Privacy-Preserving Clustering

Hannah Keller*, Helen Möllering, Thomas Schneider, Hossein Yalame
Technical University of Darmstadt

32nd Crypto Day, 15 January 2021

Clustering is an unsupervised machine learning technique that divides a data set into groups. The manifold applications of clustering range from segmenting a market using consumer preferences [1] to grouping photos of diseased organs in medical imaging [2]. These scenarios inevitably require sensitive data such as business secrets or personal health records. We consider multiple mutually distrusting parties, e.g., corporations or hospitals that do not want to or are not permitted to share their clients' or patients' data due to legislation such as GDPR [3] or HIPAA[4] or to protect their market position and business secrets. Nevertheless, these parties wish to cluster their joint data, since a larger data pool usually results in a better result providing more insights.

A growing body of research proposes privacy-preserving variants of clustering algorithms. These algorithms operate in a semi-honest security model, which assumes that all parties honestly follow the protocol. However, existing implementations are impractical for real-world usage [5] due to immense computational costs or privacy-compromising data leakage during the clustering process [6, 7, 8]. To the best of our knowledge, the few clustering schemes that fully preserve privacy [9, 10, 5] focus on the simple and efficient k-means algorithm; however, the effectiveness of k-means depends on many assumptions. For example, the number of clusters must be known in advance, and the input data cannot include nominal variables or outliers. These assumptions often do not hold for real-world clustering applications, limiting k-means' practicality.

In our work, we explore the suitability of advanced clustering algorithms for efficient crypto-oriented clustering. We identify affinity propagation [11] to be the most promising algorithm candidate. Affinity propagation and k-means have comparable computational complexities of $O(n^2)$, and affinity propagation provides more flexibility in terms of tolerance to outliers, compatibility with various data types, and the selection of tunable parameters. Furthermore, affinity propagation involves only operations that can be implemented relatively efficiently using secure computation techniques, such as addition and comparison. The algorithm was also successfully applied for privacy-relevant use cases, e.g., the detection of genes from chromosomal data [11].

Based on our insights, we design the *first* privacy-preserving affinity propagation clustering protocol using secure computation techniques. Moreover, our protocol was implemented and benchmarked with the SPDZ framework [12] and enables private clustering in a malicious security model for the first time, providing protection from strong adversaries that deviate from the protocol.

*Main author

crypto day matters 32 (2021)
Beiträge vom 32. Kryptotag, 15.01.2021, online.
doi:10.18420/cdm-2021-32-42

References

[1] A. Chaturvedi, J. Carroll, P. Green, and J. A. Rotondo, “A feature-based approach to market segmentation via overlapping k-centroids clustering,” *Journal of Marketing Research*, 1997.

[2] F. Masulli and A. Schenone, “A fuzzy clustering based segmentation system as support to diagnosis in medical imaging,” *Artificial Intelligence in Medicine*, 1999.

[3] Article 29 Data Protection Working Party, “Opinion 05/2014 on anonymisation techniques.” `https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf` (accessed: 17.12.20), 2014.

[4] American Department of Health and Human Services, “Guidance regarding methods for de-identification of protected health information in accordance with the health insurance portability and accountability act (HIPAA) privacy rule.” `https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html` (accessed: 17.12.20), 2015.

[5] A. Jäschke and F. Armknecht, “Unsupervised machine learning on encrypted data,” *SAC*, 2019.

[6] J. Liu, J. Huang, J. Luo, and L. Xiong, “Privacy preserving distributed DBSCAN clustering,” *Transactions on Data Privacy*, 2012.

[7] P. Bunn and R. Ostrovsky, “Secure two-party k-means clustering,” in *CCS*, 2007.

[8] D. Liu, E. Bertino, and X. Yi, “Privacy of outsourced k-means clustering,” in *ASIACCS*, 2014.

[9] P. Mohassel, M. Rosulek, and N. Trieu, “Practical privacy-preserving k-means clustering,” in *PETS*, 2020.

[10] W. Wu, J. Liu, H. Wang, J. Hao, and M. Xian, “Secure and efficient outsourced k-means clustering using fully homomorphic encryption with ciphertext packing technique,” *IEEE Transactions on Knowledge and Data Engineering*, 2020.

[11] B. Frey and D. Dueck, “Clustering by passing messages between data points,” *Science*, 2007.

[12] M. Keller, “MP-SPDZ: a versatile framework for multi-party computation,” in *CCS*, 2020.