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. Privacy aware Machine Learning is the discipline of applying Machine Learning techniques in such a way as to ensure the privacy of people during the process. This can most easily achieved by first anonymizing a dataset before releasing it for the purpose of statistical data mining or research; starting in June 2018, this will also be the only legally permitted way within the European Union to release data without granting any people involved the 'right to be forgotten', i.e. the right to have their data deleted on request. To governments, organizations and corporations however, this represents a serious impediment to research operations, since any anonymization results in a certain degree of information loss and therefore reduced data utility. Our work focuses on applying interactive Machine Learning to the process of anonymization based on the idea of eliciting human background knowledge to optimize an algorithms parameters as to which attributes are more or less valuable to preserve. We demonstrate that human input can yield measurably better classification results than automatic approaches, with much more room for future improvement.

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

1 Introduction and Motivation

based on our initial PAML experiments []

2 k-anonymity (and beyond)

Given the original tabular concept of anonymization, we will usually encounter three different categories of attributes within a given dataset:

- **Personal identifiers** are data items which directly identify a person without having to cross-reference or further analyze them. Examples are email address or social security number (SSN). As personal identifiers are immediately dangerous, this category of data is usually removed.

- **Sensitive data,** also called 'payload', represents information that is crucial for further data mining or research purposes. Examples for this category would be disease classification, drug intake or personal income. This data shall be preserved in the anonymized dataset and can therefore not be deleted or generalized.
- **Quasi identifiers (QI's)**, are data which in themselves do not directly reveal the identity of a person, but might be used in aggregate to reconstruct it. For instance, [15] reported in 2002 that the identity of 87% of U.S. citizens could be uncovered via just the 3 attributes *zip code*, *gender* and *date of birth*. Despite this danger, QI's may contain vital information to research applications (like ZIP code in a disease spread study); they are therefore generalized to an acceptable compromise between privacy (data loss) and information content (data utility).

Based on this categorization another formal concept of privacy was introduced as *k-anonymity* [13], in which a record is released only if its quasi-identifiers are indistinguishable from at least $k-1$ other entities in the dataset. This can be imagined like a clustering of data into so-called *equivalence classes* of at least size k, with all internal QI's being generalized to the exact same level.

This original requirement of *k-anonymity* [14] has since been extended by the concepts of *l-diversity* [10] (where every cluster must contain at least l diverse sensitive values), *t-closeness* [8] (demanding that the local distribution over sensitive values must not diverge from its global distribution by more than a factor of t) as well as *delta-presence* [12] (which incorporates the background knowledge of a potential attacker). Although all of those concepts are interesting in their own right, for the sake of comparing interactive ML algorithms to their fully automatic counterpart, we only took *k-anonymity* into consideration (at least for this work).

3 interactive Machine Learning

Interactive ML algorithms adjust their inner workings by continuously interacting with an outside *oracle*, drawing positive / negative reinforcement from this interaction [6]. Such systems are especially useful for highly-personalized predictions or decision support [7]; 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 [5].

By incorporating humans as oracles into this process, we can elicit background knowledge regarding specific use cases unknown to automatic algorithms [16]. 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 [11] 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) [17], 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.

[9] 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 [1] via on-demand group-creation in social networks [2] to even teaching algorithms suitable mappings from gestures to music-generating parameters [4].

4 Experiments

The following sections will describe our experiment in detail, encompassing the general iML setting, the chosen data set, anonymization algorithm used as well as a description of the overall pipeline employed to obtain the final results as presented.

4.1 General setting

The basic idea of our experiment was to compare different weight vectors representing attribute (quasi-identifier) importance during anonymization: Let's say that a doctor needs ro release a dataset for the purpose of studying disease-spread; in this case 'ZIP code' information is probably (but not necessarily) of much greater importance then 'occupation' or 'race'. However, if a skin cancer study is to be performed, 'race' information might be of utmost importance, whereas 'ZIP code' might be negligible.

In our experiment, the task was to classify a people dataset on the target attributes *income*, *education level* and *marital status*. Therefore, we tested an *equal* weight vector setting against two others obtained from human experiments: 1) *bias* in which the user just specified which attributes they thought would be important for a specific classification by moving sliders, and 2) *iML* in which the user was tasked to decide a series of clustering possibilities by moving a data row to one of two partly anonymized clusters presented, thereby conveying which attributes were more important to preserve than others Figure 1.

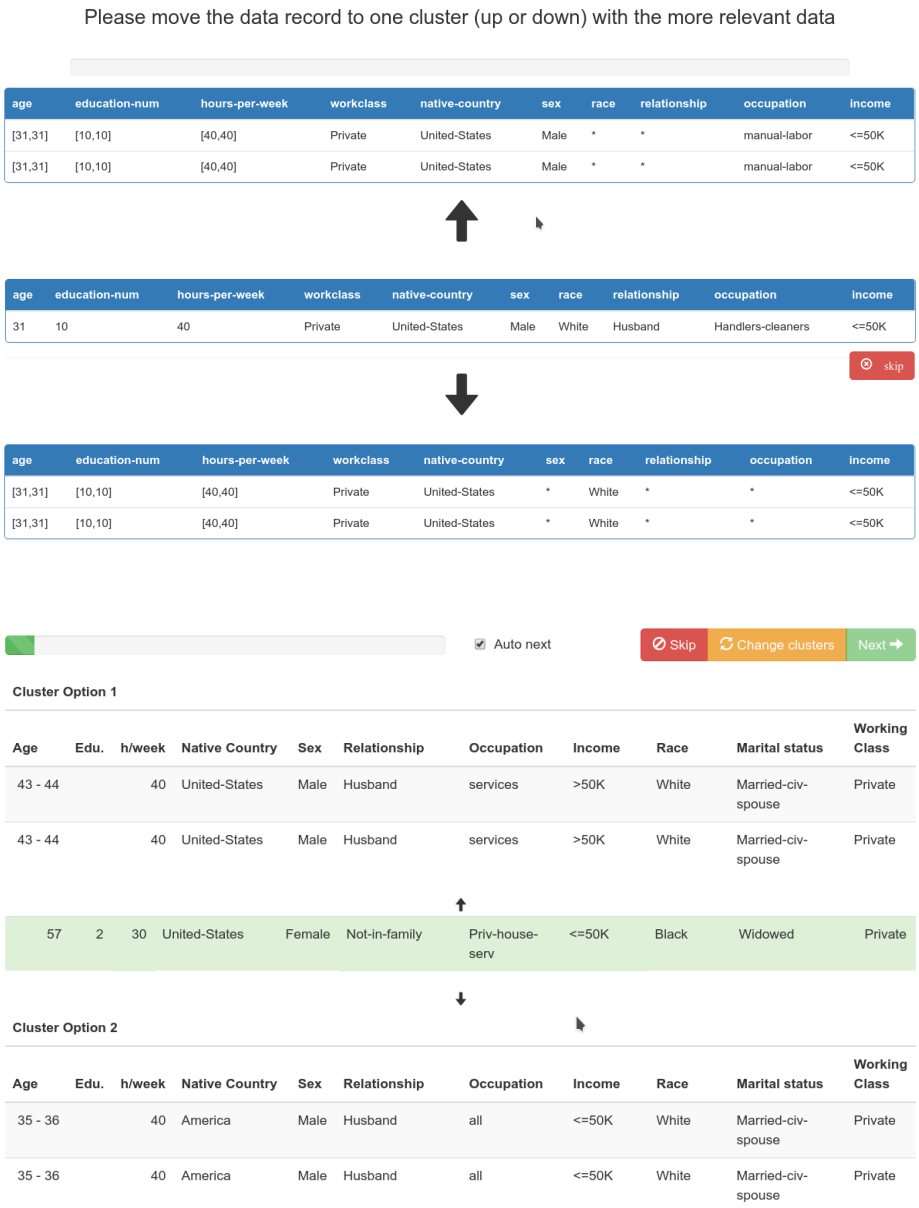

Fig. 1. Two different implementations of the iML interface design.

4.2 Data

We chose the adults dataset from the UCI Machine Learning repository which was generated from US census data from 1994 and contains approximately 50k entries in it's original; after initial preprocessing we chose the first 500 complete data rows as our iML experimental data to be presented to users via a

Web Browser UI. After obtaining bias / iml weights from the experiment, we chose the first 3k entries of the original data as the basis for producing 780 new, anonymized data sets. Although this might seem overly frugal on our part, we have asserted via random deletion of original data points that classifier performance remains stable for as little as 1.5k randomly selected rows. Of the original attributes (data columns) provided 4 were deleted: 'capital-gain' & 'capital-loss' (both were too skewed to be useful for humans), 'fnlwgt' (a mere weighting factor) as well as 'education' which was also represented by a column containing its numerical mapping ('education_num').

4.3 Algorithm

In order to conduct our experiments, it was necessary to choose an algorithm which would enable us to easily hook into its internal logic - we therefore chose a greedy clustering algorithm called *SaNGreeA* (Social network greedy clustering) which was introduced by [3] and implemented it in JavaScript. This enabled us to execute it within a browser environment during our iML experiments as well as server-side for batch-execution of all derived datasets later on.

Besides its capacity to anonymize graph structures (which we did not utilize during this work), it is a relatively simple algorithm considering *General information loss* - or GIL - during anonymization. This GIL can be interpreted by the sum of information loss occurring during generalization of continuous (range) as well as hierarchical attributes:

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

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:

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

The algorithm starts by picking a (random or pre-defined) data row as its first cluster, then iteratively picking best candidates for merging by minimizing GIL until the cluster reaches size k, at which point a new data point is chosen as the initiator for the next cluster; this process continues until all data points are merged into clusters, satisfying the k-anonymity criterion for the given dataset.

4.4 Processing pipeline for obtaining results

Once our iML experiments had yielded enough weight vectors, we had to generate a whole new set of anonymized datasets on which we subsequently applied 4 classifiers on each of the 3 target attributes (columns) described; therefore we designed the following processing pipeline:

1. Taking the first 5k rows of the original, preprocessed dataset as input and applying k-anonymization with a k-factor range of [5, 10, 20, 50, 100, 200] and 129 different weight vectors (equal, bias, iml) from our experiments on it, we produced 774 anonymized datasets (775 including the original).

5 Results & Discussion

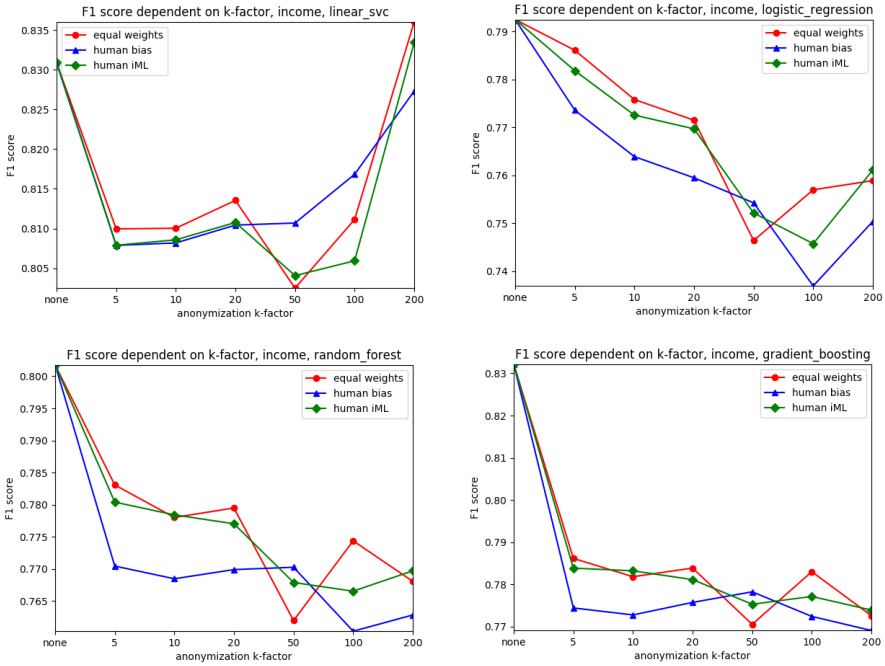

Fig. 2. Results on...

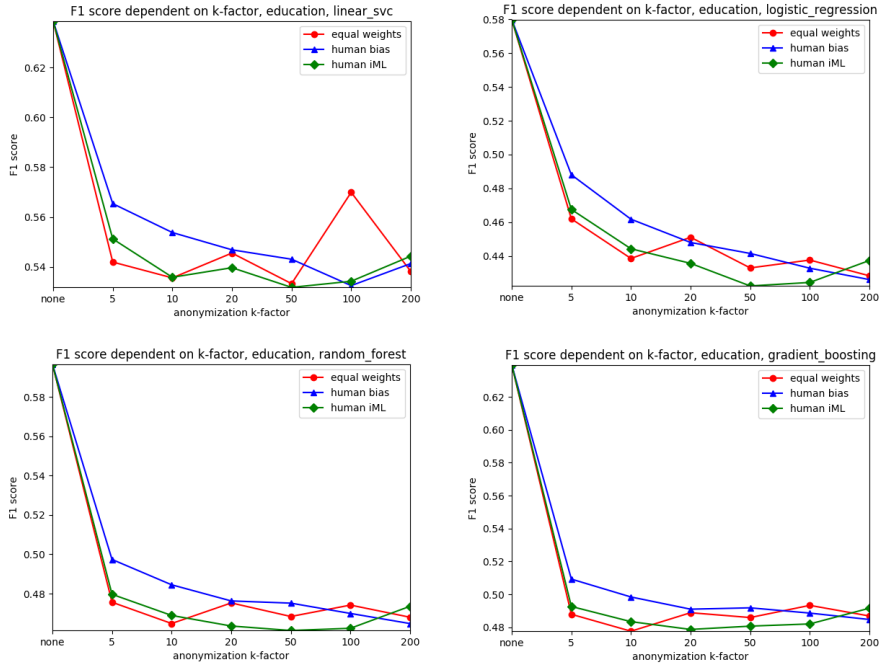

Fig. 3. Results on...

6 Open problems & future challenges

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

7 Conclusion

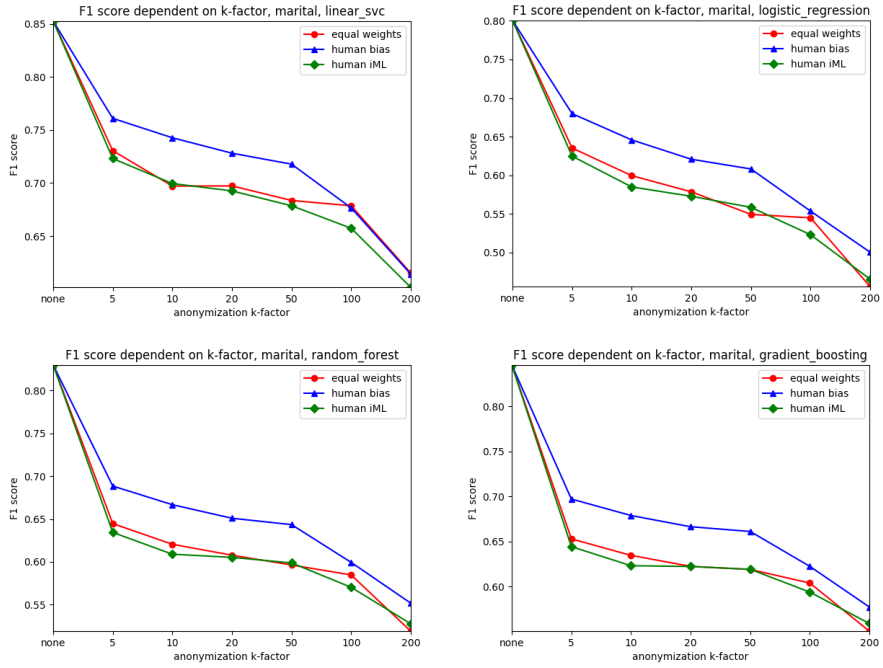

Fig. 4. Results on...

References

1. 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.
2. 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.
3. 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.
4. R. Fiebrink, D. Trueman, and P.R. Cook. A metainstrument for interactive, on-the-fly machine learning. *Proc. NIME*, 2:3, 2009.
5. A Holzinger, M Plass, K Holzinger, GC Crisan, CM Pintea, and V Palade. Towards interactive machine learning (iml): Applying ant colony algorithms to solve the traveling salesman problem with the human-in-the-loop approach. In *IFIP International Cross Domain Conference and Workshop (CD-ARES)*, pages 81–95. Springer, Heidelberg, Berlin, New York, 2016.
6. Andreas Holzinger. Interactive machine learning for health informatics: When do we need the human-in-the-loop? *Springer Brain Informatics (BRIN)*, 3(2):119–131, 2016.
7. Peter Kieseberg, Bernd Malle, Peter Frhwirt, Edgar Weippl, and Andreas Holzinger. A tamper-proof audit and control system for the doctor in the loop. *Brain Informatics*, pages 1–11, 2016.
8. Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *IEEE 23rd International Conference on Data Engineering, ICDE 2007*, pages 106–115. IEEE, 2007.
9. 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.
10. Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):1–52, 2007.
11. Carlos Moque, Alexandra Pomares, and Rafael Gonzalez. AnonymousData.co: A Proposal for Interactive Anonymization of Electronic Medical Records. *Procedia Technology*, 5:743–752, 2012.
12. M. E. Nergiz and C. Clifton. delta-presence without complete world knowledge. *IEEE Transactions on Knowledge and Data Engineering*, 22(6):868–883, 2010.
13. Pierangela Samarati. Protecting respondents identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering*, 13(6):1010–1027, 2001.
14. Latanya Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):571–588, 2002.
15. Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.
16. 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.

17. 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.