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

Introduce randomness into the data.. can never be sure if specific answer is correct, nevertheless a certain fraction of ansers will statistically be correct.

Laplace (double exponential) distribution.. close to 0 - large perturbation - further out smaller ...

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.
- **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, [8] 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

Interactive ML algorithms adjust their inner workings by continuously interacting with an outside *oracle*, drawing positive / negative reinforcement from this interaction [**?**]. Such systems are especially useful for highly-personalized predictions or decision support [**?**]; moreover many real-world problems exhibit (super)exponential algorithmic runtime; in such cases human brains dwarf machines at approximating solutions and learning from very small samples, thus enabling us to 'intuit' solutions efficiently [**?**].

By incorporating humans as oracles into this process, we can elicit background knowledge regarding specific use cases unknown to automatic algorithms [9]. This however is highly dependent on the users' experience in a certain field as well as data / classification complexity; domain experts can of course be expected to contribute more valuable decision points than laymen; likewise, a low-dimensional dataset and simple classification tasks will result in higher quality human responses than convoluted problem sets.

While the authors of [7] propose a system that interacts with a user in order to set a certain k-factor and subsequently provides a report on information loss and Kurtosis of QI distributions, the algorithm is not *interactive* by our definition in

that it does not influence the inner workings of the algorithm during the learning phase. This is also true in case of the Cornell Anonymization Toolkit (Cat) [10], which conducts a complete anonymization run and only then lets the user decide if they are content with the results. In contrast, our approach alters algorithmic parameters upon every (batch of) human decisions, letting the algorithm adapt in real-time.

[6] describe an approach incorporating humans into the anonymization process by allowing them to set constraints on attribute generalization; moreover they construct generalization hierarchies involving domain-specific ontologies. Although this technique marks a departure from wholesale automatic anonymization, it still lacks the dynamic human-computer interaction of our approach.

Apart from the field of privacy, interactive ML is today present in a wide spectrum of applications, from bordering medical fields like protein interactions / clusterings [2] via on-demand group-creation in social networks [3] to even teaching algorithms suitable mappings from gestures to music-generating parameters [5].

5 Experiments

The following sections will describe our experiment 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 **??** shows the attribute value distribution of the original input dataset with the exception of the sample weights.

5.2 Algorithm

SaNGreeA stands for *Social network greedy clustering* and was introduced by [4]. 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.

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

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:

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

The main process...

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 2
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

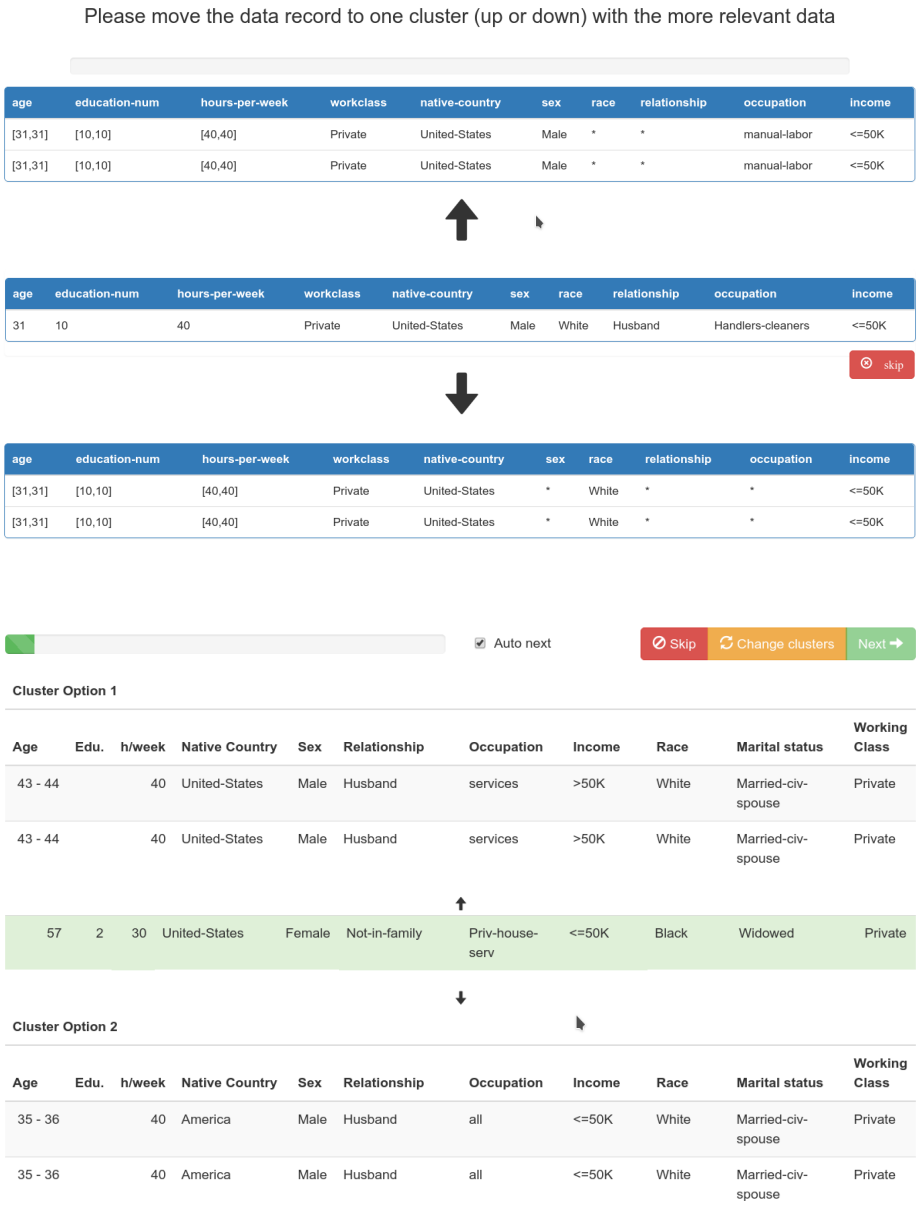
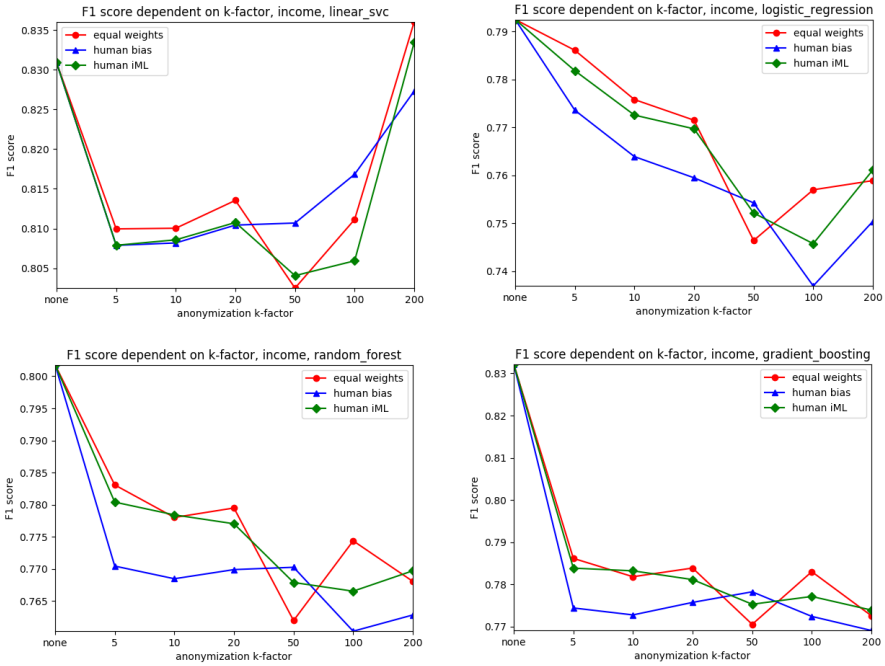


Fig. 2. Two implementations of the iML interface design...



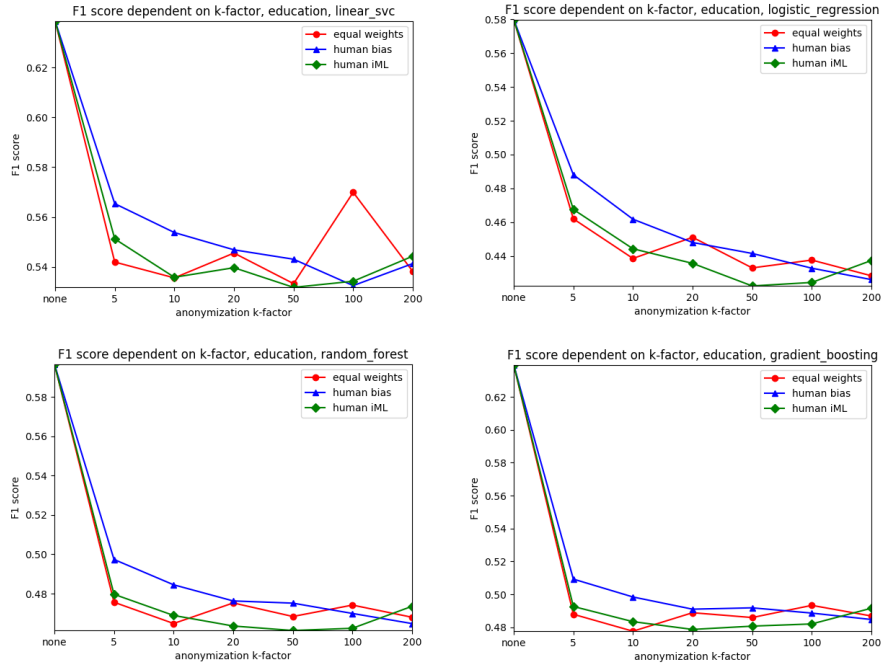


Fig. 4. Results on...

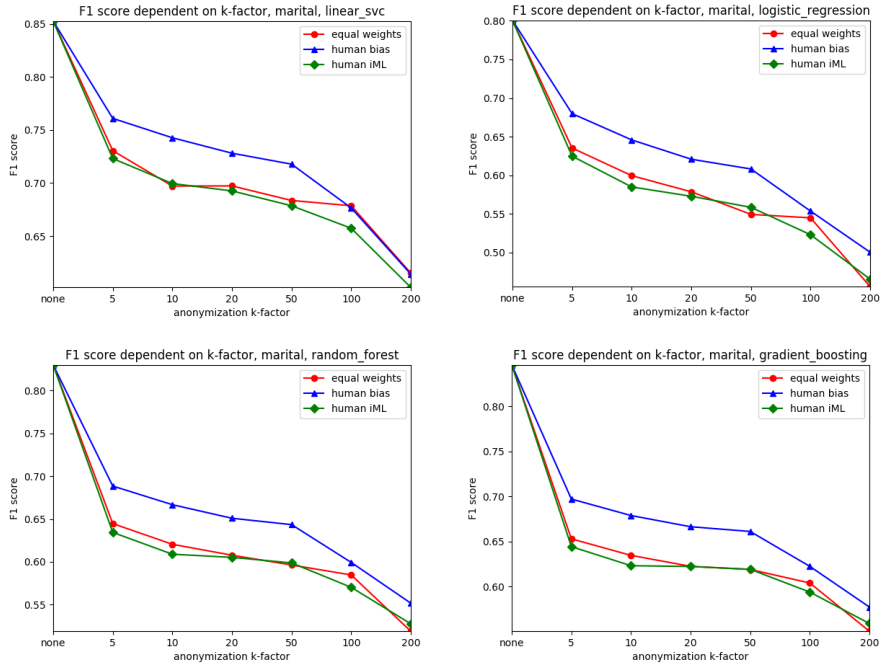


Fig. 5. Results on...

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. Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. Power to the People: The Role of Humans in Interactive Machine Learning. *AI Magazine*, 35(4):105–120, 2014.
3. Saleema Amershi, James Fogarty, and Daniel Weld. ReGroup: interactive machine learning for on-demand group creation in social networks. *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12*, page 21, 2012.
4. 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.
5. R. Fiebrink, D. Trueman, and P.R. Cook. A metainstrument for interactive, on-the-fly machine learning. *Proc. NIME*, 2:3, 2009.
6. Brian C.S. Loh and Patrick H.H. Then. Ontology-enhanced interactive anonymization in domain-driven data mining outsourcing. *Proceedings - 2nd International Symposium on Data, Privacy, and E-Commerce, ISDPE 2010*, (June):9–14, 2010.
7. Carlos Moque, Alexandra Pomares, and Rafael Gonzalez. AnonymousData.co: A Proposal for Interactive Anonymization of Electronic Medical Records. *Procedia Technology*, 5:743–752, 2012.
8. Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.
9. MALCOLM WARE, EIBE FRANK, GEOFFREY HOLMES, MARK HALL, and IAN H WITTEN. Interactive machine learning: letting users build classifiers. *International Journal of Human-Computer Studies*, 55(3):281–292, 2001.
10. Xiaokui Xiao, Guozhang Wang, and Johannes Gehrke. Interactive anonymization of sensitive data. *Proceedings of the 35th SIGMOD international conference on Management of data - SIGMOD '09*, page 1051, 2009.