Towards fuzzy c-means based microaggregation

Josep Domingo-Ferrer[1] and Vicenç Torra[2]

[1] Dept. Comput. Eng. and Maths - ETSE, Universitat Rovira i Virgili
Av Paisos Catalans 26, 43007 Tarragona (Catalonia, Spain)
e-mail: jdomingo@etse.urv.es
[2] Institut d'Investigació en Intel·ligència Artificial - CSIC
Campus UAB s/n, 08193 Bellaterra (Catalonia, Spain)
e-mail: vtorra@iiia.csic.es

Abstract. National Statistical Offices collect data from respondents and then publishes them. To avoid disclosure, data is protected before the release. One of the existing masking methods is microaggregation. This method is based on obtaining a set of clusters (clustering stage) and then aggregating the values of the elements in the cluster (aggregation stage). In this work we propose the use of fuzzy c-means in the clustering stage.

1 Introduction

National Statistical Offices (NSO) collect data from respondents and then disseminate them. Due to legal restrictions, data has to be protected so that no disclosure of sensitive data is possible. This is, it should not be possible to link sensitive data from a particular respondent to this respondent.

To avoid disclosure, data is masked before their release. Statistical Disclosure Control studies masking methods for applying some distortion to data in such a way that the data is still analytically valid. This is, the published data is valid for researchers and users because they can reach to similar conclusions than the ones inferred from the original data.

At present, there exist a large set of micro-data protecting methods. For example, among the most widely used [4], we find the following ones: sampling, top and bottom coding, recoding, data swapping, microaggregation. See [2] and [8] for an extensive review of micro-data methods and [1] for a detailed analysis comparing information loss and disclosure risk for existing methods.

This work is devoted to one of the methods for numerical micro-data protection: microaggregation.

1.1 Microaggregation

Given a datafile to be protected, microaggregation consists on obtaining microclusters of similar records (at least k records have to be included in each

cluster) and then publishing the averages of each cluster instead of publishing the original data. According to this, the method is defined by the following two steps (see [3] for details):

Step 1: Partition of the records into a set of disjoint groups so that the number of records in each group is at least k.

Step 2: For each record, the value of each variable is replaced by the aggregation of the corresponding values of all records in the same group.

Thus, this method consists on applying a clustering method to the original data and then returning for each record the prototype of the cluster instead of returning the record itself. This method can be formalized as follows:

1. Let $X = \{x_1, x_2, \cdots, x_n\}$ be a set of given data for n respondents for which p variables are observed. Thus, x_i is a p-dimensional vector $x_i = (x_{i1}, \cdots, x_{ip})$. Let $d(x, y)$ be a p-dimensional distance. Then, if we denote a partition of X into g groups by $G = \{G_1, \cdots, G_g\}$, with $|G_i|$ being the cardinality of groups G_i and $\hat{x}_i$ being the average data vector of G_i, the optimal partition G for microaggregation is the one that minimizes:

$$\sum_{i=1}^{g} \sum_{x_j \in G_i} d(x_j, \hat{x}_i) \tag{1}$$

   while satisfying $|G_i| \geq k$ for all $G_i \in G$.
2. Given an optimal solution G, x_j is replaced by $\hat{x}_i$ if x_j belongs to group G_i. elements in G_i).

Finding the optimal solution in Equation 1 is not an easy task. The following difficulties are found:

1. The size of all clusters has to be at least k and as similar to k as possible. Clusters of less records are not allowed. This constraint, not usually considered in clustering, required the development of specific clustering methods.
2. The problem of finding the best k-size partition of the domain is a NP problem. To have approximated solutions, heuristic approaches have been developed.
3. Classical microaggregation techniques assign the same values to all records in a cluster. This give clues to attackers, specially in the case of applying multivariate microaggregation to several groups of variables for the same records.

In this work, we introduce the use of fuzzy c-means for microaggregation. We define an heuristic method based on fuzzy c-means to obtain partitions of at least size k. Then, for each record, the mean value of all variables is not necessarily taken from the same cluster. Instead, using fuzzy membership

values, they can be taken from different clusters. An additional advantage of fuzzy c-means is the existence of efficient computational algorithms (as the one described [6]).

The structure of this work is as follows. In Section 2, we review fuzzy c-means. Then, in Section 3 we detail our approach. The paper finishes with the conclusions.

2 Fuzzy c-means

While classical clustering methods partition the records of a given domain into a disjoint set of clusters, fuzzy clustering methods build a set of clusters in which elements can belong at the same time to several of them. When an element belongs to more than one cluster their membership is partial. This is modeled considering membership degrees in the $[0, 1]$ interval in such a way that 0 means no membership and 1 means full membership. In this case, it is commonly assumed that for a given element in the set, the summation of the membership of this element to all clusters is 1. Formally, this is defined as follows:

Let $X = \{x_1, \cdots, x_n\}$ be the set of elements (records or respondents in our application), then a set of membership functions $A = \{A_1, \cdots, A_c\}$ is a fuzzy partition of X into c clusters if and only if:

$$\sum_{i=1}^{c} A_i(x_k) = 1 \qquad \text{for all } x_k \in X$$

Here, $A_i(x_k)$ is interpreted as the membership of the k-th element to the i-th set.

Note that restricting $A_i(x_k)$ in $\{0, 1\}$ this definition corresponds to a crisp partition.

Among existing fuzzy clustering methods, a well known one is Fuzzy c-means. This method is a generalization of the k-means clustering method. This latter method consists on finding the set of c clusters such that the following objective function is minimized:

$$J(A, V) = \sum_{k=1}^{n} \sum_{i=1}^{c} A_i(x_k) \cdot ||x_k - v_i||^2$$

subject to the following constraints: $A_i(x_k) \in \{0, 1\}$ and $\sum_{i=1}^{c} A_i(x_k) = 1$ for all $x_k \in X$. Here, $|| \cdot ||$ corresponds to the Euclidean distance, and $V = \{v_i\}_{i=1,\cdots,c}$ are the centers of the clusters.

Defining,

$$M = \{(A_i(x_k)) | A_i(x_k) \in \{0, 1\}, \sum_{i=1}^{c} A_i(x_k) = 1 \text{ for all } k\}$$

the problem to be solved is to find A and V such that $min_{A \in M} J(A, V)$.

For computing the fuzzy c-means, the constraints M are replaced by:

$$M_f = \{(A_i(x_k)) | A_i(x_k) \in [0,1], \sum_{i=1}^{c} A_i(x_k) = 1 \text{ for all } k\}$$

However, solving the same problem with these new *fuzzy* constraints lead to the same crisp solution because the objective function is linear with respect to $A_i(x_k)$, and the optimization problem is solved using linear programing. Therefore, the optimal solution for $A_i(x_k)$ is found at an extremal point in the $[0, 1]$ interval. Thus, it is a crisp value.

To obtain a fuzzy solution, the following objective function was proposed (see e.g. [7] and its references for details):

$$J(A, V) = \sum_{k=1}^{n} \sum_{i=1}^{c} (A_i(x_k))^m \cdot ||x_k - v_i||^2$$

Here, m is a real number $m \geq 1$ that influences the membership values. With $m = 1$, the solution is crisp and, then, the larger is m, the more fuzzy the clusters we obtain.

To find A and V that minimize this objective function constrained in M_f, the following algorithm is used (see [7] or [5]):

Step 1: Generate an initial A and V

Step 2: Solve $min_{A \in M_f} J(A, V)$ computing (A_{ik} stands for $A_i(x_k)$):

$$A_{ik} = \Big(\sum_{j=1}^{c} \Big(\frac{||x_k - v_i||^2}{||x_k - v_j||^2} \Big)^{\frac{1}{m-1}} \Big)^{-1}$$

Step 3: Solve $min_V J(A, V)$ computing:

$$v_i = \frac{\sum_{k=1} n (A_{i,k})^m x_k}{\sum_{k=1}^{n} (A_{i,k})^m}$$

Step 4: If the solution does not converge, go to step 2; otherwise, stop

3 At least k Fuzzy c-means

As the fuzzy c-means do not assure that clusters have at least k elements, we have defined an algorithm for clustering based on the fuzzy c-means. Being k the minimum number of records allowed in a cluster, the algorithm is defined in the Algorithm 1. Note that in this algorithm, $fcm(X, c, m)$ corresponds to standard fuzzy c-means (as described in Section 2).

Once the clusters have been obtained, the values for the variables are (randomly) replaced by the values of the center of the class. When an element belongs to several clusters, the corresponding center is selected by a random distribution proportional to the membership degree.

Algorithm 1 micro-fuzzy-c-means

Algorithm *At least* k fuzzy c-means (X, k, m) **is**
begin
  $c := n/(3 \cdot k)$;
  $P = fcm(X, c, m)$;
  $cardMin = \min_{p \in P} |p|$
  if $cardMin < k$ **then**
    decrease c, increase m and start again
  end if
  for $p \in P$ **do**
    nothing
    begin
      if $|p| > 2 \cdot k$ **then**
        define m' in terms of m and $|p|$;
        micro-fuzzy-c-means (p, k, m');
      end if
    end
  end for
end

4 Conclusions

In this work we have shown the applicability of fuzzy c-means to microaggregation. Future work is on the comparison of our approach to existing techniques following the approach in [1].

Acknowledgements

Josep Domingo-Ferrer and Vicenç Torra are partially supported by the EU project CASC: Contract: IST-2000-25069 and CICYT project STREAMOBILE (TIC2001-0633-C03-01/02)

References

1. Domingo-Ferrer, J., Torra, V., (2001), A Quantitative Comparison of Disclosure Control Methods for Microdata, 111-133, in Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies, P. Doyle, J. I. Lane, J. J. M. Theeuwes, L. M. Zayatz (Eds.), Elsevier.
2. Domingo-Ferrer, J., Torra, V., (2001), Disclosure Control Methods and Information Loss for Microdata, 91-110, in Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies, P. Doyle, J. I. Lane, J. J. M. Theeuwes, L. M. Zayatz (Eds.), Elsevier.
3. Domingo-Ferrer, J., Torra, V., (2002), Aggregation techniques for statistical confidentiality, in "Aggregation operators: New trends and applications", (Ed.), R. Mesiar, T. Calvo, G. Mayor, Physica-Verlag, Springer.

4. Felso, F., Theeuwes, J., Wagner, G. G., (2001), Disclosure Limitation Methods in Use: Results of a Survey, 17-42, in Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies, P. Doyle, J. I. Lane, J. J. M. Theeuwes, L. M. Zayatz (Eds.), Elsevier.
5. Klir, G., Yuan, B., (1995), Fuzzy Sets and Fuzzy Logic: Theory and Applications, Prentice-Hall, U.K.
6. Kolen, J. F., Hutcheson, T., (2002), Reducing the time complecity of the Fuzzy C-Means Algorithm, IEEE Trans. on Fuzzy Systems, April, 263-267.
7. Miyamoto, S., Umayahara, K., (2000), Methods in Hard and Fuzzy Clustering, pp 85–129 in Z.-Q. Liu, S. Miyamoto (Eds.), Soft Computing and Human-Centered Machines, Springer-Tokyo.
8. Willenborg, L., De Waal, T., (1996), Statistical Disclosure Control in Practice, Springer LNS 111.