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. 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 abstract into classes of similar objects is called clustering. A cluster of data objects can be treated collectively as one abstract into classes of similar objects is called clustering. A cluster of data objects can be treated collectively as one abstract into classes of similar objects is called clustering.

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 y value is user specified. The results for this algorithm are presented in Table 1, taking y=0.2 for Tarragona and Census datasets and y=1.1 for the EIA dataset as these two values of y are suggested by authors of V-MDAV in [12]. For the IVMDAV algorithm the value of y 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=LXl;
2. while (n>=3k) do
2.1 compute centroid x of remaining records in X;
2.2 find the most distant record x_i from x ;
2.3 find 2k nearest neighbors y_1,y_2,...,y_2k of x_i;
2.4 form cluster c_i with first k-neighbors y_1,y_2,...,y_k;
2.5 remove records y_1,y_2,...,y_k from dataset X;
2.6 set n = n - k;
2.7 compute centroid 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 \(x_i\);
2.8.3 find distance \(d_2\) of record \(y_j\) from \(z_i\);
2.8.4 if \((d_1 > y_{d_2})\) then
2.8.4.1 insert \(y_j\) in current cluster \(c_i\);
2.8.4.2 recomputed centroid \(x_i\) 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 \(x\) of remaining records in \(X\);
3.2 find the most distant record \(x_i\) from \(x\);
3.3 find \(k\) nearest neighbors \(y_{1}, y_{2}, \ldots, y_{k}\) of \(x_i\);
3.4 form cluster \(c\) 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 \(y_{1}, y_{2}, \ldots, y_{k}\) of the farthest record \(x_i\), from the centroid \(x\) 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 \(x_i\). 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 \(y\) whose value is slightly greater than 1.0 (in the range 1.0 - 1.20). Let, \(x_i\) be the centroid of the cluster \(c_i\). Consider the \((k+1)\)-th neighbor, \(y_{k+1}\) of \(x_i\). Let \(z_{1}, z_{2}, \ldots, z_k\) be the \(k\)-nearest unassigned neighbors of \(y_{k+1}\). Find distance \(d_1\) of \(y_{k+1}\) from \(x_i\). Find distance \(d_2\) of \(y_{k+1}\) from furthest neighbor \(z_k\). Now, if \(d_2 > y_{d_2}\) then insert \(y_{k+1}\) in cluster \(c_i\) and recomputed the centroid of the cluster. Then the test is repeated for \(y_{k+2}, y_{k+3}, \ldots, y_{2k}\). 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 \(n_2n\) unassigned records remain in dataset, complexity of the algorithm will be \(O(2n/(3k-1) (2kn/2k+2kn/2k))\) i.e. \(O(n^3)\).

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

### Table 1. Experimental results.

<table>
<thead>
<tr>
<th>Database</th>
<th>Method</th>
<th>(K=3) SSE: (IL)</th>
<th>(K=4) SSE: (IL)</th>
<th>(K=5) SSE: (IL)</th>
<th>(K=10) SSE: (IL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tarragona</td>
<td>1. MDAV</td>
<td>1835.8318</td>
<td>2119.1678</td>
<td>2435.2796</td>
<td>3598.7743</td>
</tr>
<tr>
<td></td>
<td>2. MDAVsingle</td>
<td>1839.4617</td>
<td>2139.1554</td>
<td>2473.9951</td>
<td>3601.2138</td>
</tr>
<tr>
<td></td>
<td>3. VMDA</td>
<td>1839.6440</td>
<td>2135.5903</td>
<td>2481.3201</td>
<td>3607.2572</td>
</tr>
<tr>
<td></td>
<td>4. IVMDAV</td>
<td>1839.4739</td>
<td>2139.1554</td>
<td>2473.9951</td>
<td>3601.2138</td>
</tr>
<tr>
<td></td>
<td>(proposed)</td>
<td>16.9662</td>
<td>19.7303</td>
<td>22.8186</td>
<td>33.2154</td>
</tr>
<tr>
<td>Census</td>
<td>1. MDAV</td>
<td>799.1827</td>
<td>1052.2557</td>
<td>1276.0162</td>
<td>1987.4925</td>
</tr>
<tr>
<td></td>
<td>2. MDAVsingle</td>
<td>793.7539</td>
<td>1044.7749</td>
<td>1247.3171</td>
<td>1966.5216</td>
</tr>
<tr>
<td></td>
<td>3. VMDA</td>
<td>794.9373</td>
<td>1054.9675</td>
<td>1264.4901</td>
<td>1975.8520</td>
</tr>
<tr>
<td></td>
<td>4. IVMDAV</td>
<td>791.2159</td>
<td>1039.4388</td>
<td>1246.1519</td>
<td>1965.0536</td>
</tr>
<tr>
<td></td>
<td>(proposed)</td>
<td>5.6354</td>
<td>7.4043</td>
<td>8.8757</td>
<td>13.9961</td>
</tr>
<tr>
<td>EIA</td>
<td>1. MDAV</td>
<td>217.3804</td>
<td>302.1859</td>
<td>750.1957</td>
<td>1728.3120</td>
</tr>
<tr>
<td></td>
<td>2. MDAVsingle</td>
<td>215.1095</td>
<td>301.9676</td>
<td>783.0258</td>
<td>1580.8008</td>
</tr>
<tr>
<td></td>
<td>3. VMDA</td>
<td>229.2986</td>
<td>437.8020</td>
<td>588.0341</td>
<td>1264.4328</td>
</tr>
<tr>
<td></td>
<td>4. IVMDAV</td>
<td>184.1079</td>
<td>274.5894</td>
<td>412.3063</td>
<td>1286.3228</td>
</tr>
<tr>
<td></td>
<td>(proposed)</td>
<td>0.4090</td>
<td>0.6100</td>
<td>0.9160</td>
<td>2.8577</td>
</tr>
</tbody>
</table>

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 \((O(n^2))\). 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 $y$ 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 an 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


