DO NOT DISTURB ?
Classifier behavior on perturbed datasets

No Author Given

No Institute Given

Abstract. Exponential trends in data generation are presenting todays organizations, economies and governments with challenges never encountered before, especially in the field of privacy and data security. One crucial trade-off regulators are facing regards the simultaneous need for publishing personal information for the sake of statistical analysis and Machine Learning in order to increase quality levels in areas like medical services, while at the same time protecting the identity of individuals. A key European measure will be the introduction of the General Data Protection Regulation (GDPR) in 2018, giving customers the 'right to be forgotten', i.e. having their data deleted on request. As this could lead to a competitive disadvantage for European companies, it is important to understand which effects deletion of significant data points has on the performance of ML techniques. In a previous paper we introduced a series of experiments applying different algorithms to a binary classification problem under anonymization as well as perturbation. In this paper we extend those experiments by multi-class classification and introduce outlier-removal as an additional scenario. While the results of our previous work were mostly in-line with our expectations, our current experiments revealed unexpected behavior over a range of different scenarios. A surprising conclusion of those experiments is the fact that classification on an anonymized dataset with outliers removed in beforehand can almost compete with classification on the original, un-anonymized dataset. This could soon lead to competitive Machine Learning pipelines on anonymized datasets for real-world usage in the marketplace.

Keywords: Machine learning, knowledge bases, right to be forgotten, perturbation, k-anonymity, SaNGreeA, information loss, cost weighing vector, multi-class classification, outlier analysis, variance-sensitive analysis

1 Introduction and Motivation for Research

2 K-Anonymity and Information loss

While there several data-structures which can contain and convey personal information we might want to protect (free text, audio, images, graph structures etc.) we are focusing our work on tabular data, since most unstructured documents of sensitive nature today can be mapped to tabular data and since delicate

information is most easily extracted from those. Figure 1 illustrates the original tabular concept of three different categories of data we will encounter in such tables:

- **Identifiers** directly reveal the identity of a person without having further analysis of the data. Examples are first and last names, email address or social security number (SSN). As personal identifiers are hard to generalized (see Figure 2) in a meaningful way (truncating an email address to 'host' would not yield much usable information), those columns are usually removed. The figure displays this column in a red background color.
- **Sensitive data,** or 'payload', is crucial information for statisticians or researchers and can therefore not be erased or perturbed; such data usually remains untarnished within the released dataset. The table shows one column in green background color representing such data.
- **Quasi identifiers (QI's)**, colored in the table with an orange background, do not directly identify a person (age=35), but can be used in combination to restrict possibilities to such a degree that a specific identity follows logically. For instance, [9] mentioned that 87% of U.S. citizens in 2002 could be re-identified by just using the 3 attributes *zip code*, *gender* and *date of birth*. On the other hand, this information might hold significant information for the purpose of research (e.g. zip code could be of high value in a study on disease spread). Therefore we generalize this kind of information, which means to lower its level of granularity. As an example, one could generalize grades from A+ to B- into A's and B's and then further up to encompass 'all' (also denoted as '*'), as shown in Figure 2.

Name	Age	Zip	Gender	Disease
Alex	25	41076	Male	Allergies
...	...	...	...	...

Fig. 1. The three types of data considered in (k-)anonymization

3 Related Work

A comparison of different Machine Learning algorithms on anonymized datasets was already conducted in 2014 [10] by applying 6 different algorithms on 3 datasets, with very diverse results per algorithm. The main weakness of this paper is its usage of extremely differently-sized datasets which does not easily allow comparison; moreover they only used one very low privacy setting of $k = 2$,

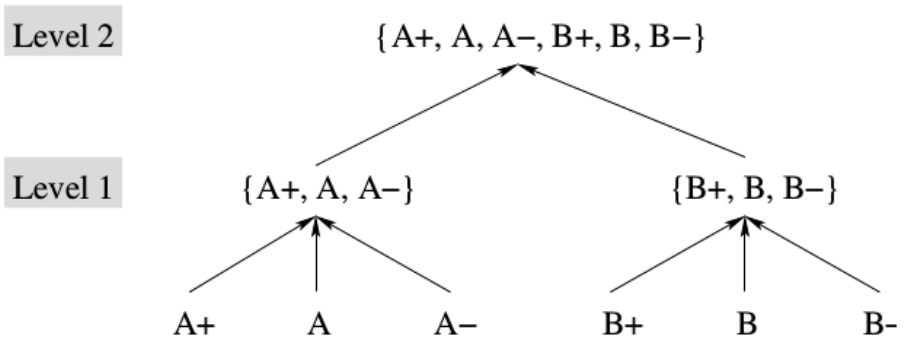

Figure 1: A possible generalization hierarchy for the attribute "Quality".

Fig. 2. Example of a typical generalization hierarchy taken from [1]

Node	Name	Age	Zip	Gender	Disease
X1	Alex	25	41076	Male	Allergies
X2	Bob	25	41075	Male	Allergies
X3	Charlie	27	41076	Male	Allergies
X4	Dave	32	41099	Male	Diabetes
X5	Eva	27	41074	Female	Flu
X6	Dana	36	41099	Female	Gastritis
X7	George	30	41099	Male	Brain Tumor
X8	Lucas	28	41099	Male	Lung Cancer
X9	Laura	33	41075	Female	Alzheimer

Node	Age	Zip	Gender	Disease
X1	25-27	4107*	Male	Allergies
X2	25-27	4107*	Male	Allergies
X3	25-27	4107*	Male	Allergies
X4	30-36	41099	*	Diabetes
X5	27-33	410**	*	Flu
X6	30-36	41099	*	Gastritis
X7	30-36	41099	*	Brain Tumor
X8	27-33	410**	*	Lung Cancer
X9	27-33	410**	*	Alzheimer

Fig. 3. Tabular anonymization: input table and anonymization result

preventing the authors from examining more interesting behavior as information content degrades further; this is a main point of our work.

The authors of [6] propose a scheme for controlling over-generalization of less identity-vulnerable QIs in diverse classes by determining the importance of QIs via Random Forest pre-computations as well as computing sensitive attribute diversity via the Simpson index [8]. Their resulting adaptive anonymization algorithm was compared to Mondrian [4] as well as IACk [5] and shows improvements w.r.t information loss as well as coverage (the number of descendant leaf nodes of generalized values in the taxonomy). Accuracy measured on classification tree, random forest and SVM shows equal or better performance when applied to a dataset anonymized by their proposed solution; it is interesting to note that their performance on large factors of k not only remains stable, but in some cases increases with k, the same behavior we also observed in some of our experiments.

The authors of a recent paper [3] propose the introduction of an additional requirement for anonymization on top of k-anonymity called h-ceiling, which simply restricts generalizations within an equivalence class to a certain level below suppression. In the case on an equivalence class being able to satisfy h-ceiling but not k-anonymity (their method applies full-domain generalization), counterfeit records are inserted into the respective group; each insertion is also collected in a journal which is eventually published with the anonymized data. Their approach unsurprisingly yields lower reconstruction error and information loss as well as more fine-grained query results due to less generalization. However, their experiments mostly fix $k = 5$ and therefore simply try to reduce information loss due to anonymization, but do not try to examine ML performance over a wider range of k factors; moreover, there seems to be some inconsistency in their predictions.

4 Experiments

The following sections will describe our series of experiments in detail, encompassing the dataset used, the algorithms chosen for classification as well as a description of the overall process employed to obtain our results.

4.1 Data

As input data we chose the training set of the adults dataset from the UCI Machine Learning repository which was generated from US census data and contains approximately 32,000 entries (30162 after deleting rows with incomplete information). All but one columns were considered for experimentation, the remaining representing duplicate information (education => education_num). Figure 4 shows the attribute value distribution of 6 arbitrarily selected columns of the original dataset.

4.2 Anonymization Algorithm

We implemented our own version of a greedy clustering algorithm called SaNGreeA (Social network greedy clustering, [2]) in JavaScript mainly for three reasons: 1) apart from 'normal' tabular anonymization tt has a network anonymization component based on stochastic reconstruction error, so it is possible for us to use this algorithm in later works regarding the impact of anonymization on graph algorithms; 2) we wanted a simple conceptual model so we could interact with the algorithm and thus conduct interactive Machine Learning experiments in the future (those experiments are well under way at the time of this writing); 3) we wanted an algorithm capable of running in the browser so we could run our experiments online especially w.r.t. 2). The main downside of this choice is the reduced algorithmic performance of $O(n^2)$ as well as a further slow-down for JS vs native code of a factor of about $3 - 4$. In the future, we will strive

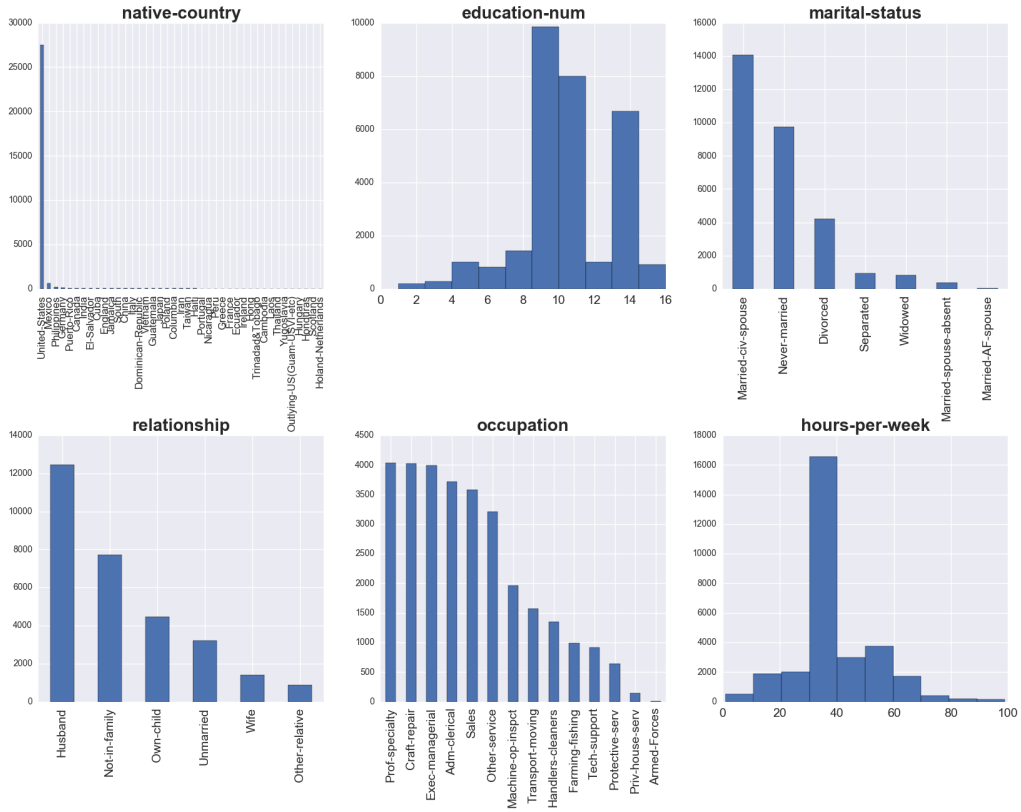

Fig. 4. Initial distribution of six selected data columns of the adult dataset.

to implement faster algorithms which nevertheless retain properties suitable for our needs, narrowing down the simplicity - performance trade-off.

As mentioned, SaNGreeA consists of two strategies for tabular as well as network anonymization, with two respective metrics for information loss. The *Generalization Information Loss* or *GIL* consists of a categorical as well as a continuous part, with the former measuring the distance of a level-of-generalization from it's original leaf node in the generalization hierarchy (taxonomy), while the latter measures the range of a continuous-valued generalization (e.g. age cohort [35-40]) divided by the whole range of the respective attribute (e.g. overall age-range [17-90]).

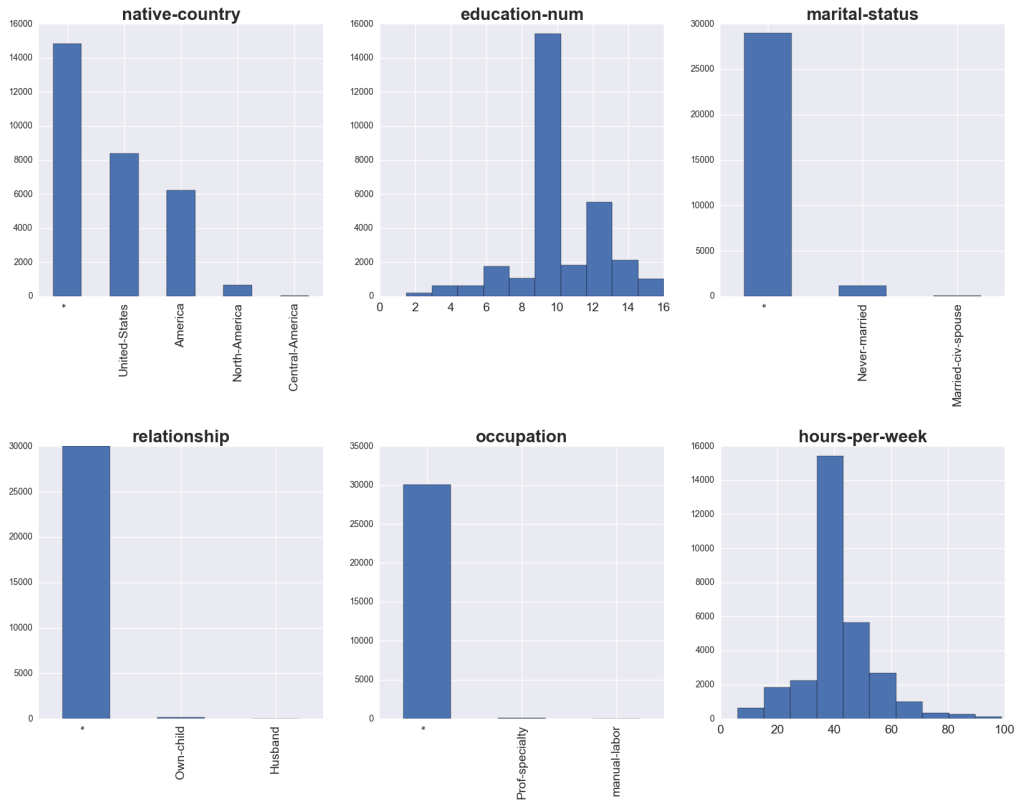

Fig. 5. Anonymized distribution of six selected data columns of the adult dataset, anonymization factor of k=19, equal weight for each attribute.

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

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:

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

As for the networking-part of this algorithm, it introduces a measure called *structural information loss* (SIL). The SIL is composed of two different components, which represent statistical errors of 1) intra-cluster as well as 2) inter-cluster reconstruction.

For the exact mathematical definitions of SIL & NSIL the reader is kindly referred to the original paper. Because the structural information loss cannot be computed exactly before the assembly of all clusters is completed, the exact computations were replaced by the following distance measures:

Distance between two nodes:

$$\mathrm{dist}(X^i, X^j) = \frac{|\{l | l = 1..n \wedge l \neq i,j; b_l^i \neq b_l^j|}{n-2}$$

Distance between a node and a cluster:

$$\mathrm{dist}(X, cl) = \frac{\sum_{X^j \epsilon cl} \mathrm{dist}(X, X^j)}{|cl|}$$

Since SaNGreeA follows the greedy-clustering paradigm, it runs in quadratic time w.r.t. the input size in number of nodes. This worked well within milliseconds for a problem size of a few hundred nodes, but took up to 60 minutes on the whole adult training dataset. Finally, as stated above, we chose SaNGreeA for its intuitive simplicity and graph anonymization capabilities, the latter of which are serving us well in a different branch of our ongoing research efforts; for the experiments in this paper, we restricted ourselves to the tabular anonymization capabilities of the algorithm.

4.3 Process

To examine the effect of perturbation, anonymization, outlier-removal as well as outlier-removal+anonymization on classifier performance, we designed the following processing pipeline:

5 Results & Discussion

5.1 Perturbed Datasets - Selective Deletion

In order to be able to compare the impact of selectively deleting the most / least important attribute values (in fact, the whole data points containing those

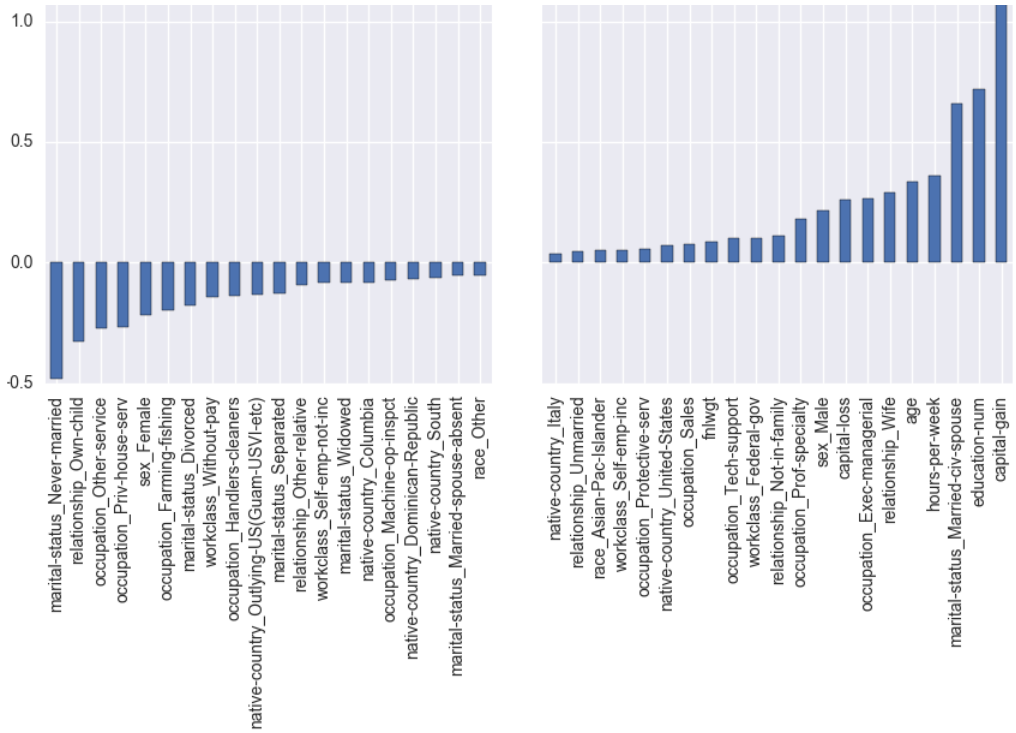

Fig. 6. Attribute values within the adult dataset which contribute highest / lowest certainty to the classification of income (truncated at 1.0). The rightmost columns represent information which enable a classifier to discern most clearly between classes, while the leftmost columns (depending on their actual score) could even confuse the algorithm. We chose this example because income is a binary decision, so the values don't change per category to predict.

values) on different classifiers, we chose to select these values via examining the logit coefficients produced during logistic regression. Although this possibly entails non-erasure of the values specifically significant for each classifier, we chose algorithmic comparison as the more insightful criterion; the implicit assumption that the same attribute values would influence all classifiers approximately equally was largely confirmed by our results.

In contrast to binary classification, determining the 'right' values to delete for a multi-class problem is not always possible: Values contributing highly to the decision boundary for one class might be less significant in the case of another - accordingly one would expect inconclusive behavior in the case of a target for which the highest / lowest log coefficients do not line up over class boundaries.

For each of the targets 'marital-status' and 'education-num' we measured those interesting coefficients in the hope of improving / degrading algorithmic

performance; that means deletion of highest logit's is supposed to remove certainty from an algorithm and decreasing performance, while deletion of lowest logit's is supposed to remove uncertainty, thus improving performance. Our analysis showed that while 'marital-status' had mainly the same most / least significant logit's across all classes, the attribute values for 'education-num' were rather diverse in this area.

In the latter case this lead to erratic behavior of the resulting performance curves, as can be seen in (Figure 7). It is interesting to note that 'income_>50k' obviously held much larger significance for Logistic Regression than for the other classifiers, as their results showed f1 score improvement with this particular value eviscerating.

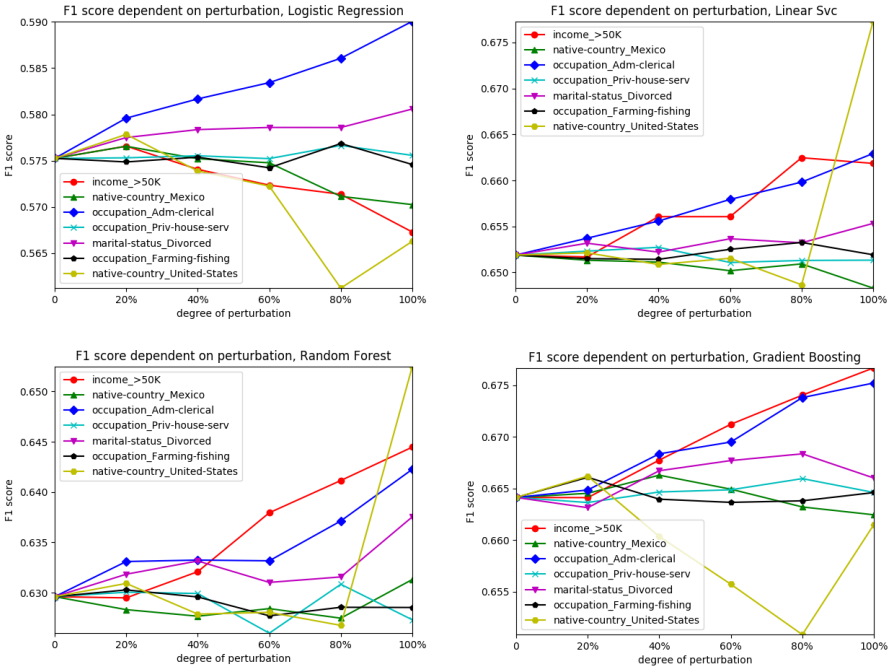

Fig. 7. Multi-class classification on target *education-num* under perturbation by selective deletion of the most / least contributing attribute values. Since different values are significant for deciding on different classes of education level, progressive deletion of this data results in indeterminate behavior.

In the case of 'marital-status' almost the same attribute values were rated as most / least significant across all classes - this results in very clear outputs with the erasure of highly important values decreasing performance drastically while deletion of confusing values leading to a significant increase in classifier performance (Figure 8). While it is not surprising that relationship information

shows high correlation with marital status, the opposite effects of *sex_Female* and *sex_Male* stand out as a slight curiosity - being a woman in this dataset seems to point less distinctly to a specific marital status than being a man.

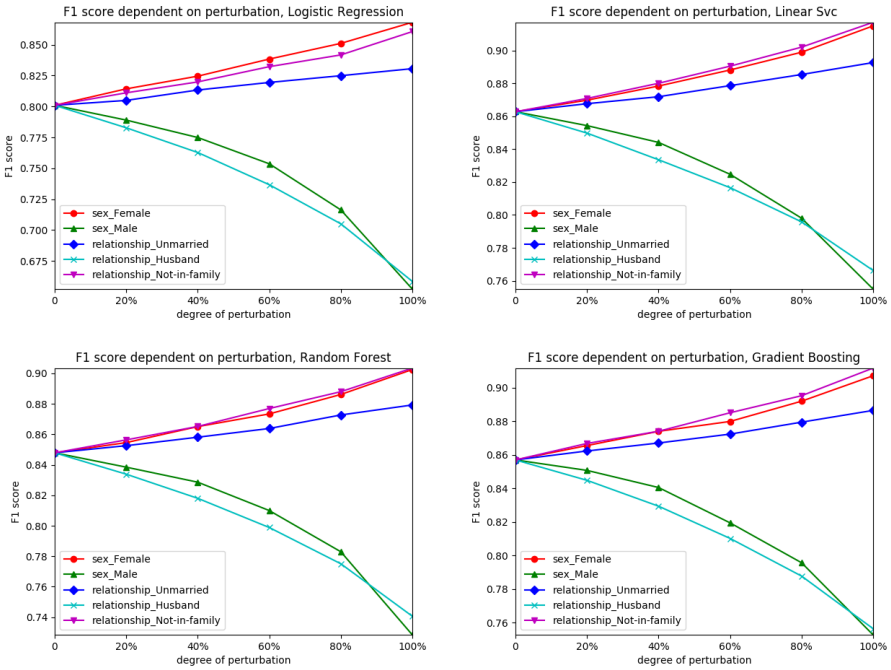

Fig. 8. Multi-class classification on target *marital-status* under perturbation by selective deletion of the most / least contributing attribute values. Since the same values are significant for deciding different classes of marital status, progressive deletion leads to orderly increase / decrease of ML performance.

5.2 Anonymized Datasets

Analogous to our previous work [7] we performed anonymization on the adult dataset for a range of values of k, but this time extending the range to $k \in \{3, 7, 11, 15, 19, 23, 27, 31, 35, 100\}$ for a broader observational basis of algorithmic behavior, especially towards higher values of k, as already conducted by other researchers [6], [3]. As we set out to examine multi-class classification performance, we chose the 'marital-status' and 'education-num' columns of the adult dataset as targets, treating income as an independent input feature. For 'marital-status' we left the 7 categorical values in the original dataset unchanged, whereas we clustered the 16 continuous 'education-num' levels into the 4 groups

'elementary school', 'high school including graduate', 'college up to Bachelors' as well as 'advanced studies'.

Our observation generally show the same type of behavior than in our previous experiments on target *income*, with one notable exception: The Random Forest classifier shows a sharp drop in algorithmic performance when operating on the very skewed 'age' and 'race' feature vectors, only to recover its discriminative power and increase in performance up to a k of 100. We also note a somewhat similar behavior for Logistic Regression, albeit not as distinctly. A possible explanation for this behavior could lie in the *bagging*-nature of Random Forest, meaning that the algorithm bootstraps by randomly sampling data-points from the overall population into possibly overlapping bags of 'local' data. As larger swaths of the input data become more and more equal with increasing levels of k, this would lead to less local over-fitting, thus making the job easier for a global averaging-strategy to filter out variance and improve generalization ability. However, if this was true, the maximum performance should not be recorded on the original (un-anonymized) dataset, thus we are currently at a loss of an adequate explanation for this specific case.

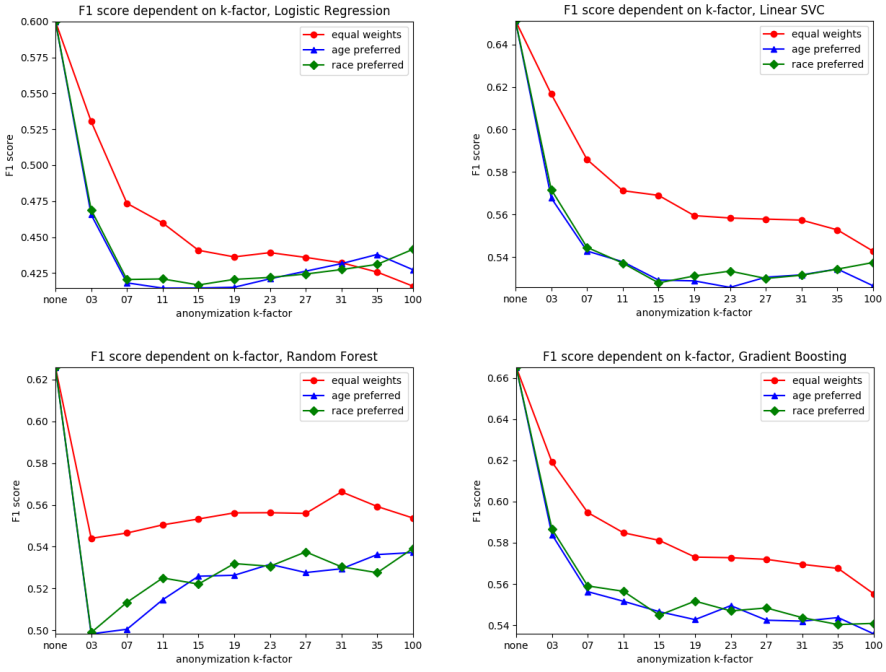

Fig. 9. Multi-class classification on target *education-num* on the adult dataset under several degrees of k-anonymization.

Classifier performance on target *marital-status* displayed the same basic behavior as above, including the mysterious conduct of the Random Forest in case of our age- and race-vectors. Moreover, the classification results are generally better than for *education-num*, which is probably caused by our somewhat arbitrary clustering of education levels during pre-processing. All in all, the pure anonymization-related results were almost in line with our expectations.

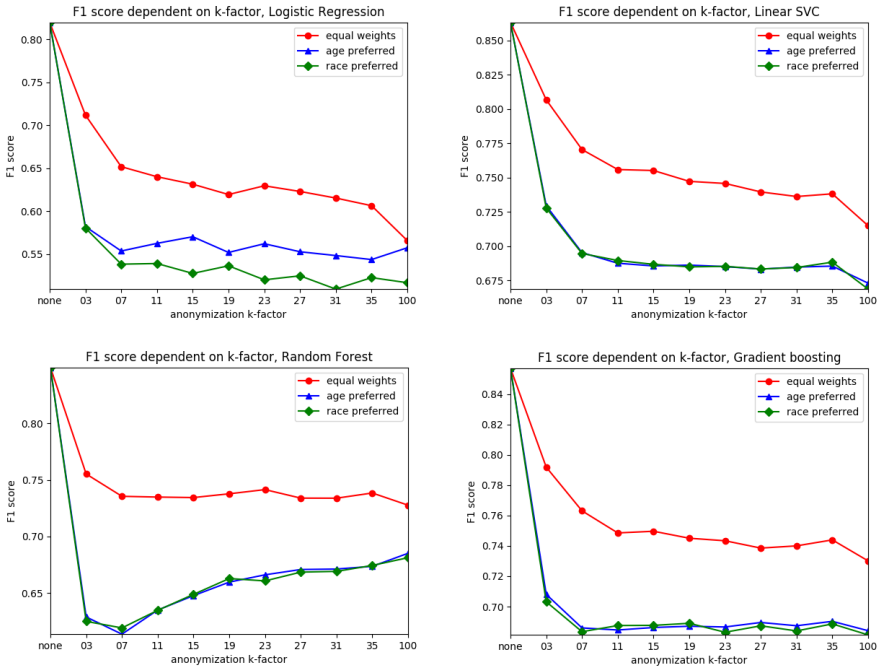

Fig. 10. Multi-class classification on target *marital-status* on the adult dataset under several degrees of k-anonymization.

5.3 "Outliers" removed

One question we didn't tackle in our previous work was the one of outlier removal; this is relevant due to the fact that e.g. people showing abnormal behavior could be supposed to exercise their 'right-to-be-forgotten' more frequently, especially in a social network scenario. For our experiments we chose the original adult dataset's income target, especially since we could thus directly compare the results with those of our previous work [7]. We used scikit-learn's Isolation-Forest classifier to identify outliers according to a given *contamination* level and performed an initial round of removing outliers in a range of $0.5\% - 5\%$.

Since ML performance decreased only marginally under those settings and we thus assumed that the dataset had been curated in such a way as to exclude significant outliers, we pivoted to a much broader investigation of examining classifier performance on a dataset with increasingly eviscerating variance. Thus we repeated the same procedure for "outlier" levels of $5\% - 95\%$, gradually diminishing the dataset's size from over 30k to about 1.5k data points. In order to account for that dramatic reduction, we compared classifier behavior with a control instance of the adult dataset with the same levels of truncation, but under random deletion of data points, thus not targeting variance in the control set.

The results are shown in Figure 11 and exhibit similar behavior to the removal of most-significant attribute values in our previous work: While performance only decreases slightly for deletion levels under 55%, we see a dramatic drop over the second half of the range. The obvious explanation for this behavior lie in the fact that more homogeneous clusters of data make it harder for any algorithm to construct a decision boundary - though it is noteworthy that this applies to all 4 classifiers the same despite their fundamentally different approaches. Lastly, the comparison set shows no significant increase / decrease of performance over the whole range of data deletion, supporting our conclusion that decreasing data set size was not the dominating influence for the observed algorithmic behavior.

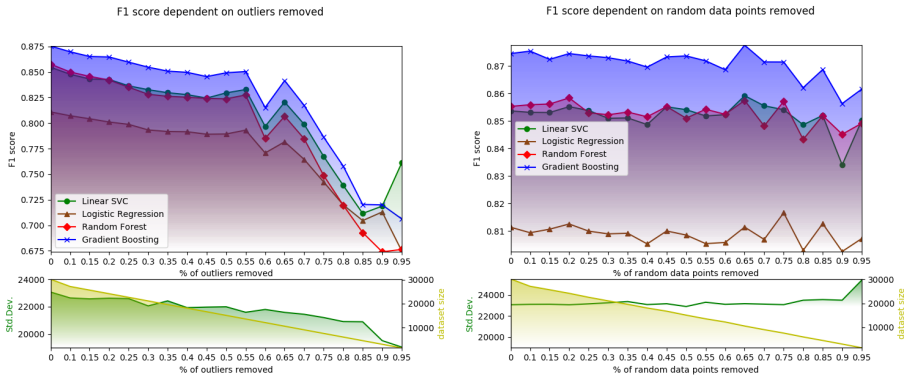

Fig. 11. Binary classification on target income based on a dataset with different degrees of outliers removed (= variance loss) vs. the same degree of data randomly deleted.

5.4 Anonymization on Outliers removed

One problem with outliers during anonymization is that it forces the algorithm to over-generalize attribute values; this can either happen towards the end-stages of a greedy-clustering procedure like SaNGreeA (in which case the damage might be limited to the outliers themselves), but could also influence a full-domain

generalizing algorithm during determination of a whole column's suitable generalization level (in which case the whole dataset would suffer significantly higher information loss). This fact in combination with our previously described results based on outlier removal gave rise to an interesting possibility: what if we *combined* outlier removal with anonymization? On the one hand classifier performance degrades with loss of variance, but for the very same reason information loss during anonymization might be limited to much more sufferable levels.

This led to our last round of experiments in which we took the adult dataset with 30% outliers removed and conducted k-anonymization as described in the respective earlier section (for time- and comparison reasons only on marital-status), the results of which can be seen in Figure 12. We were astonished to observe that - for the most part - classifiers performed better under this setting than under anonymization alone. For logistic regression, although age & race vectors performed worse then their anonymized-only counterparts, performance for equal weights was better for $k < 11$. With Random Forest, all vectors performed better than their anonymization-only counterparts, with $k = 3$ only 2% below original performance. With Linear SVC, age & race performed worse at the beginning only to recover with increasing performance towards $k = 100$, whereas the equal vector behaved about equal to it's non-outlier-removed opposite. Finally, Gradient Boosting in this setting outperforms it's anonymization-only competitor in all settings with it's $k = 3$ equal weight vector performance lying within only half a percentage point of the performance on the original, un-anonymized dataset.

As a side-note, we observe that under these settings, SVC starts to mimic Random Forest's behavior of an initial collapse in performance for the age- and range-vectors with a subsequent recovery towards higher levels of k. We do not yet have an adequate explanation for this and will investigate deeper in our future efforts.

Those amazing results raise a few burning questions: 1) Can we repeat that performance on real-world data? 2) Could we combine this technique with interactive Machine Learning / Anonymization which yield better weight vectors? 3) Do those advantages only hold for a toy algorithm or will they persist under more sophisticated Anonymization pipelines, 4) Can we further enhance those results by mixing synthetic data into the dataset, 5) Will better feature engineering compensate for our original drop in performance and thus moot our insight, and 6) can we apply this conclusion to other data structures like social networks. These points shall now briefly be discussed before concluding the paper.

6 Open problems Future challenges

7 Conclusion

We believe that this insight, in combination with work on interactive Anonymization we are currently conducting, state-of-the art anonymization techniques (we were using a rather simple algorithm for this paper), as well as the introduction

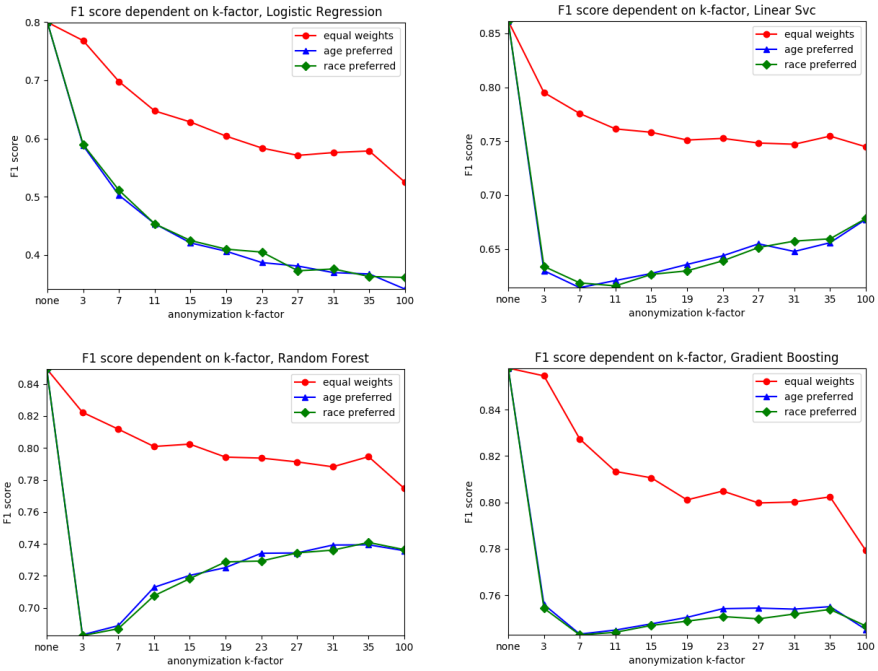

Fig. 12. Multi-class classification on target marital status based on a dataset with 30% outliers removed AND under different degrees of k-anonymization.

of synthetic data, will enable us to soon propose competitive Machine Learning pipelines for real-world usage to counterbalance any regulatory disadvantage for European companies on the marketplace.

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. Hyukki Lee, Soohyung Kim, Jong Wook Kim, and Yon Dohn Chung. Utility-preserving anonymization for health data publishing. *BMC Medical Informatics and Decision Making*, 2017.
4. Kristen LeFevre, David J DeWitt, and Raghu Ramakrishnan. Mondrian multidimensional k-anonymity. In *Data Engineering, 2006. ICDE'06. Proceedings of the 22nd International Conference on*, pages 25–25. IEEE, 2006.
5. Jiuyong Li, Jixue Liu, Muzammil Baig, and Raymond Chi-Wing Wong. Information based data anonymization for classification utility. *Data & Knowledge Engineering*, 70(12):1030–1045, 2011.

6. A Majeed, F Ullah, and S Lee. Vulnerability-and Diversity-Aware Anonymization of Personally Identifiable Information for Improving User Privacy and Utility of Publishing Data. *Sensors*, pages 1–23, 2017.
7. Bernd Malle, Peter Kieseberg, Edgar Weippl, and Andreas Holzinger. The right to be forgotten: towards machine learning on perturbed knowledge bases. In *International Conference on Availability, Reliability, and Security*, pages 251–266. Springer, 2016.
8. Edward H Simpson. Measurement of diversity. *Nature*, 1949.
9. Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.
10. Hayden Wimmer and Loreen Powell. A Comparison of the Effects of K-Anonymity on Machine Learning Algorithms. pages 1–9, 2014.