Fig. 1. Example of a typical generalization hierarchy
taken from [1]

Interactive Anonymization for Privacy aware Machine Learning

Bernd Malle[1 2], Peter Kieseberg[1 2], Andreas Holzinger[1]

[1] Holzinger Group HCI-KDD
Institute for Medical Informatics, Statistics & Documentation
Medical University Graz, Austria
b.malle@hci-kdd.org
[2] SBA Research gGmbH, Favoritenstrae 16, 1040 Wien
PKieseberg@sba-research.org

Abstract. **Keywords**: Machine Learning, Privacy aware ML, interactive ML, Knowledge Bases, Anonymization, k-Anonymity, SaNGreeA, Information Loss, Weight Vectors

1 Introduction and Motivation

2 Privacy aware Machine Learning

3 Methods of providing privacy

3.1 Perturbation

3.2 ϵ differential privacy

3.3 k-anonymity (and beyond)

Figure **??** illustrates the original tabular concept of anonymization: Given an input table with several columns, we will in all probability encounter three different categories of data:

- **Personal identifiers** are data items which directly identify a person without having to cross-reference or further analyze them. Examples are first and last names, but even more so an (email) address or social security number (SSN). As personal identifiers are dangerous and cannot be generalized (see Figure 1) in a meaningful way (e.g. one could generalize an email address by only retaining the mail provider fragment, but the result would not yield much usable information), this category of data is usually removed. The table shows this column in a red background color.

Fig. 2. Initial distribution of six selected data columns of the adult dataset.

- **Sensitive data,** also called 'payload', which is the kind of data we want to convey for statistics or research purposes. Examples for this category would be disease classification, drug intake or personal income level. This data shall be preserved in the anonymized dataset and can therefore not be deleted or generalized. The table shows this column in a green background color.
- **Quasi identifiers (QI's)**, colored in the table with an orange background, are data that in themselves do not directly reveal the identity of a person, but might be used in aggregate to reconstruct it. For instance, [3] mentioned that 87% of U.S. citizens in 2002 had reported characteristics that made them vulnerable to identification based on just the 3 attributes *zip code*, *gender* and *date of birth*. But although this data can be harmful in that respect, it might also hold vital information for the purpose of research (e.g. zip code could be of high value in a study on disease spread). The the actual point of all anonymization efforts is therefore to generalize this kind of information, which means to lower its level of granularity. As an example, one could generalize the ZIP codes 41074, 41075 and 41099 to an umbrella version 410**, as shown in Figure **??**.

4 interactive Machine Learning

5 Experiments

The following sections will describe our series of experiments in detail, encompassing the data source selected, the algorithm used as well as a description of the overall process employed to obtain our results.

5.1 Data

As input data we chose the adults dataset from the UCI Machine Learning repository which was generated from US census data of 1994 and contains approximately 32,000 entries; from those 30,162 were selected after preprocessing. Of the attributes (data columns) provided only one was deleted because it was also represented by a column containing its numerical mapping (education => education_num). Figure 2 shows the attribute value distribution of the original input dataset with the exception of the sample weights.

As one can see, there are several attributes with one value clearly dominating the others; *native-country* being the most prominent example with the entry for the United States dwarfing all other countries (which comes as no surprise given the data origin). As anonymization generalizes different countries together if necessary, it was interesting for the author to see how these distributions

Fig. 3. Anonymized distribution of six selected data columns of the adult dataset, anonymization factor of k=19, equal weight for each attribute.

would change under a relatively large k-factor. Figure 3 shows the same attribute distribution with its values anonymized by a factor of $k = 19$. Although the dominance of the United states was successfully "broken" by this method, in several instances the *generalized-to-all*-value (*) now skews the data set even more. Apart from the expected generalization information loss this is another reason why one would assume worse results from a machine learning classifier applied to an anonymized dataset.

5.2 Algorithm

SaNGreeA stands for *Social network greedy clustering* and was introduced by [2]. In addition to 'clustering' nodes of a graph according to the minimum general information loss (GIL) incurred as described in Section **??**, this algorithm also considers the structural information loss (SIL) incurred in assigning a node to a certain cluster. The SIL quantifies the probability of error when trying to reconstruct the structure of the initial graph from its anonymized version.

$$\mathrm{GIL}(cl) = |cl| \cdot \Big(\sum_{j=1}^{s} \frac{size(gen(cl)[N_j])}{size(min_{x \epsilon N}(X[N_j]), max_{x \epsilon N}(X[N_j]))} + \sum_{j=1}^{t} \frac{height(\Lambda(gen(cl)[C_j]))}{height(H_{C_j})}\Big)$$

where:
- $|cl|$ denotes the cluster cl's cardinality;
- $size([i1, i2])$ is the size of the interval $[i1, i2]$, i.e., $(i2 - i1)$;
- $\Lambda(w), w \epsilon H_{C_j}$ is the sub-hierarchy of H_{C_j} rooted in w;
- $height(H_{C_j})$ denotes the height of the tree hierarchy H_{C_j};

The total generalization information loss is then given by:

$$\mathrm{GIL}(G, S) = \sum_{j=1}^{v} \mathrm{GIL}(cl_j)$$

And the normalized generalization information loss by:

$$\mathrm{NGIL}(G, S) = \frac{\mathrm{GIL}(G, S)}{n \cdot (s + t)}$$

Fig. 4. The attribute values of the adult dataset which contribute most positively / negatively to the classification result. The columns to the right strongly indicate a yearly income of above 50k, whereas the columns to the outer left indicate a yearly income of below 50k. The least significant columns in the middle part were cut out.

5.3 Process

To examine the impact of perturbation and anonymization of datasets on the quality of a classification result, we designed the following processing pipeline:

1. Taking the original (preprocessed) dataset as input, we transformed its attributes to boolean values, so instead of *native-country* $->$ *United-States* we considered *United-States* $->$ *yes / no*.
2. We then ran 4 different classifiers on it and computed precision, recall as well as F1 score. The four classifiers used were *gradient boosting*, *random forest*, *logistic regression* and *linear SVC*.
3. From the obtained results we extracted the 3 attribute values most contributing to a "positive" (>50k) result as well as the top 3 attribute values indicating a "negative" (<=50k) prediction as depicted in Figure 4
4. For each of these 6 attribute values, we subsequently deleted a specific percentage of data rows containing that value from the original dataset, resulting in 30 reduced datasets. The 5 percentages used were 0.2, 0.4, 0.6, 0.8 as well as 1.0.
5. To each of those datasets we re-applied the four chosen classifiers successively and recorded the respective impact on the quality of the classification result. The results can be seen in Figure **??** and Figure **??**.
6. In order to measure the effects of k-anonymization on classifier performance, we used the SaNGreeA's GIL component described in the following section to generate datasets with a k-factor of $k = 3$, $k = 7$, $k = 11$, $k = 15$ as well as $k = 19$. Furthermore, we used each of these settings with 3 different weight vectors: 1) equal weights for all attributes, 2) age information preferred ($\omega(age) = 0.88$, $\omega(other\_attributes) = 0.01$) and 3) race information preferred ($\omega(race) = 0.88$, $\omega(other\_attributes) = 0.01$). We then re-executed all classifiers on the resulting 15 datasets and recorded the respective results, which can be seen in Figure **??**.

6 Results & Discussion

7 Open problems & future challenges

– **Explain the unexpected behavior** for...

8 Conclusion

References

1. Gagan Aggarwal, Tomas Feder, Krishnaram Kenthapadi, Rajeev Motwani, Rina Panigrahy, Dilys Thomas, and An Zhu. Approximation algorithms for k-anonymity. *Journal of Privacy Technology (JOPT)*, 2005.
2. Alina Campan and Traian Marius Truta. Data and structural k-anonymity in social networks. In *Privacy, Security, and Trust in KDD*, pages 33–54. Springer, 2009.
3. Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.