SAND2022-10568C

Differentially Private DBSCAN via Subsampling for Out-of-Distribution Detection

Alycia N. Carey*, Michael R. Smith*, Mitchell Negus, Nicholas D. Pattengale, Jonathan P. Roose
Sandia National Laboratories

Abstract

(150 words)

1 Introduction

Most machine learning make the implicit assumption that they exist and operate in a closed world; i.e., the data that they will encounter in the wild will be drawn from the same distribution as it was trained on. However, the probability of that assumption holding during deployment is low. Compounded with the fact that machine learning models often produce high confidence answers on examples that fall outside of their training distribution, it is prudent to have a system that can identify if a data example is suitable for the model to classify in the first place. Out-of-Distribution (OOD) detection is a family of closely-related tasks (e.g., anomaly detection, outlier detection, . . .) that have been proposed solve this problem across a multitude of settings.

We focus on the task of using outlier detection as a means of OOD. To motivate this need, we demonstrate the problems associated with OOD for detecting malware and malicious PDFs. In addition to the problem above, machine learning models have been shown to be prone to privacy leakage under many different circumstances. When data is trained on sensitive data, such as what malware a system can defend against, protecting this data is of utmost importance.

One common approach to increasing the privacy of an ML system is to implement differential privacy – a mathematical construct that protects sensitive data. Most differential privacy approaches to ML add noise to the data or in certain parts of the learning algorithm. But, recent studies have shown that in order to have sufficient privacy guarantees, large quantities of noise must be added which significantly impacts the performance of the ML algorithm. Most techniques that propose a differentially private version of outlier detection fall victim to this issue as noise is simply added to the data before being used. Nonetheless, there are techniques that do not require adding noise and still adhering to the guarantees of differential privacy. One approach of this manner is subsampling. Subsampling relies on the inherent randomness of the sampling procedure to ensure adherence to differential privacy and has been shown to increase other privacy masures (requiring less noise). We examine subsampling in addition to strategically placed noise in DBScan to produce a more efficient deferentially private outlier detection method.

Paragraph on correlation here....

Paragraph for motivating example goes here ... Will focus on malware and dynamic nature of malware. An operator needs to know when the data has shifted or when new types of malware have been discovered and should be examined by an nalayst.

To solve these issue we propose SDP-DBSCAN, a differentially private variant of Density Based Spatial Clustering of Applications with Noise (DBSCAN) based on subsampling and the strategic addition of noise in the DBSCan algorithm. Instead of adding noise to the data itself or to only the core points, we instead add it to the thresholding function which determines the final clusters. Further, subsampling amplifies privacy (CITE), when it is used in a differentially private algorithm, less noise needs to be added in order to meet the desired privacy level. Therefore, our approach is able to reach the same level of privacy as other DP-DBSCAN techniques while achieving higher utility due to adding less noise. Additionally, We do this because ... Further, we propose CSDP-DBSCAN, a variant of SDP-DBSCAN that still provides differential privacy when correlated data is present in the training set. We further examine subsampling that removes separates correlated data points.

Our major contributions include: (1) a differentially private implementation of DBSCAN that does not require the addition of noise to the training data or core data points; (2) a differentially private implementation of DBSCAN that accounts for correlation in the training data without injecting noise to the training data itself; and (3) the empirical evaluation of both SDP-DBSCAN and CSDP-DBSCAN against several baselines to show that they can guarantee the same level of privacy with higher utility with less noise.

The remainder of the paper is organized as follows. In Section ....

2 Related Works

In this section we will detail the important works related closely to our publication.

2.1 Out-of-Distribution Detection and outlier detection

- The Connection between Out-of-Distribution Generalization and Privacy of ML Models [9]

- Other previous approaches (including outlier detection, anomaly detection, ...) and their differences

2.2 Differential Privacy

- The Algorithmic Foundations of Differential Privacy [3]

- Practical applications of DP

2.3 Privacy Amplification by Subsampling

- Privacy Amplification by Subsampling: Tight Analyses via Couplings and Divergences [1]

- On the Intrinsic Differential Privacy of Bagging [8]

- Differentially Private Bagging: Improved Utility and Cheaper Privacy than Subsample-and-Aggregate [5]

2.4 Differential Privacy and Correlation

- No Free Lunch in Data Privacy [6]

- Dependence Makes You Vulnerable: Differential Privacy Under Dependent Tuples [7]

- Correlated Differential Privacy: Hiding Information in Non-IID Data Set [16]

- Differentially Private Outlier Detection in Correlated Data [2]

- Correlated Data in Differential Privacy: Definition and Analysis [15]

3 Problem Formulation

In this section, we formulate and present our approach to differentially private out of distribution detection.

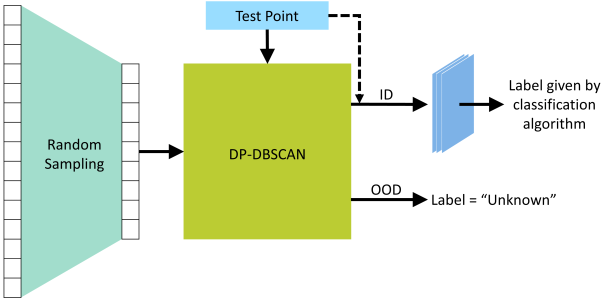

Figure 1: Overview of architecture for SDP-DBSCAN.

3.1 DBSCAN

3.2 Differential Privacy

- Introduce DP (main definition from Dwork)

- Introduce type of noise used (Laplace)

- Introduce any composition algorithms we need (or introduce later in correlated section)

3.2.1 Privacy Amplification by Subsampling

- Introduce theory behind privacy amplification

- Many do subsample and aggregate, but in tests w/o correlation taken into effect we only subsample once (i.e., not PATE). Additionaly, noise is added suring inference time which can exhaust a privacy budget and the model will need to be retrained.

3.3 SDP-DBSCAN

We now present our approach to differentially private DBSCAN using sampling, SDP-DBSCAN. An overview of the architecture can be found in Fig. 1.

- Formulate our method for non-correlated data

  - Subsample once

  - Add noise to neighbor count (DBSCAN)

  — Derive sensitivity

  - if not out of distribution feed data point to classification algorithm

  - otherwise classify as "Unknown"

- Algorithm

3.4 Correlation Analysis

- Introduce theory behind correlation and DP
- Correlated features
- Correlated Points
- More like PATE, have to so subsample and aggregate

3.5 CSDP-DBSCAN

We now present our method CSDP-DBSCAN that is able to provide DP guarantees while using SDP-DBSCAN on correlated data. An overview of the architecture can be found in Fig. 2.

- Formulate our method for correlated data

  - Subsample once

  - Add noise to neighbor count (DBSCAN)

  — Derive sensitivity

  - if not out of distribution feed data point to classification algorithm

  - otherwise classify as “Unknown"
- Diagram
- Algorithm

4 Experiments

We compare our methods against the baselines along accuracy, recall/precision, ...

4.1 Datasets

We test SDP-DBSCAN against three different datasets:

- MNIST/EMNIST
- PDFRate [12]
- Genomics [11]
- Microsoft Malware Classification Challenge.

We test CSDP-DBSCAN on synthetic data. Specifically, the synthetic dataset will be a collection of points (and their associated 'label') that have correlations so we can test CSDP-DBSCAN more concretely.

4.2 Baselines

We compare both SDP-DBSCAN and CSDP-DBSCAN against three baselines:

- **DP-DBSCAN [10, 14]**: DP-DBSCAN adds noise to each dimension of the direct density-reachable points in the dataset by differential privacy technique so that the published data can conform to the privacy budget requirement, thereafter the privacy of the data is protected during clustering. Since DP-DBSCAN publishes the approximation of data points density, the attackers cannot deduce the sensitive properties of the data points even if they grasp some information through the knowledge background. However, when the privacy budget parameter ε is small (i.e., the added noise is too large), the accuracy of DP-DBSCAN clustering algorithm will decrease. Moreover, when the data size is large and the density is non-uniform, the clustering efficiency will also decrease.
- **I-DP-DBSCAN** [4]: The Improved DP-DBSCAN (I-DP-DBSCAN) algorithm achieves differential privacy by injecting Laplacian noise into the Euclidean distance of every two data objects.
- **DP-MCDBSCAN** [10]: In order to solve the drawbacks of DP-DBSCAN where the initial core object is randomly selected, we propose a DP-MCDBSCAN (Differential Privacy Preservation Multicore DBSCAN Clustering) algorithm which determines multiple core objects as the initial object to cluster through the furthest distance selection method. Our algorithm ensures that the initial cluster centers are dispersed as far as possible so that the initial core objects selected are not in the same cluster, reducing the influence of the initial core objects selection on the clustering result.
- **DP-Kmeans [13]**: Given a d-dimensional dataset, partition the domain into M equal-width grid cells, and then releases the noisy count in each cell, by adding Laplacian noise to each cell count.

4.3 Experiments on Non-Correlated Data

In order to show that SPD-DPBSCAN provides DP guarantees without loosing utility, we perform four different tests: 1) injecting noise into the neighbor counts only; 2) subsampling plus noise injection to the neighbors; 3) injecting noise into the data only; and 4) subsampling plus noise injection to the data. We perform these experiments across three different datasets (MNIST/EMNIST, Genomics, and PDFRate) as well as on all of the mentioned baselines. Additionally, we will perform the experiments on a range of epsilon values to show that SPD-DBSCAN provides higher utility than other baselines when epsilon is small.

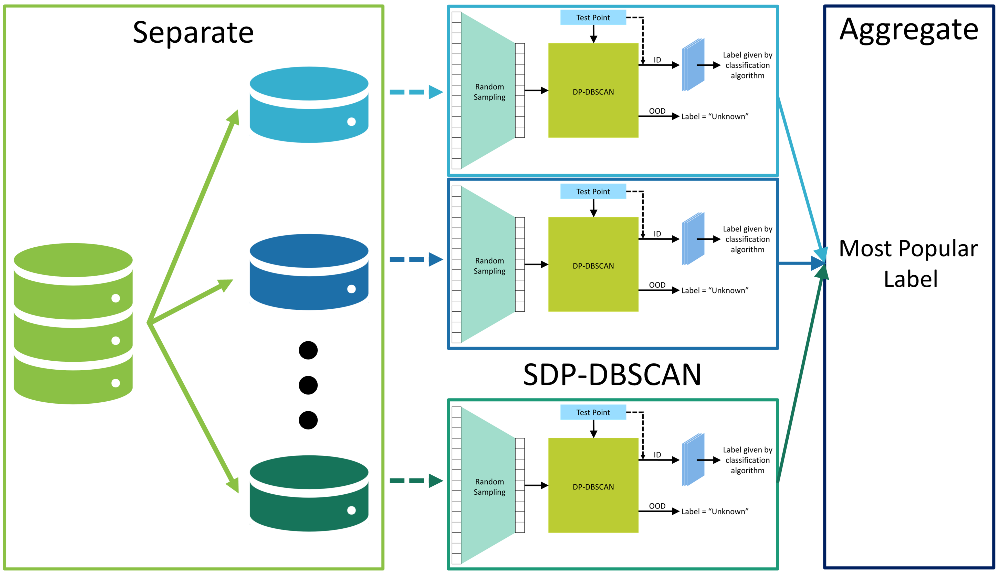

Figure 2: Overview of architecture for CSDP-DBSCAN.

4.3.1 Noisy Neighbors

In this experiment, we will test the effectiveness of our DP OOD methods by introducing noise to the neighbor counts in DBSCAN. We will choose a range of epsilon values to try and calculate the utility of the DP-DBSCAN with noisy neighbors to correctly identify outliers. In our datasets we will know which points are out of distribution before hand (e.g., if the data comes from the emnist class it is out of distribution from the training set of MNIST), so we will be able to calculate the accuracy of the DP enforced mechanism.

4.3.2 Subsampling + Noisy Neighbors

This experiment builds on the previous, but instead of using all of the training data, we will take a subsample and then perform DP OOD detection and classification. This experiment it to show that subsampling allows for less noise to be added to the neighbors in order to reach the same DP privacy level which will (most likely) result in higher accuracy.

4.3.3 Noisy Data

As a foil to our adding of noise to the neighbor counts, we will test the standard method of adding noise to the training data before processing it through DBSCAN. Again we will try multiple different epsilon levels and compare our approach against the baselines along accuracy.

4.3.4 Subsampling + Noisy Data

Additionally, we will repeat the immediate test above but with subsampling to once again show the power of subsampling.

4.4 Experiments on Correlated Data

In the correlation experiments, we will specifically partition the data into subsets that do not contain any correlated data within itself. In this manner we will then have k different partitions, and we will make use of DP composition theorems to combine the answer from all of the different partitions. The experiments are the same as above - the only changes being the forced separation of the data and the analysis on multiple subsets, not just one. Also, in addition to testing different epsilon values, we will test different number of splits of the data into non-correlated partitions.

4.4.1 Noisy Neighbors

4.4.2 Subsampling + Noisy Neighbors

4.4.3 Noisy Data

4.4.4 Subsampling + Noisy Data

5 Results

We will include the following tables, figures, and writings for our results:

- Comparison table with accuracy, precision/recall, ... for different epsilons for the non-correlated data experiments (results for SDP-DBSCAN, I-DP-DBSCAN, DP-MCDBSCAN, and DP-KMeans) on all datasets

- Comparison table with accuracy, precision/recall, ... for different epsilons for the correlated data experiments (results for CSDP-DBSCAN, I-DP-DBSCAN, DP-MCDBSCAN, and DP-KMeans) on all datasets

- Graph comparing accuracy vs epsilon for the non-correlated data experiments
- Graph comparing accuracy vs epsilon and accuracy vs num-splits for the correlated data experiments

6 Analysis

Note: this section could possibly be integrated into the methods section.

6.1 Theoretical Analysis of SDP-DBSCAN

Here we will prove why subsampling + adding noise to the neighbors adheres to differential privacy. Specifically, we will show that our derived privacy level (ε, δ) is an improvement over the other baselines.

6.2 Theoretical Analysis of CSDP-DBSCAN

Here we will prove why our separation strategy works to deal with correlated data and derive our privacy bounds.

7 Conclusion

Acknowledgments

The USENIX latex style is old and very tired, which is why there's no \acks command for you to use when acknowledging. Sorry.

Availability

USENIX program committees give extra points to submissions that are backed by artifacts that are publicly available. If you made your code or data available, it's worth mentioning this fact in a dedicated section.

References

[1] Borja Balle, Gilles Barthe, and Marco Gaboardi. Privacy amplification by subsampling: Tight analyses via couplings and divergences. *Advances in Neural Information Processing Systems*, 31, 2018.

[2] Kwassi H Degue, Karthik Gopalakrishnan, Max Z Li, and Hamsa Balakrishnan. Differentially private outlier detection in correlated data. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 2735–2742. IEEE, 2021.

[3] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407, Aug 2014.

[4] Yuzhen Jin and Shuyu Li. An improved differentially private dbscan clustering algorithm for vehicular crowdsensing. In *2019 IEEE 13th International Conference on Anti-counterfeiting, Security, and Identification (ASID)*, pages 51–55, 2019.

[5] James Jordon, Jinsung Yoon, and Mihaela van der Schaar. Differentially private bagging: Improved utility and cheaper privacy than subsample-and-aggregate. *Advances in Neural Information Processing Systems*, 32, 2019.

[6] Daniel Kifer and Ashwin Machanavajjhala. No free lunch in data privacy. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, volume 37, pages 193–204, Athens, Greece, jun 2011. ACM.

[7] Changchang Liu, Supriyo Chakraborty, and Prateek Mittal. Dependence makes you vulnerable: Differential privacy under dependent tuples. In *Proceedings of the Network and Distributed System Security Symposium, NDSS'16*, volume 1, pages 1 – 15, San Diego, California, Feb 2016. Internet Society.

[8] Hongbin Liu, Jinyuan Jia, and Neil Zhenqiang Gong. On the intrinsic differential privacy of bagging. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, volume 30, pages 2730 – 2736, Montreal, Canada, Aug 2021. IJCAI.

[9] Divyat Mahajan, Shruti Tople, and Amit Sharma. The connection between out-of-distribution generalization and privacy of ml models. *arXiv preprint arXiv:2110.03369*, 2021.

[10] Lina Ni, Chao Li, Xiao Wang, Honglu Jiang, and Jiguo Yu. Dp-mcdbscan: Differential privacy preserving multicore dbscan clustering for network user data. *IEEE access*, 6:21053–21063, 2018.

[11] Jie Ren, Peter J Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, Joshua Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. *Advances in neural information processing systems*, 32, 2019.

[12] Charles Smutz and Angelos Stavrou. Malicious pdf detection using metadata and structural features. In *Proceedings of the 28th annual computer security applications conference*, pages 239–248, 2012.

[13] Dong Su, Jianneng Cao, Ninghui Li, Elisa Bertino, and Hongxia Jin. Differentially private k-means clustering. In *Proceedings of the sixth ACM conference on data and application security and privacy*, pages 26–37, 2016.

[14] W Wu and H Huang. A dp-dbscan clustering algorithm based on differential privacy preserving. *Computer Engineering and Science*, 37(4):830–834, 2015.

[15] Tao Zhang, Tianqing Zhu, Renping Liu, and Wanlei Zhou. Correlated data in differential privacy: definition and analysis. *Concurrency and Computation: Practice and Experience*, 34(16):e6015, 2022.

[16] Tianqing Zhu, Ping Xiong, Gang Li, and Wanlei Zhou. Correlated differential privacy: Hiding information in non-iid data set. *IEEE Transactions on Information Forensics and Security*, 10(2):229–242, 2015.