Importance partitioning in micro-aggregation

G. Kokolakis, D. Fouskakis *
National Technical University of Athens, Department of Mathematics, Zografou Campus, Athens 15780, Greece

ARTICLE INFO

Article history:
Received 13 March 2007
Received in revised form 23 September 2008
Accepted 24 September 2008
Available online 4 October 2008

ABSTRACT

One of the techniques of data holders for the protection of confidentiality of continuous data is that of micro-aggregation. Rather than releasing raw data (individual records), micro-aggregation releases the averages of small groups and thus reduces the risk of identity disclosure. At the same time the method implies loss of information and often distorts the data. Thus, the choice of groups is very crucial to minimize the information loss and the data distortion. No exact polynomial algorithms exist up to date for optimal micro-aggregation, and so the usage of heuristic methods is necessary. A heuristic algorithm, based on the notion of importance partitioning, is proposed and it is shown that compared with other micro-aggregation heuristics achieves improved performance.

© 2008 Elsevier B.V. All rights reserved.

1. Introduction

One of the most important tasks of data holders, such as medical institutions, public health agencies, financial organizations, or statistical institutions, when disseminating information for public use, is to prevent any possibility of direct or indirect identification of the individuals behind the records. Their aim is to protect the records from what in Duncan and Lambert (1989) has been called the risk of identity disclosure. Frequently, this has negative implications in pursuing accurate statistical analysis or research studies for decision support purposes, especially in the area of social science research and public policy, cf. Duncan and Pearson (1991) and Fienberg (1994).

To resolve the problem of identity protection, several statistical techniques have been developed. Under the general terminology disclosure limiting techniques, or confidentiality protection techniques, these methods can be divided into two main categories; those for categorical and those for continuous variables. For an extensive presentation of statistical methods on the problem of Disclosure Control and Statistical Data Protection see Paass (1988), Panda and Nagabhushanam (1995), Willenborg and De Waal (1996), Willenborg and De Waal (2001) and Gosh et al. (2007).

Among the most recent methods for categorical variables are the following: The Post RAndomization Method (PRAM), cf. De Wolf et al. (1999) and the k-anonymity technique, cf. Sweeney (2002). In the first one, the scores on some variables for certain records are changed into different ones according to a specific probability Markov matrix, whose choice is crucial on disclosure risk limitation and on preservation of the statistical information content. In the second one, recoding or suppression is performed in such a way that every record has to be indistinguishable from at least $(k - 1)$ other records. In recoding, a category is replaced with a less specific but semantically consistent one, while in suppression a category is not released at all.

For continuous variables among the most common techniques are the grouping method, cf. Spruill (1983), noise introduction, cf. Kim (1986) and micro-aggregation, cf. Defays and Nanopoulos (1993). In the first method, original continuous variables are replaced by categories. This reduces disclosure risk, but at the same time causes information loss, in respect to both univariate and multivariate shapes and structures. In the second method an error term with mean zero

* Corresponding author. Tel.: +30 210 772 1702; fax: +30 210 772 1705.
E-mail addresses: kokolakis@math.ntua.gr (G. Kokolakis), fouskakis@math.ntua.gr (D. Fouskakis).

0167-9473/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
and known variance is added to each response variable in all records. These methods have been criticized for introducing bias and considerable data distortion, cf. Adam and Wortmann (1989).

In the method of micro-aggregation, in which we are focusing in this paper, small groups of “similar” records are formed and, instead of releasing the raw values, the mean of the group they belong is reported in their place. Ensuring that each group has at least a minimum number of observations \( k \), resulting to a partition known as \( k \)-partition, the confidentiality is protected at a satisfactory level and in addition, using an appropriate measure of similarity of the data within a group, distortion can be avoided. Nevertheless, this implies loss of information. Thus, it is crucial to choose groups, in such way that the statistical information content of the data is kept as high as possible.

The micro-aggregation problem can be divided into univariate or multivariate according to the dimension of the data involved. The problem becomes easier in the case of a single variable, where one can use the individual ranking idea, cf. Domingo-Ferrer and Mateo-Sanz (2002). Furthermore, in the univariate case, a polynomial-time optimal algorithm is given in Hansen and Mukherjee (2003). When there is more than one variable, one could project the data onto a single axis and perform the univariate ranking technique there (projected data method). This method implies considerable data distortion and massive loss of information regarding the data correlation structure. Thus, only simultaneous single-axis and perform the univariate ranking technique there (projected data method). This method implies considerable data distortion and massive loss of information regarding the data correlation structure. Thus, only simultaneous micro-aggregation on all variables, i.e., without reducing the problem to a lower dimension, would perform well (unprojected data method), and this is the case we consider here.

We have to mention that we are not focusing on the problem of assessing the individual risk of disclosure, but this can be nevertheless controlled by increasing the group size. For more information on this topic see Bethlehem et al. (1990), Fienberg and Makov (1998), Fuller (1993), Omori (1999) and Samuels (1998). A Bayesian method of optimal multivariate micro-aggregation by minimizing probabilistic distance criteria is also available in the literature, cf. Kokolakis and Nanopoulos (2001). Their method refers to multivariate normal data and provides a \( k \)-partition defined by hypercubes in \( \mathbb{R}^d \) formed by hyperplanes in \( \mathbb{R}^{d-1} \) orthogonal to the eigenvectors of the data covariance matrix. Finally, a more general class of functions measuring the similarity of the data within a group is provided by Kokolakis et al. (2006).

In Oganian and Domingo-Ferrer (2001), it was shown that, for multivariate data, optimal micro-aggregation is an NP-hard problem. Funding an exact solution, therefore, is computationally infeasible for very large data-sets. Since no exact optimal polynomial algorithms are known to date to micro-aggregate optimally, we introduce a new heuristic method of multivariate fixed-size micro-aggregation, with \( O(n^2) \) complexity, based on the notions of importance and dissimilarity measures. Section 2 introduces basic micro-aggregation concepts together with the necessary terminology. In Section 3 some mathematical results are derived on which the proposed multivariate micro-aggregation method is based. Section 4 presents experimental results and assesses the effectiveness of the derived algorithm. Finally in Section 5, conclusions are drawn.

2. The micro-aggregation problem

Micro-aggregation is a technique proposed by Defays and Nanopoulos (1993), in order to protect confidentiality without losing valuable information contained in the individual records. The idea is to cluster “similar” raw data into groups of small size and replace their initial values by the average value of the group. It is important to keep the number of individuals in each group not too large, in order to have low information loss, but also not too small, to prevent the identification of the individuals (in many countries the rules do not permit the publication of data when the number of individuals in each group is less than 3; therefore groups of sizes of \( k_0 = 3 \) or 4 are usually preferred). Thus, this partition problem differs from the classical clustering problem, cf. Hartigan (1975), whose aim is to identify “dissimilar” groups of any size.

Consider a data-set with \( d \)-dimensional continuous variables on \( n \) individuals. With these individual records, \( m \) groups are formed with \( k_i \) records in the \( i \)th group, where, according to the previous paragraph, \( k_i \geq k_0 \) and not too large for every \( i = 1, \ldots, m \), with \( \sum_{i=1}^{m} k_i = n \). Thus, we have a \( k_0 \)-partition \( G \) consisting of the groups \( G_i = \{ x_{ij}, j = 1, \ldots, k_i \} \) for \( i = 1, \ldots, m \). If \( \mathcal{G} \) is the family of all possible \( k_0 \)-partitions, our aim is to find the partition \( G^* \in \mathcal{G} \) that minimizes an appropriate criterion.

Let \( \bar{x}_i \) be the average data vector over the the \( i \)th group and \( \bar{x} \) the total mean of the whole data-set. The optimality criterion we consider here is based on the Euclidean distance. Thus, the optimal \( k_0 \)-partition is the one that minimizes over \( \mathcal{G} \) the within-group sum of squares, namely

\[
SSW = \sum_{i=1}^{m} \sum_{j=1}^{k_i} \| x_{ij} - \bar{x}_i \|^2 = \sum_{i=1}^{m} \sum_{j=1}^{k_i} \langle x_{ij} - \bar{x}_i, x_{ij} - \bar{x}_i \rangle,
\]

where \( \langle \cdot, \cdot \rangle \) denotes the standard inner product in \( \mathbb{R}^d \).

Alternatively, the optimal \( k \)-partition is the one that maximizes over \( \mathcal{G} \) the between-group sum of squares, namely

\[
SSB = \sum_{i=1}^{m} k_i \| \bar{x}_i - \bar{x} \|^2 = \sum_{i=1}^{m} k_i \langle \bar{x}_i - \bar{x}, \bar{x}_i - \bar{x} \rangle,
\]

since the total sum of squares \( \text{SST} \) is analyzed as follows:

\[
\text{SST} = \text{SSW} + \text{SSB} = \sum_{i=1}^{n} \| x_i - \bar{x} \|^2 = \sum_{i=1}^{n} \langle x_i - \bar{x}, x_i - \bar{x} \rangle.
\]
Finally the measure of information loss standardized between 0 and 1, can be provided by the quantity:

\[ L = \frac{SSW}{SST}. \]  
(4)

Thus, the optimal partition is the one that minimizes \( L \) over \( \mathcal{G} \).

The more dissimilar the data within the group \( G \) are, the larger the quantity \( D_i \equiv \sum_{j=1}^{k} ||x_{ij} - \bar{x}_i||^2 \) becomes, and thus it is called the dissimilarity measure of the group \( G \). Similarly, the quantity \( l_i \equiv ||\bar{x}_i - \bar{x}||^2 \) specifies the average, per observation within the group \( G \), importance of the group \( G \), since the larger it is, the highest the reduction implied by the group \( G \) on the SST, and thus it is called the importance measure of the group \( G \). When dealing with continuous data in \( \mathbb{R}^d \), a strict ordering of all groups is implied by the importunity measure \( I \). In other words, we have a ranking of individual groups in an analogous way we have a ranking of individual data in the univariate case. Thus, with \( SSW = \sum_{i=1}^{m} D_i \) and \( SSB = \sum_{i=1}^{m} k_i l_i \), the total sum of squares can be written as follows:

\[ SST = \sum_{i=1}^{m} (D_i + k_i l_i). \]  
(5)

Consequently, the optimal partition is the one that minimizes over \( \mathcal{G} \) the quantity \( \sum_{i=1}^{m} D_i \), or equivalently maximizes \( \sum_{i=1}^{m} k_i l_i \).

The pairs \((D_i, l_i), i = 1, \ldots, m\), will have a key role in the new algorithm proposed.

3. Multivariate fixed-size micro-aggregation

As mentioned in the previous paragraph, our aim is to find the optimal partition \( G^* \in \mathcal{G} \) that minimizes the information loss \( L \) defined in (4). We are focusing here on the fixed-size micro-aggregation problem, i.e., the groups of any partition in \( \mathcal{G} \) have equal size \( k \), with \( n \) a multiple of \( k \). In this case, the number of partitions is \( N = n!/(k!^m \times m!) \) and therefore full search over the whole \( \mathcal{G} \) is in most cases analytically and computationally infeasible. For example, with \( n = 1000, m = 300 \) and \( k = 3 \), this requires \( N = 10^{4422} \) comparisons. Thus, the use of a heuristic optimization algorithm is necessary.

The procedure we propose in this paper is based on a sequential optimization algorithm. At the first stage our aim is to produce a “good” group of \( k \) elements among all two-group partitions of \( k \) and \((n-k)\) elements respectively. Such a group should be close to the optimal group \( G \) of \( k \) elements, i.e., corresponding to the optimal partitioning of the initial data-set into two groups of \( k \) and \((n-k)\) elements respectively, and at the same time should have a very low dissimilarity measure \( D \), in order to be close to a group that belongs to the optimal \( G^* \). Then, on the \((n-k)\) remaining elements we repeat the above procedure in order to create a new “good” group of \( k \) elements, and so on until the exhaustion of the data-set.

The final partition created by the above technique will not necessarily be the optimal \( G^* \), but we expect it to be a sub-optimal solution, based on the applied procedure.

The next theorem characterizes the optimal two-group partitioning, namely \((G, G^*)\).

**Theorem 1.** The optimal two-group partitioning of \( k \) and \((n-k)\) elements is provided by maximizing the quantity:

\[ I(G) = ||\bar{x}_G - \bar{x}||^2. \]  
(6)

**Proof.** From (5) with only two groups \( G \) and \( G^* \) of \( k \) and \((n-k)\) elements respectively we have:

\[ SST = D(G) + D(G^*) + k l(G) + (n-k) l(G^*), \]  
(7)

where

\[ D(G) = \sum_{x_i \in G} ||x_i - \bar{x}_G||^2, \quad D(G^*) = \sum_{x_i \in G^*} ||x_i - \bar{x}_{G^*}||^2, \]  
(8)

and

\[ I(G) = ||\bar{x}_G - \bar{x}||^2, \quad I(G^*) = ||\bar{x}_{G^*} - \bar{x}||^2. \]  
(9)

Therefore we have the equivalence:

\[ \min_G \left\{ \sum_{x_i \in G} ||x_i - \bar{x}_G||^2 + \sum_{x_i \in G^*} ||x_i - \bar{x}_{G^*}||^2 \right\} \Leftrightarrow \max_G \left\{ k ||\bar{x}_G - \bar{x}||^2 + (n-k) ||\bar{x}_{G^*} - \bar{x}||^2 \right\}. \]  
(10)

Since

\[ k(\bar{x}_G - \bar{x}) + (n-k)(\bar{x}_{G^*} - \bar{x}) = 0, \]  
(11)
we have
\[ k\|\bar{x}_G - \bar{x}\|^2 + (n - k)\|\bar{x}_G - \bar{x}\|^2 = \frac{nk}{n-k} \|\bar{x}_G - \bar{x}\|^2. \] (12)

and therefore
\[ \text{SST} = D(G) + D(G') + \frac{nk}{n-k} I(G). \] (13)

Thus, the optimal partition is provided by maximizing the quantity:
\[ I(G) = \|\bar{x}_G - \bar{x}\|^2. \]

**Corollary 1.** The optimal two-group partitioning of \( k \) and \( (n - k) \) elements is provided by minimizing the criterion:
\[ R(G) \equiv \frac{D(G) + D(G')}{I(G)}. \] (14)

Our method starts from the most extreme point \( x_1 \) from the total mean \( \bar{x} \), and creates a group around \( x_1 \) with a high similarity measure, by introducing sequentially a new point closer to the mean of those already introduced. The fact that this group has a high similarity is based on the following argument. Suppose we have a group with \( v \) vectors \( x_l, l = 1, \ldots, v \), with \( 1 \leq v < k \) and group-mean \( \bar{x}_G(v) \), and we want to introduce a new one \( x_{v+1} \) into the group. Using the properties of the inner product, it can be easily proved that the within group sum of squares can be expressed as follows:
\[ \sum_{i=1}^{v+1} \|x_i - \bar{x}_{(u+1)}\|^2 = \sum_{i=1}^{v} \|x_i - \bar{x}_{(v)}\|^2 + \frac{v}{v+1} \|x_{v+1} - \bar{x}_{(v)}\|^2. \] (15)

Thus, by choosing the new vector \( x_{v+1} \) to be the closest one to the group mean \( \bar{x}_{(v)} \) of \( v \) initial points, we have the lowest possible augmentation of the within group sum of squares and therefore the lowest possible group dissimilarity measure \( D \), for the augmented data-set. It can be mentioned here that the above result can also be taken from expression (13) with \( n = v + 1 \) and \( k = 1 \).

Thus, the resulting group is expected to have very low dissimilarity measure and at the same time very high importance measure, since we started from the most extreme point \( x_1 \).

We have to mention here that if, alternatively, at each step, we had created groups around the most distant points, this would restrict the algorithm to a far smaller number of possibilities than necessary, and, in fact, could be eliminating the optimal choice. In addition if, at each step, we had created the groups with points that are closer to the extremes \( x_1 \), this would result to higher importance measures, but at the same time, to higher dissimilarity measures and the final algorithm would have a stronger dependence on the extreme points rather than the more robust group means.

The following theorem explains the logic behind our proposed methodology. For notational convenience let \( X_0 \) denote the initial data-set and \( X_i \) the remaining data-set after removing the first \( i \) groups of an arbitrary \( k \)-partition \( G \), with \( i = 1, \ldots, m - 1 \), i.e., \( X_i = X_0 \setminus \bigcup_{j=1}^{i} G_j \). Let also \( \text{SST}(X_i) \) be the total sum of squares of the group \( X_i \).

**Theorem 2.** The following recursive relation holds:
\[ \frac{\text{SST}(X_{j-1}) - \text{SST}(X_j)}{I(G_j)} = \frac{D(G_j)}{I(G_j)} + c_j, \quad j = 1, \ldots, m - 1, \] (16)

with \( c_j = k(1 + \frac{k}{n_j}), n_j = n_{j-1} - k \) and \( n_0 = n \).

**Proof.** For \( j = 1 \) the above result is derived from (13) by realizing that \( D(G') = \text{SST}(X_1) \). By deduction, we then get (16).

**Remark 1.** The left hand side of expression (16) represents the relative reduction of SST at the \( j \)-th stage. Since, as we proceed, the number of possible choices decreases, this reduction should be kept as low as possible at each step. This can be achieved by requiring \( D(G_j) \) to be the smallest possible for a given \( I(G_j) \).

Our proposed new algorithm is the following:

**Algorithm** (*Importance Partitioning (IP)*).
- Choose the most distant point \( x_1 \) from the total mean \( \bar{x} \).
- Construct the first group by finding, initially, the closest point to \( x_1 \), and then introduce each time a new one that is closer to the mean of the already selected points, until a group of \( k \) points is formed.
Table 1
Comparison of information loss between the UF and IP algorithms, for four different group sizes. In the first part of the table we have 880 data points of dimension 6 and in the second part we have 640 data points of dimension 8.

<table>
<thead>
<tr>
<th>Group size, k</th>
<th>Information loss L</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>UF</td>
<td>IP</td>
<td></td>
</tr>
<tr>
<td>(N = 880, d = 6)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.0906</td>
<td>0.0899</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.1238</td>
<td>0.1212</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.1526</td>
<td>0.1456</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.1706</td>
<td>0.1681</td>
<td></td>
</tr>
<tr>
<td>(N = 640, d = 8)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.1326</td>
<td>0.1314</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.1732</td>
<td>0.1698</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.2067</td>
<td>0.2014</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.2304</td>
<td>0.2300</td>
<td></td>
</tr>
</tbody>
</table>

- Exclude this group from the data-set. If they remain more than \(k\) data points, find the new overall mean and repeat Step 1 and Step 2.
- If they remain \(k\) data points, these form the last group.
- If they remain less than \(k\) data points, introduce each one to its closest group.
- End.

It is useful to mention the following corollary which refers to the optimality of the above algorithm in the one dimensional case.

**Corollary 2.** For Univariate data-sets with \(n = mk\) the Importance Partitioning algorithm is optimal.

**Proof.** With univariate data and \(n = mk\) the importance measure in (6) produces the optimal partition, since the resulting groups \(G_i, i = 1, \ldots, m\), form non-overlapping intervals and thus imply a minimum SSW. \(\blacksquare\)

4. Experimental results

Initially, we implemented our algorithm (IP) together with the one proposed by Domingo-Ferrer and Mateo-Sanz for the multivariate fixed-size micro-aggregation problem on unprojected data (UF), cf. Domingo-Ferrer and Mateo-Sanz (2002), and we compared their performances. A real data-set of 1000 households in Greece was used, based on the “1998’s Household Budget Continuous Survey (HBCS)” collected by the National Statistical Service of Greece (NSSG). For each household 12 quantitative variables were collected, representing expenses for: Food and Non-Alcoholic Beverages (E1), Alcoholic Beverages and Tobacco (E2), Clothing and Footwear (E3), Housing (E4), Furniture and Household Equipment (E5), Health (E6), Transport (E7), Communications (E8), Recreation and Culture (E9), Education (E10), Hotels, Cafés and Restaurants (E11) and Miscellaneous (E12).

We used the information loss \(L\) in (4) as a measure of comparison, for fixed group sizes of \(k = 3, 4, 5\) and \(6\). Since there were missing data in some of the variables, we performed two different comparisons, one with 880 households and 6 variables, namely (E1), (E4), (E5), (E8), (E9) and (E12), and another one with 640 households and 8 variables, namely (E1), (E3), (E4), (E5), (E8), (E9), (E11) and (E12).

The proposed algorithm was implemented in C under Linux on a Pentium Celeron machine with 3.66 GHz of CPU speed and 1 GB of RAM.

Table 1 gives the resulting information loss in each case. The results indicate that our algorithm outperforms that of UF in all cases examined. Additionally, the CPU time of IP was much smaller than that of UF, due to a lower computational complexity. The complexity of IP is \(O(n^2)\) while that of UF is \(O(n^3)\), making the former much faster. Actually, the differences between the two algorithms’ CPU times can be very substantial when dealing with very large data-sets.

In order to check the effectiveness of the proposed algorithm, we also performed a series of comparisons with other micro-aggregation heuristics, using three reference data-sets from the European project CASC (Brand et al. 2002). We have considered the “Tarragona” data-set that contains 834 records and 13 quantitative variables, the “Census” data-set which contains 1080 records and 13 quantitative variables, and the “EIA” data-set with 4092 records and 11 quantitative variables. Comparisons were made with the following micro-aggregation algorithms:

- An improved version of UF, called Maximum Distance to Average Point (MDAV), cf. Domingo-Ferrer and Torra (2005).
- The centroid-based fixed-size micro-aggregation (CBFS) algorithm and the minimum spanning tree-based heuristic (M-d), for both see Laszlo and Mukherjee (2005).
- The maximum distance algorithm (MD), cf. Domingo-Ferrer and Mateo-Sanz (2002).
- The NPN-MHM, MD-MHM, MDAV-MHM and CBFS-MHM heuristics presented in Domingo-Ferrer et al. (2006).
Table 2
Comparison of information loss ($L$) for three data-sets and various microaggregation heuristics.

<table>
<thead>
<tr>
<th>Method</th>
<th>Tarragona data-set ($n = 834, \ d = 13$)</th>
<th>Census data-set ($n = 1080, \ d = 13$)</th>
<th>EIA data-set ($n = 4092, \ d = 11$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$k = 3$</td>
<td>$k = 5$</td>
<td>$k = 3$</td>
</tr>
<tr>
<td></td>
<td>$L$</td>
<td>$L$</td>
<td>$L$</td>
</tr>
<tr>
<td>IP</td>
<td>0.1561</td>
<td>0.2245</td>
<td>0.0534</td>
</tr>
<tr>
<td>NPN-MHM</td>
<td>0.1739</td>
<td>0.2702</td>
<td>0.0634</td>
</tr>
<tr>
<td>MD</td>
<td>0.1698</td>
<td>0.2252</td>
<td>0.0571</td>
</tr>
<tr>
<td>MD-MHM</td>
<td>0.1698</td>
<td>0.2252</td>
<td>0.0569</td>
</tr>
<tr>
<td>MDAV</td>
<td>0.1693</td>
<td>0.2246</td>
<td>0.0565</td>
</tr>
<tr>
<td>MDAV-MHM</td>
<td>0.1693</td>
<td>0.2246</td>
<td>0.0565</td>
</tr>
<tr>
<td>CBFS</td>
<td>0.1697</td>
<td>0.2282</td>
<td>0.0568</td>
</tr>
<tr>
<td>CBFS-MHM</td>
<td>0.1697</td>
<td>0.2282</td>
<td>0.0567</td>
</tr>
<tr>
<td>M-d</td>
<td>0.1663</td>
<td>0.2450</td>
<td>0.0611</td>
</tr>
<tr>
<td>$\mu$-Approx</td>
<td>0.1710</td>
<td>0.2604</td>
<td>0.0625</td>
</tr>
</tbody>
</table>

All the above algorithms have $O(n^2)$ complexity except of MD and MD-MHM that have $O(n^3)$. We used the information loss $L$ as a measure of comparison, for fixed group sizes of $k = 3$ and $k = 5$. Table 2 gives the resulting information loss in each case for all three data-sets. Comparable results for the algorithms CBFS-MHM and M-d are not available with the “EIA” data-set. In all the trials of the “Tarragona” and “Census” data-sets, IP outperformed its competitors, showing its advantages and usefulness. Regarding the CPU time, IP was faster than MD and MD-MHM, and ran on approximately the same CPU time with the rest of the algorithms.

The third reference data-set (“EIA”), is a non-homogeneous data-set, with clustered records. In such cases, heuristics producing varying-size partitionings could be more appropriate. For $k = 3$, all the algorithms produced very similar results, with MDAV-MHM to be the best performer. As $k$ increases fixed-size micro-aggregation techniques become less appropriate, especially for data-sets with small-size clusters. When $k = 5$, $\mu$-Approx clearly outperformed its competitors, followed by NPN-MHM, MDAV-MHM, MD-MHM and IP, indicating the existence of small-size clusters. As a topic of further research, it will be interesting to extend IP to the more difficult problem of non fixed-size groups. Alternatively, one can apply the IP algorithm on each cluster separately.

5. Conclusions

No exact polynomial algorithms have been published up to date for optimal micro-aggregation. In this paper, we focused on the case of multivariate unprojected data, and we worked with groups of equal size. We proposed a heuristic based on importance grouping and compare its usefulness with known micro-aggregation heuristics of the literature on four different data-sets, three of which are widely used as reference data-sets. We found that our new micro-aggregation heuristic achieves substantial improvements for data-sets without small-size natural clusters.

Acknowledgments

We are grateful to the National Statistical Service of Greece for making the data available to us, to Dr. Katerina Potika and to Mr. George Kouvaras for helping with the programming of the algorithms. Furthermore, we are very thankful to the referees for their helpful and valuable comments.

References