An Investigation of Data Privacy and Utility Preservation using KNN Classification as a Gauge

Kato Mivule[1] and Claude Turner PhD[2]
[1]mivulek0220@students.bowiestate.edu,[2]cturner@bowiestate.edu
Computer Science Department, Bowie State University, Bowie, MD, USA

Abstract – It is obligatory that organizations by law safeguard the privacy of individuals when handling datasets containing personal identifiable information (PII). Nevertheless, during the process of data privatization, the utility or usefulness of the privatized data diminishes. Yet achieving the optimal balance between data privacy and utility needs has been documented as an NP-hard challenge. In this study, we investigate data privacy and utility preservation using KNN machine learning classification as a gauge.

Keywords: Data Privacy Preservation, Data Utility, Machine Learning, KNN Classification.

I. Introduction

During the process of data privatization, the utility or usefulness of the privatized data diminishes. Yet achieving the optimal balance between data privacy and utility needs has been documented as an NP-hard challenge [1] [2]. In this study, we investigate data privacy and utility preservation using KNN machine learning classification as a gauge. As Cynthia Dwork succinctly and aptly stated [6]:

> "Perfect privacy can be achieved by publishing nothing at all, but this has no utility; perfect utility can be obtained by publishing the data exactly as received, but this offers no privacy".

In this study, we investigate data privacy and utility preservation using KNN machine learning classification as a gauge [4].

Noise addition: is a data privacy perturbative method that adds a random value, usually selected from a normal distribution with zero mean and a very small standard deviation, to sensitive numerical attribute values to ensure privacy [3] [8]. The general expression of noise addition as defined:

$$X + \varepsilon = Z \qquad (1)$$

Where X is the original numerical dataset and ε is the set of random values (noise) with a distribution $e \sim N(0, \sigma^2)$ that is added to X, and finally Z is the privatized dataset.

This work was supported in part by the U.S. Department of Education HBGI Grant.

Claude Turner, PhD is an Associate Professor of Computer Science and Director for the Center for Cyber Security and Emerging Technologies at Bowie State University. (E-mail: cturner@bowiestate.edu).

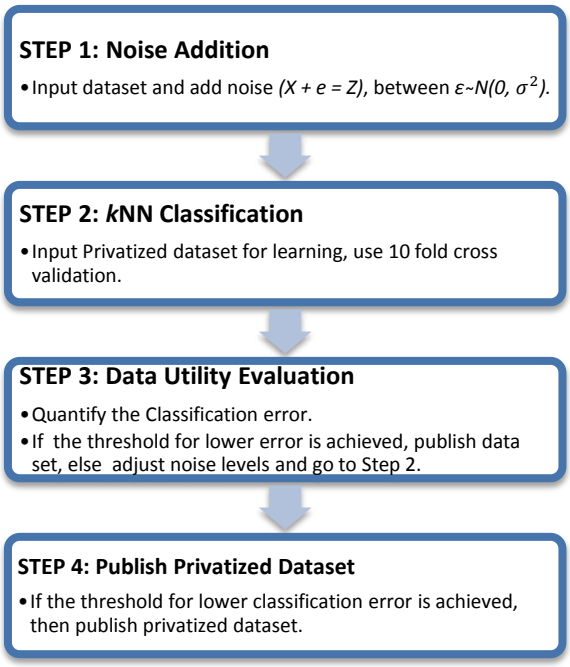
Kato Mivule is a doctoral candidate, Computer Science Department, Bowie State University. (E-mail: mivulek0220@students.bowiestate.edu).

K Nearest Neighbors (KNN): is a classification method that matches items in the test data to those in the training data by measuring the distance between the two items. Any k items that are closer to each other are then placed in the same class. The Euclidean distance is the normally used distance measure for KNN expressed as follows [5]:

$$distance(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad (2)$$

II. Methodology

In the first stage of our approach, we apply a data privacy procedure, in this case, noise addition, on the Iris dataset for privacy [7]. The privatized Iris dataset is then sent to the KNN machine learning classifier for training and testing using 10 fold cross validation; the classification error is quantified. If the classification error is lower or equal to a threshold, then better utility might be achieved, otherwise, we adjust the data privacy parameters and re-classify the results.



III. Experiment

In our experiment, we used the Iris dataset from the UCI machine learning repository as our original dataset [9]. We then privatized the dataset by using the noise addition data privacy technique. We then used KNN classification and quantified the classification error. We adjusted the noise

levels and run the privatized dataset through the KNN classifier after which we published the results. We used MATLAB for both noise addition and KNN classification.

IV. Results

As shown in our initial results, only 4 percent of records from the original Iris dataset were misclassified. When noise addition was chosen between the mean and standard deviation for the privatized dataset, 32 per cent of records got misclassified. However, when noise addition was reduced to mean = 0 and standard deviation = 0.1 for the privatized dataset, 26 percent of records got misclassified, a 6 point reduction in classification error.

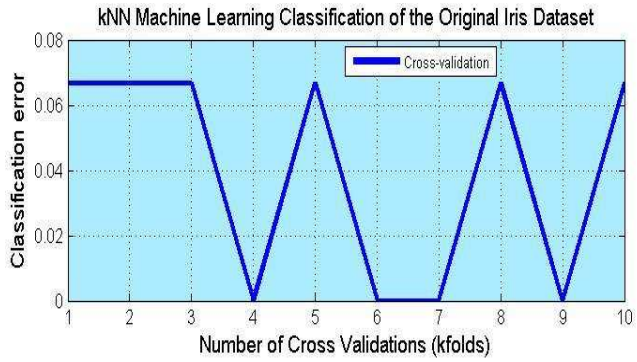


Fig 1: KNN classification of the original Iris dataset with classification error at 0.0400 (4 percent misclassified data)

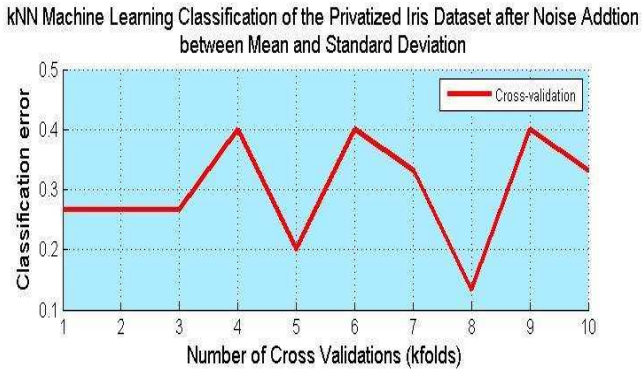


Fig 2: KNN classification of the privatized Iris dataset with noise addition between the mean and standard deviation.

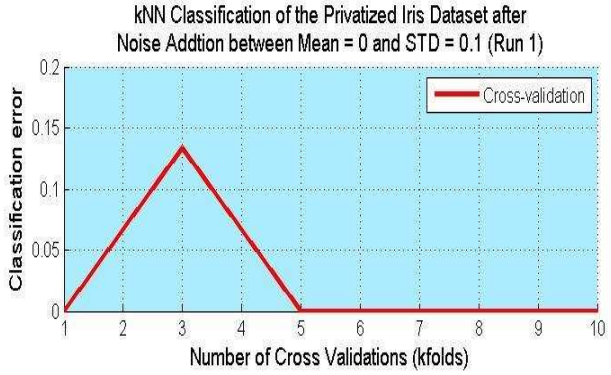


Fig 3: KNN classification of the privatized Iris dataset with reduced noise addtion between mean = 0 and standard deviation = 0.1

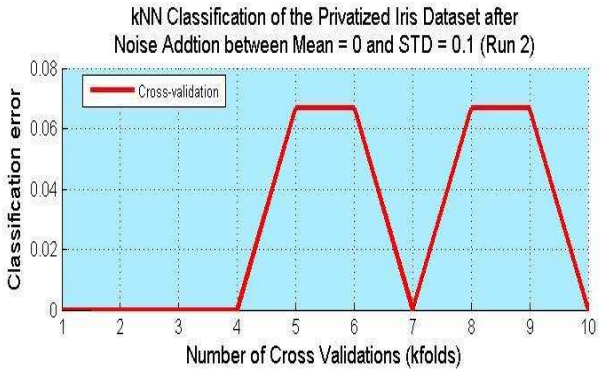


Fig 4: A second run of the kNN classification of the privatized Iris dataset with reduced noise addition between mean = 0 and standard deviation = 0.1.

V. Conclusion and Discussion

The initial results from our investigation show that a reduction in noise levels does affect the classification error rate. However, this reduction in noise levels could lead to low risky privacy levels. Finding the optimal balance between data privacy and utility needs is still problematic.

Acknowledgment

Special thanks to Dr. Claude Turner and the Computer Science Department at Bowie State University for making this work possible.

References

[1] R. C.-W. Wong, A. W.-C. Fu, K. Wang, and J. Pei, "Minimality Attack in Privacy Preserving Data Publishing," Proceedings of the 33rd international conference on Very large data bases, pp. 543–554, 2007.

[2] A. Krause and E. Horvitz, "A Utility-Theoretic Approach to Privacy in Online Services," Journal of Artificial Intelligence Research, vol. 39, pp. 633–662, 2010.

[3] J. Kim, "A Method For Limiting Disclosure in Microdata Based Random Noise and Transformation," in Proceedings of the Survey Research Methods, American Statistical Association,, 1986, vol. Jay Kim, A, no. 3, pp. 370–374.

[4] M. Banerjee, "A utility-aware privacy preserving framework for distributed data mining with worst case privacy guarantee," University of Maryland, Baltimore County, 2011.

[5] B. Liú, Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, Datacentric Systems and Applications. Springer, 2011, pp. 124–125.

[6] C. Dwork, "Differential Privacy," in Automata languages and programming, vol. 4052, no. d, M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, Eds. Springer, 2006, pp. 1–12.

[7] K. Mivule, C. Turner, and S.-Y. Ji, "Towards A Differential Privacy and Utility Preserving Machine Learning Classifier," in Procedia Computer Science, 2012, vol. 12, pp. 176–181.

[8] K. Mivule, "Utilizing Noise Addition for Data Privacy, an Overview," in Proceedings of the International Conference on Information and Knowledge Engineering (IKE 2012), 2012, pp. 65–71.

[9] Frank, A., Asuncion, A. Iris Data Set, UCI Machine Learning Repository [http://archive.ics.uci.edu/ml/datasets/Iris]. Department of Information and Computer Science, University of California, Irvine, CA (2010).