ENHANCING UTILITY AND PRIVACY WITH NOISY MINIMAX FILTERS

Jihun Hamm

The Ohio State University
Department of Computer Science and Engineering
Columbus, OH 43210, USA

ABSTRACT

Preserving privacy of continuous and/or high-dimensional data such as images, videos and audios is challenging. Syntactic anonymization methods were proposed typically for discrete data types and can be unsuitable. Differential privacy, which provides a stricter type of privacy, has shown more success in sanitizing continuous data. However, both syntactic and differential privacy are susceptible to inference attacks, i.e., an adversary can accurately guess sensitive attributes from insensitive attributes. On the other hand, minimax filters were proposed previously to minimize the accuracy of inference while maximizing utility at the same time. The paper presents *noisy minimax filter* that combines minimax filter and differentially private mechanism, which can attain high average utility and protection against inference attacks and a formal worst-case privacy guarantee. The proposed algorithm is demonstrated with real databases of faces, voices, and motion data.

Index Terms— syntactic anonymity, differential privacy, minimax optimization, postprocessing, machine learning

1. INTRODUCTION

Privacy is becoming an important issue in mining, processing, and learning from individuals' data. In response to growing privacy concerns, various privacy-preserving methods have been proposed by privacy researchers (e.g., see [1] for a review.) Earlier methods such as k-anonymity [2] and l-diversity [3] focussed on *syntactic anonymization* of sensitive attributes and quasi-identifiers in static databases. However, it was shown that syntactic anonymization is susceptible to several types of attacks such as the DeFinetti attack [4], in which an adversary is able to *infer* sensitive attributes of individuals accurately from insensitive, sanitized attributes. Another challenge for syntactic methods is anonymizing high-dimensional data. For example, k-anonymity is ineffective for high-dimensional sparse databases [5]. Furthermore, syntactic anonymization is proposed mainly for databases of discrete attributes, and can be unsuitable for protecting continuous-valued data, such as photo, video, audio, and biometric data. *Differential privacy* [6, 7, 8] was proposed to address many weaknesses of syntactic methods, and has since gained popularity due to its strong guarantees. However, similar to syntactic anonymization, differential privacy is not immune from inference attack [9], as it only prevents an adversary from gaining *additional* knowledge by inclusion/exclusion of a participant [10].

Recently, Hamm [11] proposed *minimax filters* which address main challenges discussed so far. The algorithm learns a linear or nonlinear filter from continuous and/or high-dimensional data and produces a dimensionality-reduced, 'filtered' representation of data. The filter allows useful information to pass through but prevents other information that can be used for inference attack from passing through.[1]. One disadvantage of minimax filters is that the privacy guarantee it provides is only in expectation (or in empirical average), which may be considered weaker than others such as differential privacy. Since minimax filter and differential privacy have almost independent goals and mechanisms to achieve them, it is natural to ask if the two methods can be combined.

This paper presents *noisy minimax filter*, which combines minimax filter with additive noise perturbation to satisfy the differential privacy criterion. Two methods of combination – preprocessing and postprocessing – are proposed (see Fig. 2.) In the preprocessing approach, a minimax filter is applied *before* perturbation to reduce the sensitivity of transformed data, so that the same level of differential privacy is achieved with less noise. However, it requires that the curator who trains the minimax filter is trusted by participants. In the postprocessing approach, a minimax filter is applied *after* perturbation. Since postprocessing cannot worsen differential privacy [10], this approach has the advantages of not requiring a trusted curator and no leakage of information through the released filter.

Experiments with real databases yield intuitive results. Differential privacy and resilience to inference are indeed different goals, such that using differentially private mechanism alone to achieve the latter requires a large amount of noise that destroys data utility. In contrast, minimax filters can suppress inference attack with little loss of utility. Noisy

[1] Originally, minimax filters were demonstrated for filtering out identity information, but they can be used for filtering out any sensitive attribute.

minimax filters, therefore, can provide a formal differential privacy in the worst case, and high on-average utility and protection against inference attacks.

2. NOISY MINIMAX FILTER

2.1. Minimax filter

Minimax filter [11] is a (deterministic) non-invertible transformation of input signal such that the transformed data has an optimal utility-privacy tradeoff. Formally, let $\mathcal{X} \subset \mathbb{R}^D$ be the space of features and let $g(x;u) : \mathcal{X}^D \to \mathbb{R}^d$ a filter parameterized by u. Let z be a target variable such as medical diagnosis that is of interest to participants or the public. If a researcher has a predictor $z^{\text{pred}}(g(x);w)$ (parameterized by w), then the expected dis-utility of the filter and the predictor can be measured by the expected risk:

$$f_{\text{util}}(u,w) = E[l(z^{\text{pred}}(g(x;u);w), z^{\text{true}})]. \quad (1)$$

At the same time, an adversary can make a prediction $y^{\text{pred}}(g(x);v)$ (parameterized by v) of a private variable y, which can be an identifier (in re-identification attacks) or a sensitive attribute (in inference attacks) from the filtered features $g(x)$. The expected privacy of the filter and the predictor can also be measured by the expected risk [2]:

$$f_{\text{priv}}(u,v) = E[l(y^{\text{pred}}(g(x;u);v), y^{\text{true}})]. \quad (2)$$

The goal of a filter designer is to find a filter with parameter u that achieves the two objectives – *minimum dis-utility*

$$\min_u \min_w f_{\text{util}}(u,w) = \min_u[-\max_w -f_{\text{util}}(u,w)], \quad (3)$$

and *maximum privacy*

$$\max_u \min_v f_{\text{priv}}(u,v) = -\min_u[\max_v -f_{\text{priv}}(u,v)]. \quad (4)$$

To achieve the two opposing goals, one solves the following:

$$\min_u[-\rho \max_w -f_{\text{util}}(u,w) \; + \; \max_v -f_{\text{priv}}(u,v)], \quad (5)$$

where ρ determines the relative importance of utility versus privacy. The solution to (5) is referred to as **Minimax filter** [11] and is by definition an optimal filter for utility-privacy tradeoff in terms of expected risks, given the family of filters and the family of losses/classifiers. Fig. 1 is an example with multilayer neural network as a filer and classifiers.

2.2. Differentially private minimax filter

2.2.1. Local differential privacy for continuous data

A randomized algorithm that takes data $\mathcal{D}$ as input and outputs a function f is called ϵ-differentially private if

$$P(f(\mathcal{D}) \in \mathcal{S}) \leq e^{\epsilon} P(f(\mathcal{D}') \in \mathcal{S}) \quad (6)$$

[2]Notations in this paper are different from [11].

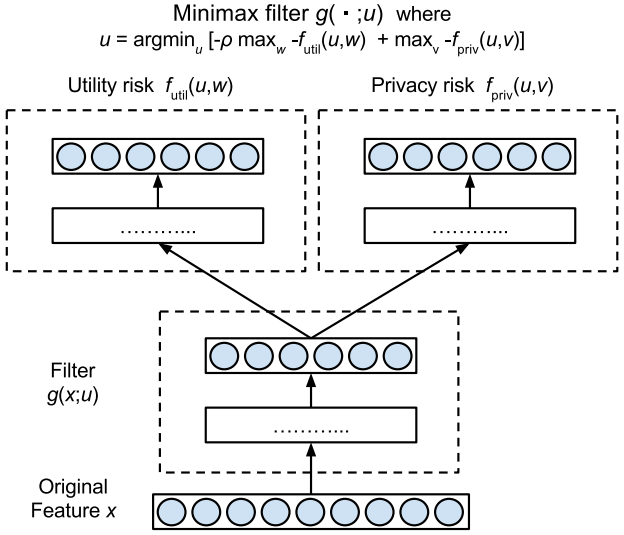

Fig. 1. Minimax filter.

for all measurable $\mathcal{S} \subset \mathcal{T}$ of the output range and for all datasets $\mathcal{D}$ and $\mathcal{D}'$ differing in a single item, denoted by $\mathcal{D} \sim \mathcal{D}'$. That is, even if an adversary knows the whole dataset $\mathcal{D}$ except for a single item, she cannot infer much more about the unknown item from the output f of the algorithm. When an algorithm outputs a real-valued vector $f \in \mathbb{R}^D$, its global sensitivity [7] can be defined as

$$S(f) = \max_{\mathcal{D} \sim \mathcal{D}'} \|f(\mathcal{D}) - f(\mathcal{D}')\| \quad (7)$$

where $\|\cdot\|$ is the norm such as the Euclidean norm. An important result from [7] is that a vector-valued output f with sensitivity $S(f)$ can be made ϵ-differentially private by perturbing f with an additive noise vector ξ whose density is

$$P(\xi) \propto e^{-\frac{\epsilon}{S(f)}\|\xi\|}. \quad (8)$$

The definition of adjacency $\mathcal{D} \sim \mathcal{D}'$ depends on problem setting. In this paper, privacy of a single sample $\mathcal{D} = \{x\}$ is considered, and the participants do not trust the data collector/curator and therefore apply perturbation before sending data to the collector/curator, also known as *local* differential privacy [12]. In such a setting, $f(\cdot)$ is the identity function, and the sensitivity (7) is simply the diameter of data[3]

$$S = \max_{x,x' \in \mathcal{X}} \|x - x'\|. \quad (9)$$

Without making any fragile assumption on the data, the paper proposes to directly clamp the diameter of $\mathcal{X}$ with a *bounding function* $b : \mathbb{R}^D \to \mathbb{R}^D$:

1. Hard-bound: $b(x) = \min\{1,\ 1/\|x\|\}x$
2. Soft-bound: $b(x) = \tanh(a\|x\|)x$

[3]Assuming $\mathcal{X}$ is compact.

3. Normalization: $b(x) = x/\|x\|$.

When the norm of features does not contain crucial information, the bounding functions can limit the sensitivity of transmitting the original data without negative effects.

2.2.2. *Preprocessing vs postprocessing*

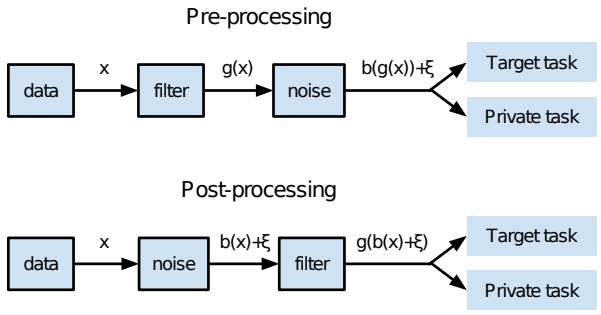

Fig. 2. Pre vs postprocessing approach to differentially private minimax filtering.

Minimax filters can be made locally differentially private by additive noise (8) in the signal chain of the filter. Depending on where the noise is injected, there are *preprocessing* and *postprocessing* approaches (see Fig. 2.) In preprocessing, the original feature x is first filtered $g(x)$, then made ϵ-differential private by bounding and perturbation $b(g(x)) + \xi$. In postprocessing, the original feature x is first made ϵ-differentially private by bounding and perturbation $b(x) + \xi$, followed by filtering $g(b(x) + \xi)$. These two approaches have different advantages and disadvantages.

A scenario when preprocessing is preferable to postprocessing is as follows. For convenience of explanation, let's assume that the private variable y is subject identity. Define *between-subject sensitivity* as the max distance of two samples from different subjects that have the same labels $z = z'$:

$$S_b = \max_{y \neq y', z = z'} \|x - x'\|. \tag{10}$$

Similarly, define *within-subject sensitivity* as the maximum distance of two samples from the same subject that have different labels $z \neq z'$:

$$S_w = \max_{y = y', z \neq z'} \|x - x'\|. \tag{11}$$

If $S_b > S_w$ for a given problem (Fig. 3a,), then after filtering we need to add less noise to achieve the same ϵ-level compared to the amount of noise required before filtering. This will results in better utility of the preprocessing approach over the postprocessing approach. From the same reasoning, if $S_w > S_b$ (Fig. 3b,) then filtering will have little effect on sensitivity, and preprocessing will offer no advantage over postprocessing. However, there are other aspects to consider as well. In preprocessing, the training of a minimax filter by solving (5) is not itself a differentially private procedure and requires a trusted collector/curator. In addition, the learned filters, when released, can leak information about private data. In contrast, postprocessing makes the process much simpler. Any postprocessing – including training of minimax filters – does not worsen differential privacy guarantees [10], and there is no need for a trusted collector/curator.

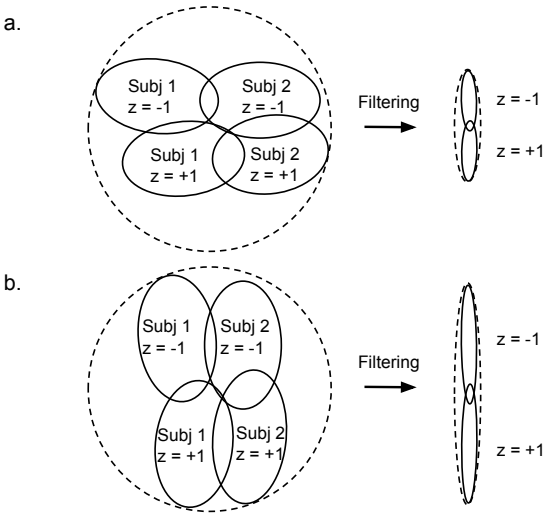

Fig. 3. Two examples of a dataset with the same data diameter and different distributions. a. When between-subject sensitivity is large. b. When within-subject sensitivity is large.

3. EXPERIMENTS

Four types of noisy filters are compared: PCA-pre, PCA-post, minimax-pre, and minimax-post. PCA is chosen as a non-minimax reference filter, which preserves the original signal in the minimum mean-squared-error sense. PCA-pre/post means that PCA is applied before/after the perturbation similarly to minimax-pre/post (see Fig. 2.) For minimax filter, a linear filter of the same dimension d as PCA is used. Tests are performed for different values of d, and only the case of $d = 20$ is reported due to lack of space. The tradeoff coefficient (5) of $\rho = 10$ is used. For all classifiers, logistic regression with a regularization factor (10^{-6}) is used throughout the tests. Optimization of (5) is done similarly to [11]. All tests are repeated 10 times for independent noise samples of (8), for each of 10 random training/test splits.

3.1. Datasets

Three public datasets of face, voice, and motion data are used for evaluating noisy minimax filter (see [11] for details.)

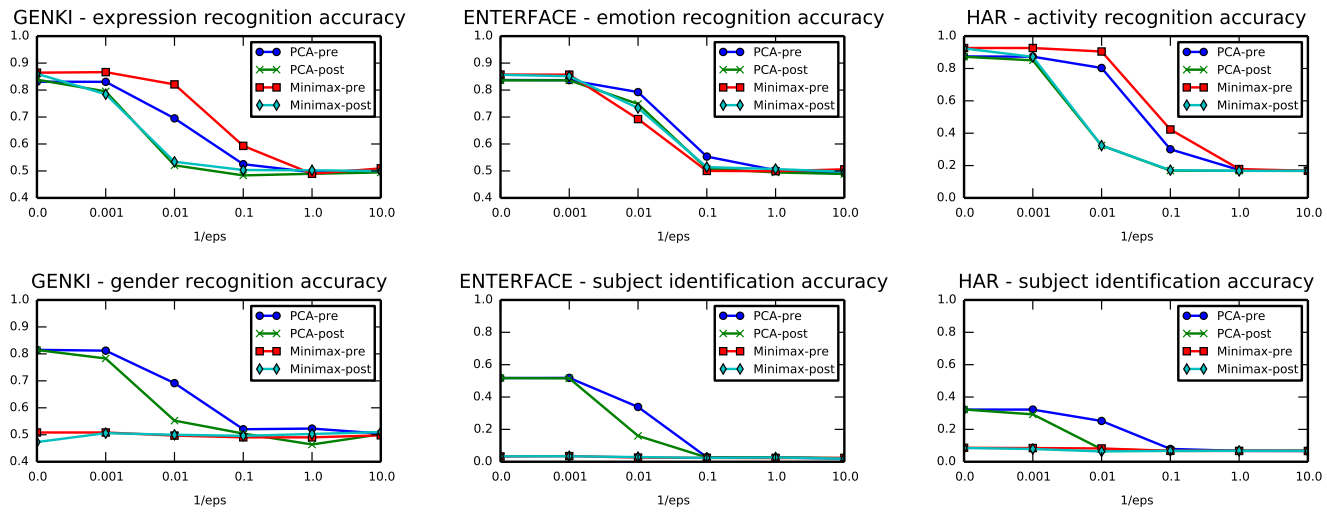

Fig. 4. Impact of four noisy filters (PCA-pre/post and Minimax-pre/post) on the accuracy of target and private tasks for three datasets (GENKI, ENTERFACE, HAR), over the range of $\epsilon^{-1} = \{0, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}\}$. Top row is the target task accuracy (higher is better) and bottom row is the private task accuracy (lower is better.) Minimax filters can limit the accuracy of inference attack (bottom row) to almost chance levels, in all ranges of ϵ.

Gender/expression recognition from face. The GENKI database [13] consists of face images with varying poses and facial expressions, with $D = 256$ and $N = 1740$ for training and $N = 100$ for testing. The target task is binary facial expression classification and the private task is binary gender classification.

Emotion recognition from speech. The ENTERFACE database [14] is an audiovisual emotion databases of English sentences. There are $N = 427$ samples of $D = 52$ dimensional feature vectors from $S = 43$ subjects. The target task is binary classification of 'happy' vs 'non-happy' emotions from speech, and the private task is multiclass $(S = 43)$ subject identification.

Activity recognition from accelerometry. The HAR database [15] is a collection of motion sensor data on a smartphone by multiple subjects performing different activities. There are $N = 10299$ samples of $D = 561$ dimensional feature vectors. Among these, randomly-chosen 15 subjects are used. The target task is multiclass $(K = 6)$ classification of activity, and the private task is multiclass $(S = 15)$ subject identification.

3.2. Results

From Fig. 4 we can make the following observations. Firstly, within each plot, increasing the privacy level from left (ϵ^{-1}=0) to right (ϵ^{-1}=10) lowers the accuracy of both target and private tasks for all filter types and datasets, which is intuitive. Secondly, target task accuracy (top row) shows that the four filters are equally accurate with no noise (ϵ^{-1}=0). In GENKI and HAR, preprocessing is better than postprocessing for both PCA and Minimax, and Minimax-pre performs the best. In ENTERFACE, preprocessing and postprocessing approaches perform similarly, and the difference among filters is relatively small. This result may be ascribed to the discussion in Sec. 2.2.2. Thirdly, and most importantly, private task accuracy (bottom row) is quite different between the proposed (Minimax-pre/post) and the reference (PCA-pre/post) methods. For both Minimax-pre and Minimax-post, the private task accuracy is almost as low as the chance accuracy of each dataset (0.5, 0.03, 0.07) regardless of the noise level ϵ, which demonstrates that minimax filter can prevent inference attacks with little help of noise. In contrast, the non-minimax filters (PCA-pre/post) allow an adversary to infer private variables quite accurately (0.8, 0.5, 0.3) when no noise is used (ϵ^{-1}=0). Preventing such an attack for non-minimax filters requires a large amount of noise (e.g., ϵ^{-1}=0.1), which destroys data utility.

4. CONCLUSION

The paper presented preprocessing and postprocessing approaches to impart differential privacy to minimax filters. Due to different properties of the two concepts, the combination can inherit advantages from both sides – high on-average utility and protection against inference attacks, and a formal privacy guarantee in the worst case. Future work will focus on refining the proposed methods. In particular, the postprocessing approach can potentially use the knowledge of noise distribution to improve learning, analogous to [16] where probabilistic estimators are learned from the noisy output of a mechanism.

5. REFERENCES

[1] Benjamin Fung, Ke Wang, Rui Chen, and Philip S Yu, "Privacy-preserving data publishing: A survey of recent developments," *ACM Comp. Surveys (CSUR)*, vol. 42, no. 4, pp. 14, 2010.

[2] Latanya Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557–570, 2002.

[3] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 1, no. 1, pp. 3, 2007.

[4] Daniel Kifer, "Attacks on privacy and definetti's theorem," in *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*. ACM, 2009, pp. 127–138.

[5] Arvind Narayanan and Vitaly Shmatikov, "Robust deanonymization of large sparse datasets," in *Security and Privacy, 2008. SP 2008. IEEE Symposium on*. IEEE, 2008, pp. 111–125.

[6] Cynthia Dwork and Kobbi Nissim, "Privacy-Preserving Data Mining on Vertically Partitioned Databases," in *Proc. CRYPTO*. Springer, 2004.

[7] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography*, pp. 265–284. Springer, 2006.

[8] Cynthia Dwork, "Differential privacy," in *Automata, languages and programming*, pp. 1–12. Springer, 2006.

[9] Graham Cormode, "Personal privacy vs population privacy: learning to attack anonymization," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2011, pp. 1253–1261.

[10] Cynthia Dwork and Aaron Roth, "The algorithmic foundations of differential privacy," *Foundations and Trends in Theoretical Computer Science*, vol. 9, no. 3-4, pp. 211–407, 2014.

[11] Jihun Hamm, "Preserving privacy of continuous high-dimensional data with minimax filters," in *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2015.

[12] John C Duchi, Michael I Jordan, and Martin J Wainwright, "Local privacy and statistical minimax rates," in *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*. IEEE, 2013, pp. 429–438.

[13] Jacob Whitehill and Javier Movellan, "Discriminately decreasing discriminability with learned image filters," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2488–2495.

[14] Olivier Martin, Irene Kotsia, Benoit Macq, and Ioannis Pitas, "The enterface´05 audio-visual emotion database," in *Data Engineering Workshops, 2006. Proceedings. 22nd International Conference on*. IEEE, 2006, pp. 8–8.

[15] Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, and Jorge L Reyes-Ortiz, "Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine," in *Ambient assisted living and home care*, pp. 216–223. Springer, 2012.

[16] Oliver Williams and Frank McSherry, "Probabilistic inference and differential privacy," in *Advances in Neural Information Processing Systems*, 2010, pp. 2451–2459.