An Improved Variable –Sized Microaggregation Algorithm for Privacy Preservation (IV-MDAV)

Gajendra Singh Rawat, Dr. Bhogeshwar Borah
Department of Computer Science and Engineering, Tezpur University, India
Associate Professor, Department of Computer Science and Engineering, Tezpur University, India
Email: rawatsgajendra@gmail.com

Abstract: **Micro aggregation is a technique used to protect privacy in databases and location-based services. We propose a new hybrid technique for multivariate micro aggregation. Our technique combines a heuristic yielding fixed-size groups and a genetic algorithm yielding variable-sized groups. Fixed-size heuristics are fast and able to deal with large data sets, but they sometimes are far from optimal in terms of the information loss inflicted. On the other hand, the genetic algorithm obtains very good results (i.e. optimal or near optimal), but it can only cope with very small datasets. Our technique leverages the advantages of both types of heuristics and avoids their shortcomings. First, it partitions the data set into a number of groups by using a fixed-size heuristic. Then, it optimizes the partitions by means of the genetic algorithm. As an outcome of this mixture of heuristics, we obtain a technique that improves the results of the fixed-size heuristic in large data sets**.

I. INTRODUCTION

Over the last twenty years, there has been an extensive growth in the amount of private data collected about individuals. This data comes from a number of sources including medical, financial, library, telephone, and shopping records. Such data can be integrated and analyzed digitally as it's possible due to the rapid growth in database, networking, and computing technologies. On the one hand, this has led to the development of data mining tools that aim to infer useful trends from this data. But, on the other hand, easy access to personal data poses a threat to individual privacy. This has lead to concerns that the personal data may be misused for a variety of purposes. Detailed person-specific data in its original form often contains sensitive information about individuals, and publishing such data immediately violates individual privacy.

II. DATA MINING AND PRIVACY

Privacy is defined as the freedom from intrusion or public. It is the quality or condition of being isolated from the presence or view of others. The boundaries and content of what is considered private differ among cultures and individuals, but share basic common themes.

III. CLUSTERING

From a practical perspective clustering plays an important role in data mining application. The process of grouping a set of physical or abstracts into classes of similar objects is called clustering. A cluster of data objects can be treated collectively as one group and so may be considered as a form of data compression. It models data by its clusters. Cluster analysis has wide applications, including market or customer segmentation, pattern recognition, biological studies, spatial data analysis, web scientific data exploration, information retrieval, text mining, medical diagnostics, computational biology, and many others.

3.1 Objectives

Following are the objectives of work:

- To study the existing privacy preserving data mining methods.
- To analyze experimentally some of the popular preserving techniques.
- To evaluate the performance of the existing methods in terms of security and Information loss.

3.2 Statistical Methods for Disclosure Control[23] Statistical Disclosure Control (SDC) techniques can be defined as the set of methods to reduce the risk of disclosing information on individuals, businesses or other organizations. Such methods are only related to the dissemination step and are usually based on restricting the amount of or modifying the data released. SDC techniques can be applied to two types of data but I work on Statistical data for privacy preserving.

3.3 Micro aggregation: Micro aggregation is a set of procedures that distort empirical data in order to guarantee the factual anonymity of the data. Micro aggregation is one of the most employed micro data protection methods. The idea is to build clusters of at least k original records, and then replace them with the centroid of the cluster. When the number of attributes of the datasets is large, a common practice is to split the dataset into smaller blocks of attributes. Micro aggregation is successively and independently applied to each block. Strict application of confidentiality rules leads to replacing individual values with values computed on small aggregation.

IV. PROPOSED ALGORITHM

Proposed algorithm is an extension of the *MDAV-single-group* algorithm presented in the previous section (*algorithm-2*) to make it variable-size. We have selected *MDAV-single-group* algorithm as the basis for our variable-size algorithm because it smoothly handles less than k residuals records Variable-size algorithms show better performance for datasets with clustering tendency. This is the reason for achieving greater reduction of information loss for *EIA* dataset. *Tarragona* is a scattered dataset exhibiting no tendency for clustering, that is why reduction of information loss by proposed algorithm is very less for the *Tarragona* dataset.

For the *V-MDAV* algorithm γ value is user specified. The results for this algorithm are presented in Table 1, taking γ=0.2 for *Tarragona* and *Census* datasets and γ=1.1 for the *EIA* dataset as these two values of γ are suggested by authors of *V-MDAV* in [12]. For the *IVMDAV* algorithm the value of γ is set to 1.16.

The experimental results presented here show that proposed *IVMDAV* is a good algorithm producing micro aggregated datasets with lower information loss.

The proposed algorithm called *IVMDAV* is presented below.

(The *IVMDAV* algorithm)

1. set i=1; n=|X|;
2. while (n>=3k) do
2.1 compute centroid $\bar{x}$ of remaining records in X;
2.2 find the most distant record x_r from $\bar{x}$;
2.3 find 2k nearest neighbors $y_1,y_2,\ldots,y_{2k}$ of x_r;
2.4 form cluster c_i with first k-neighbors $y_1,y_2,\ldots,y_k$;
2.5 remove records $y_1,y_2,\ldots,y_k$ from dataset X;
2.6 set $n = n - k$; $j = k+1$;
2.7 compute centroid $\bar{x}_i$ of cluster c_i;
2.8 while (j<=2k) do
2.8.1 find k-nearest neighbors $z_1, z_2,\ldots,z_k$ of y_j in X;
2.8.2 find distance d_1 of record y_j from x_r;
2.8.3 find distance d_2 of record y_j from z_k;
2.8.4 if ($d_2 > \gamma d_1$) then
2.8.4.1 insert y_j in current cluster c_i;
2.8.4.2 recomputed centroid $\bar{x}_i$ of cluster c_i;
2.8.4.3 remove record y_j from X;
2.8.4.4 set n=n-1;
2.8.4.5 end if
2.8.5 end while
2.9 set i=i+1;
2.10. end while
3. if (n>2k) then
3.1 compute centroid $\bar{x}$ of remaining records in X;
3.2 find the most distant record x_r from $\bar{x}$;
3.3 find k nearest neighbors $y_1,y_2,\ldots,y_k$ of x_r;

3.4 form cluster c_i with the k-neighbors $y_1,y_2,\ldots,y_k$;

3.5 remove records $y_1,y_2,\ldots,y_k$ from dataset X;

3.6 set $n=n-k$; $i=i+1$;

3.7 end if

4. if ($n>0$) then

4.1 form a cluster c_i with the n remaining records;

4.2 $i=i+1$;

4.3 end if

5. end algorithm

The *IVMDAV* algorithm iterates so long as at least $3k$ records remain unassigned. In each iteration the algorithm finds $2k$ nearest neighbors, denoted by $y_1,y_2,\ldots,y_{2k}$ of the farthest record x_r from the centroid $\bar{x}$ of the remaining records in dataset X. Current cluster, c_i is formed with the first k-neighbors $y_1,y_2,\ldots,y_k$ of x_r. Each of the other k neighbors is tested for inclusion in the currently formed cluster by computing a heuristic. This algorithm also uses a constant γ whose value is slightly greater than 1.0 (in the range 1.0 - 1.20). Let, $\bar{x}_i$ be the centroid of the cluster c_i. Consider the $(k+1)$-th neighbor, y_{k+1} of x_r. Let $z_1,z_2,\ldots,z_k$, the k-nearest unassigned neighbors of y_{k+1}. Find distance d_1 of y_{k+1} from x_r. Find distance d_2 of y_{k+1} from furthest neighbor z_k. Now, if $d_2>\gamma d_1$ then insert y_{k+1} in cluster c_i and recomputed the centroid of the cluster. Then the test is repeated for $y_{k+2},\ldots,y_{2k}$. For y_{2k}, if the cluster c_i has already $2k$-1 records in it then the test should be skipped and record y_{2k} should not be inserted in the cluster c_i.

4.1 Complexity analysis

In each iterations between k and $2k$-1 records are grouped, on average $(3k-1)/2$ records. The algorithm will perform at most $2n/(3k-1)$ iterations. In each iteration, it needs to compute $2k$ nearest neighbors in the remaining records followed by extension of the cluster created k times. In each of the extension process k nearest neighbors need to be found in the remaining records of the dataset. If we assume that on average $n/2$ unassigned records remain in dataset, complexity of the algorithm will be $O(2n/(3k-1)(2kn/2+kkn/2))$ i.e. $O(kn^2)$.

4.2 Comparing variable-size MDAV algorithms

Fourth and fifth rows for each dataset in Table 1 present the results for the proposed variable-sized *IVMDAV* algorithm along with the results for the other algorithms we have implemented. It is clear from the table that *IVMDAV* performs better than *V-MDAV* producing lesser information loss. We have extended the *MDAV*-single group algorithm for developing the variable-size *IVMDAV* algorithm, so performance of the proposed algorithm should be compared to this algorithm also. It can be seen from Table 1 that the proposed algorithm produces lower information loss than the *MDAV-single-group* algorithm. In fact for the *EIA* as well as *Census* datasets the *IVMDAV* algorithm shows better results than any of the presented algorithms.

Variable-size algorithms show better performance for datasets with clustering tendency. This is the reason for achieving greater reduction of information loss for *EIA* dataset. *Tarragona* is a scattered dataset exhibiting no tendency for clustering, that is why reduction of information loss by proposed algorithm is very less for the *Tarragona* dataset. For the *V-MDAV* algorithm γ value is user specified. The results for this algorithm are presented in Table 1, taking γ=0.2 for *Tarragona* and *Census* datasets and γ=1.1 for the *EIA* dataset as these two values of γ are suggested by authors of *V-MDAV* in [12]. For the *IVMDAV* algorithm the value of γ is set to 1.16.

Table 1. Experimental results.

Datase	*Method*	*K=3* *SSE : (IL)*	*K=4* *SSE : (IL)*	*K=5* *SSE : (IL)*	*K=10* *SSE : (IL)*
Tarragona	**1.** *MDAV*	*1835.8318* *(16.9326)*	*2119.1678* *(19.545)*	*2435.2796* *(22.461)5*	*3598.7743* *(33.1929)*
	2. *MDAVsingle*	*1839.4617* *(16.9661)*	*2139.1554* *(19.7303)*	*2473.9951* *(22.8186)*	*3601.2138* *(33.2154)*
	3. *VMDAV*	*1839.6440* *(16.9678)*	*2135.5903* *(19.6974)*	*2481.3201* *(22.8862)*	*3607.2572* *(33.2711)*
	4. *IVMDAV* *(proposed)*	*1839.4739* *16.9662*	*2139.1554* *19.7303*	*2473.9951* *(22.8186)*	*3601.2138* *(33.2154)*
Census	**1.** *MDAV*	*799.1827 (5.6922)*	*1052.2557* *(7.4947)*	**1276.0162 (9.0884)**	*1987.4925* *(14.1559)*
	2. *MDAVsingle*	*793.7595 (5.6536)*	*1044.7749* *(7.4414)*	*1247.3171 (8.8840)*	*1966.5216* *(14.0066)*
	3. *VMDAV*	*794.9373 (5.6619)*	*1054.9675* *(7.5140)*	*1264.5801 (9.0070)*	*1975.8520* *(14.0730)*
	4. *IVMDAV* **(proposed)**	*791.2159 (5.6354)*	*1039.4388* *(7.4034)*	*1246.1519 8.8757*	*1965.0536* *13.9961*
EIA	**1.** *MDAV*	*217.3804 (0.4829)*	*302.1859* *(0.6713)*	*750.1957 (1.6667)*	*1728.3120 (3.8397)*
	2. *MDAVsingle*	*215.1095 (0.4779)*	*301.9676* *(0.6709)*	*783.0258 (1.7396)*	*1580.8008 (3.5120)*
	3. *VMDAV*	*229.2986 (0.5094)*	*437.8020* *(0.9726)*	*588.0341 (1.3064)*	*1264.4328 (2.8091)*
	4. *IVMDAV* **(proposed)**	*184.1079 (0.4090)*	*274.5894* *(0.6100)*	*412.3063 (0.9160)*	*1286.3228* *(2.8577)*

V. CONCLUSION AND FUTURE WORK

In this dissertation we have proposed an improved variable-size *MDAV* algorithm named *IVMDAV* that produces lower information loss with little increase in computational complexity ($O(kn^2)$). Fixed-size algorithms have complexity $O(n^2)$. This is acceptable as k is usually a small integer.

Proposed algorithm is a modification of the *MDAV* algorithm to make it variable-size. The algorithm computes $2k$ nearest neighbors of the farthest record from the centroid of the remaining unassigned records in the dataset. First k of the $2k$ neighbors form a cluster and it is extended up to a size of $2k$-1 records by including some of the remaining k neighbors based on a heuristic. The *IVMDAV* algorithm requires a user defined factor γ to be used for the cluster extension process. It can be easily determined as it need to be slightly greater than 1.0 (possible values in the range 1.0 – 1.20).

In future the following considerations can be made to further improve the algorithm. To form a single cluster $2k$ nearest neighbors of the currently selected record for cluster formation is considered. It is possible to consider $3k$ neighbors instead of $2k$ as the algorithm iterates so long as there are at least $3k$ neighbors yet to be assigned to any

cluster. This will increase computation time slightly while producing better results as more records are considered for inclusion in the cluster extension. Another possibility for modification of the algorithm is to test whether the current record considered for group formation i.e. the furthest record from the centroid of the remaining records in the dataset is a outlier or not. If it is a outlier than the group formed by the record will remain as a group of k records and it should not be extended to contain up to 2k-1 records

REFERENCES

[1] Agrawal R., Srikant "R. Privacy-Preserving Data Mining". *ACM SIGMOD Conference*,2000.

[2] CHARU C. AGGARWAL and PHILIP S. YU "PRIVACY-PRESERVING DATA MINING:MODELS AND ALGORITHMS"

[3] Sweeney L.: Replacing Personally Identifiable Information in Medical Records, the Scrub System. *Journal of the American Medical Informatics Association*, 1996.

[4] Sweeney L.: Guaranteeing Anonymity while Sharing Data, the Datafly System. *Journal of the American Medical Informatics Association*,1997.

[5] J.M. Mateo-Sanz and J. Domingo-Ferrer, "A Method for Data Oriented Multivariate Microaggregation," *Proc. Statistical Data Protection'* 98,pp. 89-99,1999.

[6] A. Hundepool, A. V. deWetering, R. Ramaswamy, L. Franconi, A. Capobianchi, P.-P. DeWolf, J.Domingo-Ferrer, V. Torra, R. Brand & S. Giessing, (2003) "µ-ARGUS version 3.2 Software and User's Manual", Voorburg NL: Statistics Netherlands, http://neon.vb.cbs.nl/casc.

[7] M. Laszlo & S. Mukherjee, (2005) "Minimum spanning tree partitioning algorithm for microaggregation", IEEE Transactions on Knowledge and Data Engineering, 17(7), pp. 902-911.

[8] Domingo-Ferrer J, Mateo-Sanz J., Practical data-oriented microaggregation for statistical disclosure control. *IEEE Transactions on Knowledge and Data Engineering* 2002; 14(1):189–201

[9] Malin B., Sweeney L.: Determining the identifiability of DNA database entries. *Journal of the American Medical Informatics Association*, pp. 537–541, November 2000

[10] Laszlo, M., Mukherjee, S.: Minimum spanning tree partitioning algorithm for microaggregation. IEEETrans. Knowl. Data Eng. 17(7), 902–911 (2005)

[11] J. Domingo-Ferrer and V. Torra, "Ordinal, continuous and heterogenerous *k*-anonymity through microaggregation," *Data Mining and Knowledge Discovery*, vol. 11, no. 2, pp. 195–212, 2005.

[12] A. Solanas & A. Mart´ınez-Ballest´e, (2006) "V-MDAV: A multivariate microaggregation with variable group size", Seventh COMPSTAT Symposium of the IASC, Rome.

[13] J. Domingo-Ferrer, A. Solanas & A. Mat´_nez-Ballest´e, 2006 "Privacy in statistical databases: kanonymity through microaggregation", in IEEE Granular Computing' 06. Atlanta. USA, pp. 774-777. 118 Computer Science & Information Technology (CS & IT).

[14] Newton E., Sweeney L.,Malin B.: Preserving Privacy by De-identifying Facial Images. *IEEE Transactions on Knowledge and Data Engineering, IEEE TKDE*, February 2005.

[15] A. Hundepool, A. V. deWetering, R. Ramaswamy, L. Franconi, A. Capobianchi, P.-P. DeWolf, J. Domingo-Ferrer, V. Torra, R. Brand, and S. Giessing, *µ-ARGUS version 4.0 Software and User's Manual*. Voorburg NL: Statistics Netherlands, May 2005, http://neon.vb.cbs.nl/casc.

[16] Sweeney L.: Privacy-Preserving Bio-terrorism Surveillance. *AAAI Spring Symposium, AI Technologies for Homeland Security*, 2005

[17] Sweeney L.: Privacy Technologies for Homeland Security. *Testimony before the Privacy and Integrity Advisory Committee of the Deprtment of Homeland Scurity*, Boston, MA, June 15, 2005

[18] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam "*l*-diversity: Privacy beyond *k*-anonymit", In *Proceedings of the 22nd IEEE International Conference on Data Engineering (ICDE 2006)*, 2006

[19] Solanas A, Mart´ınez-Ballest´e A. V-MDAV: A multivariate microaggregation with variable group size. *Seventh COMPSTAT Symposium of the IASC*, Rome, 2006

[20] Sweeney L.: AI Technologies to Defeat Identity Theft Vulnerabilities. *AAAI Spring Symposium, AI Technologies for Homeland Security*, 2005

[21]Benjamin C. M. Fung *Concordia University, Montreal*, Rui Chen *Simon Fraser University, Burnaby* and Philip S. Yu *University of Illinois at Chicago"* Privacy-Preserving Data Publishing: A Survey of Recent Developments" ACM Computing Surveys, Vol. 42, No. 4, Article 14, Publication date: June 2010

[22] Privacy-Preserving Data Mining, *Models and Algorithms* Edited by Charu C. Aggarwal *IBM T.J. Watson Research Center, USA* and Philip S. Yu *University of Illinois at Chicago, USA, Springer 2008*

[23] Ebaa Fayyoumi and B. John Oommen " A survey on statistical disclosure control and micro-aggregation techniques for secure statistical databases." *Softw. Pract. Exper.* 31 May 2010;

[24] P. Samarati and L. Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical report, CMU, SRI, 1998.

[25] Josep domingo-Ferrer, Agusti Solanas. "Privacy in Statistical Databases:k-Anonymity Through Microaggregation,"IEEE 2006

[26] Domingo-Ferrer, J., Sebé, F., & Solanas, A. (2008). A polynomial-time approximation to optimal multivariate microaggregation. Computer and Mathematics witt Applications, 55(4), 714–732.

[27] Chang, C.-C., Li, Y.-C., & Huang, W.-H. (2007). TFRP: An efficient microaggregation algorithm for statistical disclosure control. Journal of Systems and Software, 80(11), 1866–1878.

[28] Ebaa Fayyoumi and B. John Oommen "A survey on statistical disclosure control and micro-aggregation techniques for secure statistical databases" Published online in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/spe.992

[29] Sweeney L., Gross R.: Mining Images in Publicly-Available Cameras for Homeland Security. *AAAI Spring Symposium, AI Technologies for Homeland Security*, 2005.

[30] Domingo-Ferrer, J., Martínez-Ballesté, A., Mateo-Sanz, J. M., & Sebé, F. Efficient multivariate data-oriented microaggregation. The VLDB Journal, 15(4), 355–369. (2006)