of anonymity. To check if an RFD is suitable for anonymization, the RFD is applied to the original dataset and the anonymity level k of the obtained anonymized dataset is computed using the k-anonymity model. If the anonymity level k of the obtained anonymized dataset is equal or greater than the user-defined threshold t, then the RFD satisfies the minimum anonymization requirements, and it is considered in the anonymization process; otherwise, the RFD is discarded.

The RFDS capture only basic correlations in the data, hence limiting the data utility that can be achieved through their application. To this end, we analyze the attributes involved in the RFDS and define a coverage strategy to increase the number of selected attributes to be used for the anonymization of the dataset. Our strategy compares the RFDS and determines which ones can be combined to improve data utility. The intuition is that joining RFDS allows to account for multiple data correlations simultaneously, hence increasing the number of attributes that can be used. Since combined RFDS have to be valid on the considered dataset, not all RFDS can be combined.

We introduce the notion of *compatible RFDS*, which specifies when two RFDS can be joined. Intuitively, two RFDS are compatible if and only if their left-hand side attributes are disjoint or occur with the same generalization level, as formalized in Definition 2.

Definition 2 (RFD Compatibility)**.** Let $X_\Phi \to C_{\leq 0}$ and $X'_{\Phi'} \to C_{\leq 0}$ be two RFDS such that $X = \{A_1, \dots, A_n\}$, $X' = \{B_1, \dots, B_m\}$, and each attribute A_i (B_j) is associated with a generalization level ϕ_i (ϕ'_j) in Φ (Φ'). We say that the two RFDS are compatible if and only if:

- $X \cap X' = \emptyset$, or
- $\forall A_i \in X$ and $B_j \in X'$, such that $A_i = B_j \in X \cap X'$, then $\phi_i = \phi'_j$.

Notice that, in our approach we consider possible generalization rules according by guaranteeing the following constraints: *(i)* a set of RFDS can be combined into a new RFD if and only if each RFD is compatible with the others; and *(ii)* if the combination of two RFDS does not meet the minimum anonymity level, any combination of RFDS that includes those RFDS will not satisfy the minimum anonymity level, and hence, it will be discarded. Identifying the optimal candidate rules can be seen as a multi-objective optimization problem and, thus, we use the notion of Pareto-optimality and Pareto frontier [22] to guide the data owner in the selection of suitable generalization rules.

4. Experiments

We performed a number of experiments to evaluate the approach proposed in Section 3. In particular, we studied whether joining RFDS and, thus, accounting for a larger set of attributes, result in anonymization strategies that allow to obtain anonymized datasets with higher data utility. Moreover, we investigated the trade-off between anonymization and data utility that can be achieved by using generalization rules and how to devise strategies for selecting the generalization rules to be used for data anonymization. More specifically, our experiments were driven by the following research questions:

RQ1: What is the impact of combining generalization rules on data utility?

RQ2: Which trade-off between privacy and data utility can be achieved by using generalization rules?

RQ3: How much effort is required by a data owner to identify the generalization rule to apply?

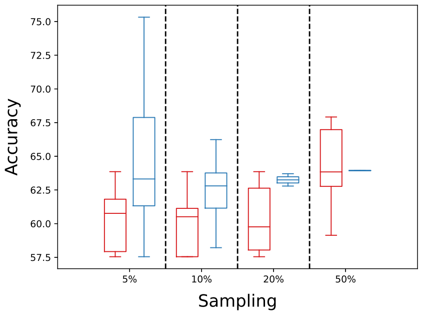

Figure 1: Variation of the number of RFDS at the increase of the anonymity level.

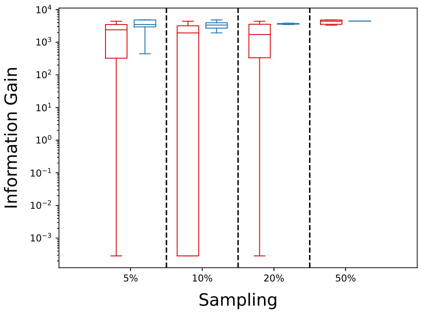

Figure 2: Variation of the number of RFDS at the increase of the anonymity level.

The first research question (**RQ1**) aims to test our hypothesis and provide insights on the impact that combined generalization rules produce on the data utility. **RQ2** aims to assess the trade-off between anonymization and data utility that can be achieved using generalization rules. **RQ3** aims to evaluate the effort required to a data owner to determine the generalization rule to apply for the anonymization of her dataset, in terms of the number of rules returned by our approach.

4.1. Results

In this section, we present the results of experiments and answer our research questions. In particular, to determine the anonymity level offered by a generalization rule, we apply the generalization rule to the original dataset and compute the minimum number of tuples in the generalized dataset that are indistinguishable with respect to the quasi-identifiers, representing the k-anonymity level that the generalized dataset can guarantee by applying such a generalization rule. Moreover, data utility is measured in terms of classification accuracy and information gain. In the experiment, classification accuracy was computed using both the decision tree and the Support Vector Machine classifiers, whereas information gain was computed by using the decision tree classifier only. The remainder of this section discusses only a portion of the obtained results, whose complete version can be found in [15].

RQ1: What is the impact of combining generalization rules on data utility? This research question aims to evaluate the benefits of combining generalization rules, represented through RFDS, to generate strategies for data anonymization, which maximize data utility while guaranteeing a desired level of privacy. We expected that, on average, the combination of RFDS provides generalization rules with higher data utility compared to those directly extracted from the data. To measure this, we compare such sets of rules in terms of classification accuracy and information gain. In the analysis, we consider all generalization rules that achieve an anonymity level of at least 2 (i.e., $k \geq 2$). Figures 1 show the accuracy that can be achieved using the generalization rules directly extracted from the data (red boxes) and using the combined rules (blue boxes) at the varying of sampling percentage for the Electricity dataset[4]. We can observe that, combining generalization rules improves the accuracy (obtained using the ID3 decision tree classifier) for all sampling percentages, except for the 50% sampling percentage for the Electricity dataset. This is because many generalization rules extracted for this sampling percentage contain the same attributes with different generalization levels and, thus, they are incompatible or their combination

[4] https://datahub.io/machine-learning/electricity

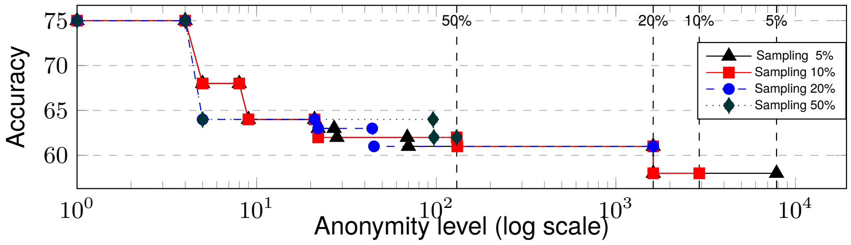

Figure 3: Trade-off between privacy and accuracy (ID3).

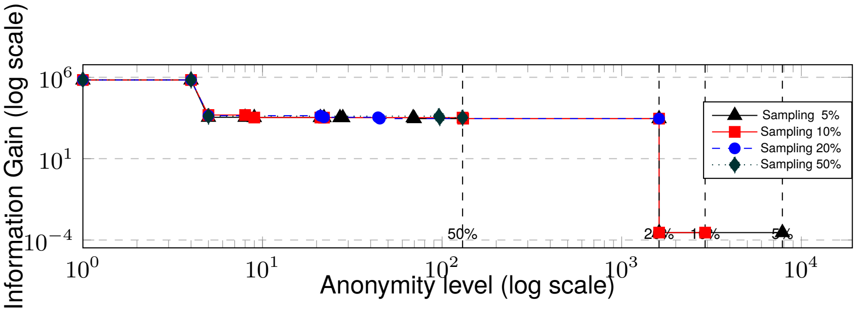

Figure 4: Trade-off between privacy and information gain.

violated the privacy requirement over k (i.e., $k < 2$). Our experiments also show that combining generalization rules improves information gain for the Electricity dataset, as illustrated in Figures 2.

RQ2: Which trade-off between privacy and data utility can be achieved using generalization rules? We expect that data utility decreases when the anonymity level increases. This is because achieving a higher level of anonymity requires higher generalization levels, leading to less specificity of data. To understand which trade-off between privacy and data utility can be achieved, we quantify these effects by showing how accuracy and information gain vary when the anonymity level increases. Figures 3 show the trade-off between accuracy and anonymity level for the Electricity dataset. The x-axis reports the anonymity levels (in log scale), whereas the y-axis reports the best accuracy that can be achieved by applying the generalization rules that satisfy a given anonymity level. The baseline accuracy is obtained over the non-anonymized version of the datasets. Each vertical dashed line in the plots represents the maximum anonymity level that can be achieved using a given sampling percentage (5%, 10%, 20%, and 50%). As expected, we can observe that, for the Electricity dataset, the accuracy decreases when the anonymity level increases, and that the highest anonymity level is achieved for the 5% sampling. Figure 4 shows the trade-off between information gain and anonymity level for the Electricity dataset. Similarly to the results obtained for accuracy, information gain decreases when the anonymity level increases, and the highest anonymity level is achieved for the 5% sampling.

RQ3: How much effort is required by a data owner to identify the generalization rule to apply? A large number of generalization rules can be potentially returned by our approach, leaving the data owner with the burden to identify which generalization rule should be applied. To assist the data owner in this task, we employed an approach based on Pareto-optimality to identify those rules providing a suitable trade-off between privacy and data utility. Next, we evaluate such

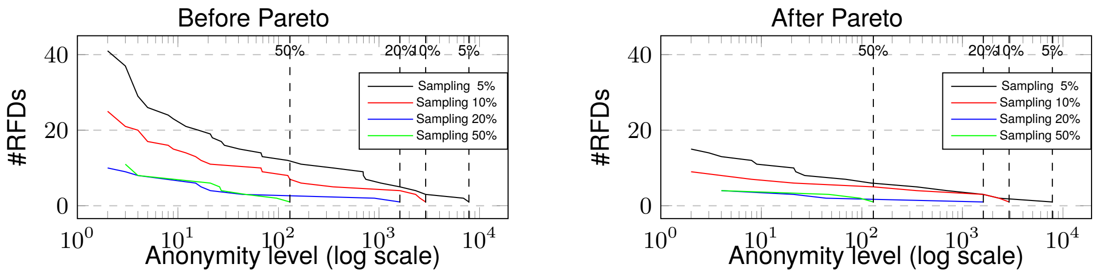

Figure 5: Variation of the number of RFDs at the increase of the anonymity level.

approach and, in general, the effort required to a data owner to determine the generalization rule to apply, in terms of the number of rules returned by our approach. Figure 5 reports the total number of RFDs obtained using our approach at the increase of the anonymity level for each sampling percentage, before (left plot) and after (right plot) the application of Pareto-optimality, for the Electricity dataset. We can observe that the sampling percentage has a large impact on the number of rules, the use of lower sampling percentages typically results in a larger number of generalization rules. An in-depth analysis (not reported here for lack of space) shows that the number of combined generalization rules is also higher for lower sampling percentages. This is mainly due to the fact that generalization rules obtained for lower sampling percentages typically involve few attributes, yielding many possibilities to combine them with each other. The results also show that the application of Pareto-optimality significantly reduces the number of generalization rules to be considered by data owners when anonymizing their datasets. For example, the use of Pareto-optimality yields a reduction of the total number of generalization rules, which achieve at least an anonymity level equal to 2, between 60% and 63% for the Electricity dataset where the largest reduction is obtained for the 5% sampling. When deriving rules using low sampling percentages, Pareto-optimality tends to preserve more combined rules than rules directly extracted from the data, whereas this consideration is reversed when the sampling percentage increases.

5. Conclusion

This work presents a decision-support framework for data anonymization with application to machine learning processes. The approach extracts RFDs from the data to define possible generalization rules and combine them to derive anonymization strategies guaranteeing a higher data utility. Pareto-optimality is then employed to identify those generalization rules that provide optimal trade-offs between privacy and data utility. Results show that the proposed approach enables a data owner to identify effective anonymization strategies.

In the future, we plan to investigate the application of other data utility and privacy metrics and study their impact on the trade-off between anonymization and data utility, as well as the applicability of our approach to other data sharing contexts [23].

Acknowledgments

This Publication was produced with the co-funding of the European union - Next Generation EU: NRRP Initiative, Mission 4, Component 2, Investment 1.3 – Partnerships extended to universities, research centers, companies and research D.D. MUR n. 341 del 5.03.2022 – Next Generation EU (PE0000014 - "Security and Rights In the CyberSpace - SERICS" - CUP: H93C22000620001).

References

[1] Y. J. Meijaard, B. C. M. Cappers, J. G. M. Mengerink, N. Zannone, Predictive analytics to prevent voice over IP international revenue sharing fraud, in: Data and Applications Security and Privacy XXXIV, volume 12122 of *LNCS*, Springer, 2020, pp. 241–260.

[2] S. Rathore, P. K. Sharma, V. Loia, Y.-S. Jeong, J. H. Park, Social network security: Issues, challenges, threats, and solutions, Information sciences 421 (2017) 43–69.

[3] L. Caruccio, O. Piazza, G. Polese, G. Tortora, Secure IoT analytics for fast deterioration detection in emergency rooms, IEEE Access 8 (2020) 215343–215354.

[4] L. Caruccio, D. Desiato, G. Polese, G. Tortora, GDPR compliant information confidentiality preservation in big data processing, IEEE Access 8 (2020) 205034–205050.

[5] D. Desiato, G. Tortora, A methodology for gdpr compliant data processing., in: SEBD, 2018.

[6] A. Zigomitros, F. Casino, A. Solanas, C. Patsakis, A survey on privacy properties for data publishing of relational data, IEEE Access 8 (2020) 51071–51099.

[7] C. Ni, L. S. Cang, P. Gope, G. Min, Data anonymization evaluation for big data and IoT environment, Information Sciences 605 (2022) 381–392.

[8] J. Li, X. Kuang, S. Lin, X. Ma, Y. Tang, Privacy preservation for machine learning training and classification based on homomorphic encryption schemes, Information Sciences 526 (2020) 166–179.

[9] M. I. Pramanik, R. Y. Lau, M. S. Hossain, M. M. Rahoman, S. K. Debnath, M. G. Rashed, M. Z. Uddin, Privacy preserving big data analytics: A critical analysis of state-of-the-art, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 11 (2021) e1387.

[10] L. Sweeney, Achieving k-anonymity privacy protection using generalization and suppression, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10 (2002) 571–588.

[11] F. Ashkouti, K. Khamforoosh, A. Sheikhahmadi, DI-Mondrian: Distributed improved mondrian for satisfaction of the l-diversity privacy model using apache spark, Information Sciences 546 (2021) 1–24.

[12] T. K. Esmeel, M. M. Hasan, M. N. Kabir, A. Firdaus, Balancing data utility versus information loss in data-privacy protection using k-anonymity, in: Conference on Systems, Process and Control, IEEE, 2020, pp. 158–161.

[13] R. Wang, Y. Zhu, C.-C. Chang, Q. Peng, Privacy-preserving high-dimensional data publishing for classification, Computers & Security 93 (2020) 101785.

[14] K. LeFevre, D. J. DeWitt, R. Ramakrishnan, Workload-aware anonymization, in: Proceedings of SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2006, pp. 277–286.

[15] L. Caruccio, D. Desiato, G. Polese, G. Tortora, N. Zannone, A decision-support framework for data anonymization with application to machine learning processes, Information Sciences 613 (2022) 1–32.

[16] L. Caruccio, V. Deufemia, F. Naumann, G. Polese, Discovering relaxed functional dependencies based on multi-attribute dominance, IEEE Transactions on Knowledge and Data Engineering 33 (2021) 3212–3228.

[17] A. Majeed, S. Lee, Anonymization techniques for privacy preserving data publishing: A

comprehensive survey, IEEE Access 9 (2021) 8512–8545.

[18] H. Goldstein, N. Shlomo, A probabilistic procedure for anonymisation, for assessing the risk of re-identification and for the analysis of perturbed data sets, Journal of Official Statistics 36 (2020) 89–115.

[19] P. Samarati, L. Sweeney, Generalizing data to provide anonymity when disclosing information, in: Symposium on Principles of Database Systems, ACM, 1998, p. 188.

[20] T. Calders, R. T. Ng, J. Wijsen, Searching for dependencies at multiple abstraction levels, ACM Transactions Database Systems 27 (2002) 229–260.

[21] L. Caruccio, V. Deufemia, G. Polese, Mining relaxed functional dependencies from data, Data Mining and Knowledge Discovery 34 (2020) 443–477.

[22] S. Petchrompo, D. W. Coit, A. Brintrup, A. Wannakrairot, A. K. Parlikad, A review of pareto pruning methods for multi-objective optimization, Computers & Industrial Engineering 167 (2022) 108022.

[23] J. Feng, L. T. Yang, N. J. Gati, X. Xie, B. S. Gavuna, Privacy-preserving computation in cyber-physical-social systems: A survey of the state-of-the-art and perspectives, Information Sciences 527 (2020) 341–355.