V-MDAV: A Multivariate Microaggregation With Variable Group Size

Agusti Solanas and Antoni Martínez-Ballesté

CRISES Research Group. Department of Computer Engineering and Mathematics.
Rovira i Virgili University. Av.Països Catalans 26. 43007 Tarragona. Catalonia.
Spain. {agusti.solanas,antoni.martinez}@urv.net

Summary. Microaggregation is a clustering problem with minimum size constraints on the resulting clusters or groups; the number of groups is unconstrained and the within-group homogeneity should be maximized. In the context of privacy in statistical databases, microaggregation is a well-known approach to obtaining anonymized versions of confidential microdata. Optimally solving microaggregation on multivariate data sets is known to be difficult (NP-hard). Therefore, heuristic methods are used in practice. This paper presents a new heuristic approach to multivariate microaggregation, which provides variable-sized groups (and thus higher within-group homogeneity) with a computational cost similar to the one of fixed-size microaggregation heuristics.

1 Introduction

Many among the usual transformations performed on data sets (data mining, knowledge discovery, statistical disclosure control for database privacy, etc.) can be viewed as clustering processes with different kinds of constraints [3, 6, 13, 18]. In this article we address the problem of microaggregation, a special kind of clustering problem where there are constraints on the minimum size of clusters or groups, but not on their number, and the within-groups homogeneity should be maximized. Microaggregation is a problem appearing in statistical disclosure control (SDC), where it is used to cluster a set of records in groups of at least k records, with k being a user-definable parameter. The collection of groups is called a k-partition of the data set. The microaggregated data set is built by replacing each original record by the centroid of the group it belongs to. The microaggregated data set can be released without jeopardizing the privacy of the individuals which form the original data set: records within a group are indistinguishable in the released data set. The higher the within-group homogeneity in the original data set, the lower the information loss incurred when replacing records in a group by their centroid; therefore within-group homegeneity is inversely related to information loss caused by microaggregation. The inverse of the within-groups sum of squares SSE is the most usual measure for clustering homogeneity [7, 9, 10, 14, 19]. SSE can be computed as

$$SSE = \sum_{i=1}^{s} \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)'(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i) \quad (1)$$

where s is the number of generated sets, n_i is the number of records in the i-th set, $\mathbf{x}_{ij}$ is the j-th record in the i-th set and $\bar{\mathbf{x}}_i$ is the centroid of the i-th set. In terms of SSE, the microaggregation problem consists of finding a k-partition with minimum SSE.

Microaggregation has been used for several years in different countries: it started at Eurostat [2] in the early nineties, and has since then been used in Germany [16] and several other countries [8]. These years of experience and practice devoted to microaggregation have bequeathed us a variety of approaches, which we next briefly summarize.

- Optimal methods:
  - Univariate case: In [11] a polynomial algorithm for optimal univariate microaggregation was presented. But optimal multivariate microaggregation was shown to be NP-hard in [15].
  - Multivariate case: Optimal multivariate microaggregation was shown to be NP-hard in [15]. So the only practical multivariate microaggregation methods are heuristic.
- Heuristic methods:
  - Fixed-size heuristics: The best-known example of this class is Maximum Distance to Average Vector (MDAV) [4, 6, 12]. MDAV produces groups of fixed cardinality k and, when the number of records is not divisible by k, one group with a cardinality between k and $2k-1$. MDAV has proven to be the best performer in terms of time and one of the best regarding the homogeneity of the resulting groups.
  - Variable-size heuristics: These yield k-partitions with group sizes varying between k and $2k-1$. Such a flexibility can be exploited to achieve higher within-group homogeneity. Variable-size methods include the genetic-inspired approach for small data sets in [18] and also [3, 13, 5].

To sum up, the challenge in microaggregation is to design good heuristics for the multivariate case, where "goodness" refers to combining high group homogeneity and computational efficiency. This paper is about a new method along this line. Since optimal microaggregation via exhaustive search is not a feasible benchmark for data sets of more than 15 records, we take as benchmark heuristic to compare with our new proposal the MDAV.

MDAV is, to the best of our knowledge, one of the best heuristic methods for multivariate microaggregation. However, being a fixed-size heuristic, there are situations in which it yields a k-partition far from the optimal one. This is illustrated by the next toy example.

Example 1. Let be D our data set composed by 13 records having 2 attributes.

$$D = \{(2.4, 3), (1.68, 4.9), (3.18, 5.54), (5.32, 3.6), (18.68, 11.49), (20.14, 9.56), (19.85, 12.33), (13.67, 18.9), (17.11, 21), (16.07, 19.23), (21.28, 18.9), (22, 21), (23, 18.5)\}$$

Considering $k = 3$, Figure 1 illustrates the output of the MDAV algorithm when applied over our toy-example. The figure depicts the records in D *(circles)* and the global centroid *(triangle)*. The group marked in red is very scattered and causes the overall 3-partition to be poor. This example shows that the fixed-size nature of MDAV may fail to suitably adapt the resulting k-partition to the particular data set.

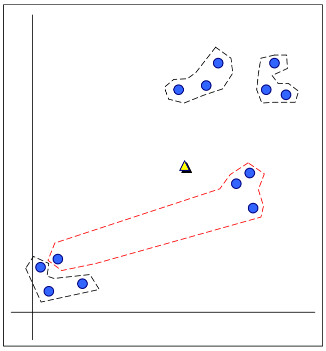

Fig. 1. Output of the MDAV algorithm on the toy example

1.1 Contribution and plan of the paper

In this paper we propose a new heuristic method for multivariate microaggregation called V-MDAV. V-MDAV stands for Variable-size Maximum Distance to Average Vector. It improves on the well-known MDAV method in terms of lower SSE while maintaining an equivalent computational cost.

The rest of the paper is organized as follows. In Section 2 we describe the proposed algorithm. Section 3 shows the experimental results obtained on several data sets. In Section 4 an assessment of the method's computational complexity is given. Finally, Section 5 is a conclusion.

2 Description of V-MDAV

As mentioned above, MDAV generates groups of fixed-size k and, thus, it lacks flexibility for adapting the group size to the distribution of the records in the data set, which may result in poor within-group homogeneity. V-MDAV is a new algorithm that intends to overcome this limitation by computing a variable size k-partition with a computational cost similar to the MDAV cost. Next, the main steps of V-MDAV are described in detail.

2.1 Group Generation

This is the first step of V-MDAV and it follows a similar strategy to MDAV, this is, firstly a matrix of distances containing the distances between any two records is built *(line 2 in the Algorithm)*. Next the global centroid is computed and the most

Algorithm 1 Variable group size algorithm

```
01) function V-MDAV in:DataSet $D$, Integer $k$; out:microaggregatedSet $M$ is
02)   Compute_Distances_Matrix($D$);
03)   $C$ = ComputeCentroidOfDataSet($D$);
04)   while (ThereAreMoreThan[$k-1$]RecordsToAssign) do
05)      $e$ = SelectTheMostDistantRecordToCentroid ($D$,$C$);
06)      $g_i$ = BuildGroupFromRecord($e$,$D$,$k$);
07)      $g_i$ = ExtendTheGroup($g_i$,$D$,$k$);
08)   end while
09)   $g_1 \dots g_s$ = AssignRemainingUnassignedRecords($D$, $g_1 \dots g_s$);
10)   $M$ = BuildMicroaggregatedDataSet($D$, $g_1 \dots g_s$);
11)   return $M$
12) end function
```

distant record from it is searched *(lines 3 and 5)*. Once this record is found, a group of k records is formed by selecting the $k-1$ records closest to the initial one *(line 6)*. At this point MDAV would apply this strategy until all records in the data set are assigned to a group; in V-MDAV there is an additional step *(line 7)* that allows it to adapt to the data set distribution and generate variable-size groups. As to complexity, the most important differences between MDAV and V-MDAV are:

- MDAV computes a centroid in each iteration. V-MDAV only computes the data set centroid at the beginning. This results in a computational time improvement.
- MDAV does not build a matrix of distances; on the contrary, it computes distances as many times as needed. Thus, V-MDAV is faster.

2.2 Extension of the group

Given a group g with p records, the record e_{min} among unassigned records outside g nearest to g and the minimum distance d_{in} between e_{min} and g are defined by Equation (2) and (3):

$$d_{in} = \min_{j \in [1, N_{un}]} [(d(e_i^g, e_j)], \forall i \in [1, p]] \tag{2}$$

$$e_{min} = \arg \min_{j \in [1, N_{un}]} [(d(e_i^g, e_j)], \forall i \in [1, p]] \tag{3}$$

where e_i^g denotes the i-th record in group g, e_j means the j-th record in the unassigned set of records and N_{un} is the number of unassigned records, this is, the number of records which have not yet been assigned to any group. If Equation (3) is satisfied by more than one record, one of them is randomly selected as e_{min}. Next, the minimum distance d_{out} from the selected record e_{min} to any of the remaining unassigned records is found using Equation (4):

$$d_{out} = \min_{j \in [1, N_{un}], e_{min} \neq e_j} [(d(e_{min}, e_j)] \tag{4}$$

Finally, in order to decide on the inclusion of e_{min} into group g, we compare its distance d_{in} to g with its distance d_{out} to the closest unassigned neighbor. Expression (5) gives the decision criterion:

$$ADD\_RECORD = \begin{cases} \text{YES if } \; d_{in} \; < \; \gamma \, d_{out} \\ \\ \text{NO} \quad \text{otherwise} \end{cases} \tag{5}$$

where γ is a *gain factor* that has to be tuned in order to improve the adaptability of V-MDAV. How to determine the best values of γ is not straightforward and due to space limitations it will not be discussed in this paper. We will expand a little more on this aspect in Section 3. The extension process is repeated until the group size equals $2k - 1$ or the condition in Expression (5) is not satisfied, because it was shown in [4] that, in an optimal k-partition, each group includes between k and $2k - 1$ records.

2.3 Addition of the last records

Similarly to MDAV, the proposed method can leave some records unassigned at the end of the main loop. Thus, it is necessary to assign these records to a group before ending the algorithm. The remaining records are assigned to their closest group *(line 9)*. At the end of all these steps, a microaggregated data set M is built from the resulting k-partition represented in the Algorithm 1 as $(g_1 \ldots g_s)$ *(line 10)*.

3 Experimental results

The experiments which have been carried out use very different data sets ranging from real to synthetic data up to 20 dimensions and from 400 records to almost 6000. Specifically, we have used the *Tarragona*, *Census* and *EIA* data sets [1], because they have become the usual reference data sets to test multivariate microaggregation [4, 5, 13]. Additionally, three new clustered data sets [1] have been synthetically created, as in [3], in order to show how MDAV and V-MDAV behave on this kind of data sets.

Our experiments demonstrate that V-MDAV is the best performer when applied to clustered data sets, because it can adapt to the structure of data. However, the distinction between *clustered* and *scattered* data does not only depend on the data but on the value of k. This variation in the outlook of the data set is similar to the effect of looking at a picture from the near or far distance. As an example let us consider the effect that a "zoom in" produces on an image. When we are very close to an image we are able to distinguish a greater number of details and some previously hidden clusters could become visible. On the other hand, a "zoom out" can produce a similar effect because, in some situations, in which we are very close to the image, the trees hide the wood. These "zoom in" and "zoom out" effects are equivalent to the ones produced by the variation of k.

For the above reasons, deciding whether a data set is scattered or clustered is not an easy task. In [17] a study on the type of data sets is presented and seems to indicate the existence of a clear dichotomy between the behaviors of clustered and scattered data sets. This research line is still open and we take some of its ideas to determine the data set type (Table 1, left-hand side). The determination of the best value of γ for a given data set is not straightforward. However, there is empirical evidence that values of γ close to zero are effective when the data are scattered because a low γ causes the algorithm to expand groups very conservatively, that is,

[1] We say that a data set is *clustered* when its records form natural clusters. Otherwise, we call the data set *scattered*: no natural clusters are apparent.

towards very close records only. In fact, for $\gamma = 0$ V-MDAV is equivalent to MDAV. On the contrary, when the data set is clustered the best values for γ are usually close to one. Then, V-MDAV takes more risks in expanding a group. In our experiments we have selected a $\gamma = 0.2$ for scattered data sets and a $\gamma = 1.1$ for clustered data sets. Table 1 (right-hand side) shows the information loss caused by each method measured using the following expression

$$I_{loss} = \frac{SSE}{SST} \cdot 100 \qquad (6)$$

where SST is the total sum of squares (sum of squared Euclidean distances from all records to the data set centroid). The lower I_{loss}, the better. From Table 1, it can be seen that the results obtained by MDAV and V-MDAV are very similar when working with scattered data sets: V-MDAV outperforms or matches MDAV, except for ”Census” and ”EIA” with $k = 3$ and $k = 4$, where it is *slightly* outperformed (by at most 0.08%). This is not a bad result, because MDAV is known to perform very very well on scattered data sets. On the other hand, V-MDAV clearly improves the result of MDAV when working on clustered data sets. More specifically, a spectacular improvement by almost 8% information loss reduction is achieved for the *Synthetic 2* data set with $k = 3$; the improvement is almost 7% for *Synthetic 1* with $k = 3$, etc. After analyzing these results, we can conclude that V-MDAV behaves as well as MDAV on scattered data and clearly outperforms it on clustered data.

Table 1. Results of the experiments. *Left:* Classification of the data sets into clustered or scattered according to k. *Right:* Information loss caused by MDAV and V-MDAV for different data sets and values of k.

Dataset	$k = 3$	$k = 4$	$k = 5$	$k = 10$
”Census”	S	S	S	S
”Tarragona”	S	S	S	S
”EIA”	S	S	C	C
”Synthetic 1”	C	S	S	S
”Synthetic 2”	C	S	S	S
”Synthetic 3”	S	C	C	S

Dataset	Method	$k = 3$	$k = 4$	$k = 5$	$k = 10$
”Census”	MDAV	5.66	7.51	9.01	14.07
	V-MDAV	5.69	7.52	8.98	14.07
”Tarragona”	MDAV	16.96	19.70	22.88	33.26
	V-MDAV	16.96	19.70	22.88	33.26
”EIA”	MDAV	0.49	0.67	1.78	3.54
	V-MDAV	0.53	0.75	1.30	2.82
”Synthetic 1”	MDAV	7.64	8.84	12.20	30.21
	V-MDAV	0.94	5.64	11.83	30.03
”Synthetic 2”	MDAV	9.54	13.35	17.34	39.28
	V-MDAV	1.63	8.21	16.49	38.91
”Synthetic 3”	MDAV	5.95	7.66	9.03	22.36
	V-MDAV	2.17	1.85	5.21	21.41

4 Computational analysis

Let n be the number of records in the data set. The dimension of the data set will not be considered in the analysis because it is usually much smaller than n and it only has a slight computational impact on distance computations. The breakdown of complexity is as follows:

1. Computing the distances matrix can be done with $O(n^2)$ (proportional to the number of distances to be computed).
2. Computing the centroid has a cost of $O(n)$.
3. Computing the distance from each point to the centroid requires n distance computations. Thus, it has $O(n)$ cost.
4. *Main loop.* In each iteration, between k and $2k-1$ records are microaggregated. Since on average $(3k-1)/2$ records are grouped, about $2n/(3k-1)$ iterations are needed. Each iteration consists of:
   a) Selecting the most distant record from the centroid, which costs $O(n)$.
   b) Building a group of size k. Since $k-1$ records are to be added and $n/2$ records, on average, remain unassigned, $((k-1)n)/2$ comparisons are required. Thus, the cost is $O(n)$.
   c) Extending the group. Up to $k-1$ records can be added to the current group, say $(k-1)/2$ on average. This requires performing $((k-1)n)/4$ comparisons, with a cost $O(n)$.

   Therefore the computational cost of the main loop is

$$O(\frac{2n}{3k-1}(n+n+n)) = O(\frac{6n^2}{3k-1}) = O(n^2)$$

5. Finally, as some records may remain unassigned, a final step is performed. For each unassigned record, n distances are computed to determine the group to which it will be added. This results in a $O(n)$ cost.

Thus the overall complexity of V-MDAV is $O(n^2)$.

5 Conclusion and further work

V-MDAV, a new method for multivariate microaggregation has been presented. V-MDAV overcomes the fixed group size constraint of previous heuristics with a similar computational cost. Therefore, V-MDAV offers increased flexibility without computational over cost. Moreover, the way in which V-MDAV expands the groups can be tuned by using the γ *gain factor*. Experimental results show substantial performance improvement by V-MDAV when working on clustered data sets. A number of research issues remain open and will be addressed in future work: *a)* Analyze the behavior of the γ *gain factor* in detail and determine the optimal value of γ for a given data set. *b)* Study the dichotomy between clustered and scattered data sets.

Acknowledgement. The authors wish to thank Prof. Josep Domingo-Ferrer and Dr. Josep M. Mateo-Sanz for their insightful comments and kind suggestions. Partial support by the Government of Catalonia under grant 2005 SGR 00446 and by the Spanish Ministry of Education and Science under project SEG2004-04352-C04-01 ”PROPRIETAS” is also gratefully acknowledged.

References

1. R. Brand, J. Domingo-Ferrer, and J. M. Mateo-Sanz. Reference data sets to test and compare sdc methods for protection of numerical microdata, 2002. European Project IST-2000-25069 CASC, http://neon.vb.cbs.nl/casc.

2. D. Defays and P. Nanopoulos. Panels of enterprises and confidentiality: the small aggregates method. In *Proc. of 92 Symposium on Design and Analysis of Longitudinal Surveys*, pages 195–204, Ottawa, 1993. Statistics Canada.
3. J. Domingo-Ferrer, A. Martínez-Ballesté, and J. M. Mateo-Sanz. Efficient multivariate data-oriented microaggregation. *Manuscript*, 2005.
4. J. Domingo-Ferrer and J. M. Mateo-Sanz. Practical data-oriented microaggregation for statistical disclosure control. *IEEE Transactions on Knowledge and Data Engineering*, 14(1):189–201, 2002.
5. J. Domingo-Ferrer, F. Sebé, and A. Solanas. A polynomial-time approximation to optimal multivariate microaggregation. *Manuscript*, 2005.
6. J. Domingo-Ferrer and V. Torra. Ordinal, continuous and heterogenerous k-anonymity through microaggregation. *Data Mining and Knowledge Discovery*, 11(2), 2005.
7. A. W. F. Edwards and L. L. Cavalli-Sforza. A method for cluster analysis. *Biometrics*, 21:362–375, 1965.
8. Economic Commission for Europe. Statistical data confidentiality in the transition countries: 2000/2001 winter survey. In *Joint ECE/Eurostat Work Session on Statistical Data Confidentiality*, 2001. Invited paper n.43.
9. A. D. Gordon and J. T. Henderson. An algorithm for euclidean sum of squares classification. *Biometrics*, 33:355–362, 1977.
10. P. Hansen, B. Jaumard, and N. Mladenovic. Minimum sum of squares clustering in a low dimensional space. *Journal of Classification*, 15:37–55, 1998.
11. S. L. Hansen and S. Mukherjee. A polynomial algorithm for optimal univariate microaggregation. *IEEE Transactions on Knowledge and Data Engineering*, 15(4):1043–1044, July-August 2003.
12. A. Hundepool, A. Van de Wetering, 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*. Statistics Netherlands, Voorburg NL, may 2005. http://neon.vb.cbs.nl/casc.
13. M. Laszlo and S. Mukherjee. Minimum spanning tree partitioning algorithm for microaggregation. *IEEE Transactions on Knowledge and Data Engineering*, 17(7):902–911, 2005.
14. J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297, 1967.
15. A. Oganian and J. Domingo-Ferrer. On the complexity of optimal microaggregation for statistical disclosure control. *Statistical Journal of the United Nations Economic Comission for Europe*, 18(4):345–354, 2001.
16. M. Rosemann. Erste ergebnisse von vergleichenden untersuchungen mit anonymisierten und nicht anonymisierten einzeldaten amb beispiel der kostenstrukturerhebung und der umsatzsteuerstatistik. In *G. Ronning and R. Gnoss (editors) Anonymisierung wirtschaftsstatistischer Einzeldaten, Wiesbaden: Statistisches Bundesamt*, pages 154–183, 2003.
17. Agusti Solanas. On the type of data dets: Clustered vs. scattered. *Manuscript*, 2006.
18. Agusti Solanas, Antoni Martínez-Ballesté, Josep M. Mateo-Sanz, and Josep Domingo-Ferrer. Multivariate microaggregation based on a genetic algorithm. *Manuscript*, 2006.
19. J. H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58:236–244, 1963.