# Cluster-specific information loss measures in data privacy: a review

*(Invited Paper)*

Vicenç Torra* and Susana Ladra**
* IIIA-CSIC, Campus UAB, E-08193 Bellaterra, Catalonia, Spain
E-mail vtorra@iiia.csic.es
** Database Laboratory, University of A Coruña,
Campus de Elviña, 15071 A Coruña (Galicia, Spain)

## Abstract

*Data protection mechanisms need to find a trade-off between information loss and disclosure risk. To this end, information loss and disclosure risk measures have been developed.*

*Due to the fact that when data is published it is usual to ignore which kind of analyses a user will pursue with the data, generic information loss measures are used to analyse the impact of the perturbation method onto the data. Such generic information loss measures are defined in terms of a few general-enough statistics.*

*Nevertheless, a more fine-grained analysis is needed for particular data uses.*

*In this paper we provide the reader with a review of a few results on cluster-specific information loss measures. More specifically, we consider the case of using fuzzy clustering to the perturbated data.*

## 1. Introduction

Data protection mechanisms [4], [24] modify an original data set with the goal of ensuring the privacy of the respondents. That is, they do changes to the data to ensure that sensitive information cannot be inferred from the modified data. Nevertheless, in practical applications a complete protection cannot be achieved unless the changes to the data are so large that the modified data is useless for real analyses.

Due to this, in real applications, a trade-off is sought between the protection and the data utility. That is, good protection mechanisms are the ones that ensure a reasonable level of privacy while permitting, at the same time, the user to perform almost correct analyses and inferences from the data.

To measure the goodness of the methods, two families of measures have been studied [5]. They are the disclosure risk and the information loss measures.

Disclosure risk measures [5] are to determine to what extent a protected file ensures privacy. In short, the measure is *proportional* to the *amount* of relevant information that can be inferred from the protected data. One of the approaches to define risk is using record linkage algorithms. Formally, a subset of the original data (the one that is assumed an intruder might have) is linked against the protected data (the one that will be published). In this case, when all the records of the original data are linked with the corresponding records in the protected file, we have zero protection. In contrast, when no record can be linked, we have total protection. In general, the proportion of correct links corresponds to a measure of the risk. It has to be said that other approaches exist for measuring the risk, as computing the uniqueness of the records.

Information loss measures are to determine to what extent the perturbated data is useful for doing the same analyses and inferences that a user would like to carry out with the original data.

When defining information loss measures, an important aspect to be taken into account is the intended use of the data by the user. Nevertheless, such use is usually not known. In fact, the use might be rather diverse for a single file, as sometimes data is published in the web and different users will apply different techniques and analyses. Due to this, some *generic* information loss measures have been developed. They are measures for a non specific use. They are defined as the divergence of a few statistics between the original data file and the protected data file. The probabilistic information loss measure [13] is an example of such measures. See [3], [5], [25] for details on such measures.

Nevertheless, although such generic measures play an important role, real analyses and inferences are based on the application of particular methods, and the behavior of such methods might diverge in a relevant

0-7695-3102-4/08 $25.00 © 2008 IEEE
DOI 10.1109/ARES.2008.87

way from what is stated by such generic measures. Due to this, it is important to develop analyses of the influence of the protection methods on particular tools for data analysis.

In this paper we study this problem, focusing on the case of clustering, and, more particularly, on fuzzy clustering. That is, we consider the analysis of how protection mechanisms influence the results of clustering. This problem can be formulated in terms of *specific* information loss measures and, more specifically, on *cluster-specific* information loss measures. That is, measures that evaluate the extent to which the results of clustering are influenced by the changes introduced in the data.

The structure of the paper is as follows. In Section 2, we review fuzzy clustering. Then, in Section 3, we consider the definition of cluster-specific information loss. Finally, the paper finishes with some conclusions and lines for future research.

## 2. Preliminaries

In this section we review a few topics that are needed later on. In particular, we review fuzzy partitions and fuzzy clustering algorithms.

### 2.1. Fuzzy sets and fuzzy partitions

We review here the concept of fuzzy partitions as fuzzy clustering algorithms give, as a result, a fuzzy partition.

*Definition 1:* Let $X$ be a reference set. Then $\mu : X \rightarrow [0,1]$ is a membership function.

*Definition 2:* Let $X$ be a reference set. Then, a set of membership functions $\mathcal{M} = \{\mu_1, \ldots, \mu_m\}$ is a fuzzy partition of $X$ if for all $x \in X$ it holds

$$\sum_{i=1}^{m} \mu_i(x) = 1$$

### 2.2. Fuzzy clustering

Typically, clustering methods are to partition a set of data into disjoint sets. In the case of fuzzy clustering, a fuzzy partition is built instead of a crisp one. In this paper we will mainly focus on Fuzzy $c$-means, although other algorithms for fuzzy clustering will also be considered. See e.g. [7], [14], [16] for details on fuzzy clustering. Fuzzy $c$-means, that was first proposed in [1], is described in most books on fuzzy sets and fuzzy clustering. See, e.g., the above mentioned references.

**Algorithm 1** Fuzzy $c$-means

Step 1: Generate initial $\mu$ and $V$

Step 2: Solve $min_{\mu \in M} J(\mu, V)$ computing:

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

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

$$v_i = \frac{\sum_{k=1}^{n} (\mu_{ik})^m x_k}{\sum_{k=1}^{n} (\mu_{ik})^m}$$

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

We describe below Fuzzy $c$-means. In the description we will use the following notation. We have a set of objects $X = \{x_1, \ldots, x_n\}$ and we want to build $c$ clusters from this data. Then, the method builds a fuzzy partition of $X$. The fuzzy partition (the clusters) are represented by membership functions $\mu_{ik}$, where $\mu_{ik}$ is the membership of the $k$th object ($x_k$) to the $i$th cluster.

Fuzzy $c$-means needs an additional value $m$ that should satisfy $m \geq 1$. When $m$ is near to 1, solutions tend to be crisp (with the particular case that $m = 1$ corresponds to the crisp $c$-means, or $k$-means). In contrast, when $m$ is *large*, solutions tend to be clusters with large fuzziness in their boundaries.

Formally, fuzzy $c$-means constructs the fuzzy partition $\mu$ from $X$ solving the minimization problem stated below. In the formulation of the problem, $v_i$ is used to represent the cluster center, or centroid, of the $i$-th cluster.

$$J_{FCM}(\mu, V) = \{ \sum_{i=1}^{c} \sum_{k=1}^{n} (\mu_{ik})^m ||x_k - v_i||^2 \} \quad (1)$$

subject to the constraints $\mu_{ik} \in [0,1]$ and $\sum_{i=1}^{c} \mu_{ik} = 1$ for all $k$.

A (local) optimal solution of this problem is obtained using an iterative process that interleaves two steps. One that estimates the optimal membership functions of elements to clusters (when centroids are fixed) and another that estimates the centroids for each cluster (when membership functions are fixed). This iterative process is described in Algorithm 1.

Noise clustering (NC), possibilistic $c$-means (PCM) and fuzzy possibilistic $c$-means are some of the variations of fuzzy $c$-means. We have used them on our analyses. Noise clustering was introduced in [2] to reduce the effects of noisy data. To do so, the method introduces a special noise cluster. Possibilistic $c$-means

also includes some noise clusters but in this case there is a noise cluster for each regular cluster. This method was introduced in [9]. Fuzzy possibilistic $c$-means is a variation of PCM, introduced in [17], to avoid coincident clusters and to make the final clusters less sensitive to initializations.

## 3. Information loss measures for clustering

Given an original file $X$ and the corresponding protected file $X'$, information loss measures are based on the comparison of the results of a few statistics (or data analyses) on both $X$ and $X'$. For example, we can compare the mean of $X$ and $X'$ for the different variables in the files. Then, the larger is the difference, the larger the information loss.

Similar approaches can be applied to any other data analysis tool. This is also the case for clustering. Let us consider a given clustering algorithm $clust$ with parameters $par$, and let denote its application to the data file $X$ by $clust_{par}(X)$. Then, we can define the information loss of $clust_{par}$ applied to the data file $X$ and its protected data file $X'$ as the divergence or distance between $clust_{par}(X)$ and $clust_{par}(X')$. That is,

$$IL(X, X') = distance(clust_{par}(X), clust_{par}(X')).$$

Naturally, the larger the divergence, the larger the loss.

In the case of partitive crisp clustering (that is, a method that returns a partition of the objects), there are a few tools for comparing the clusters (see e.g. [8], [18]). To name a few, there exist the Rand [19] and the Adjusted Rand index, the Jaccard index, and the Mántaras distance [12]. We can define the loss as proportional to the distance, or inversely proportional to the above mentioned indices.

### 3.1. Comparison of fuzzy clusters

Nevertheless, there is no such variety of methods for comparing fuzzy clusters. In the rest of this section we describe a few approaches we have introduced for tackling this problem.

A first approach [10] was to consider the transformation of the fuzzy partition into crisp sets applying $\alpha$-cuts. Recall that the $\alpha$-cut of a fuzzy set for a given $\alpha \in [0, 1]$ is a standard subset (the set of elements with a membership function larger than $\alpha$). However, this approach presents a problem as an $\alpha$-cut of a fuzzy partition does need to be a partition. Therefore, we cannot apply directly the indices and distances for fuzzy partitions. Therefore, we need to apply an adhoc approach. In our experiments we used three $\alpha$-cuts (with $\alpha_1 = 0.9$, $\alpha_2 = 0.5$ and $\alpha_1 = 0.1$) and then we used the distance between the resulting crisp clusters.

| | FCM | NC | PCM | PFCM | PIL |
|---|---|---|---|---|---|
| 0.1 | 0.0037 | 0.0030 | 0.0036 | 0.0029 | 4.1310 |
| 0.2 | 0.0084 | 0.0049 | 0.0072 | 0.0063 | 6.4298 |
| 0.4 | 0.0153 | 0.0092 | 0.0136 | 0.0192 | 9.2348 |
| 0.6 | 0.0209 | 0.0141 | 0.0197 | 0.0188 | 12.6145 |
| 0.8 | 0.0310 | 0.0165 | 0.0261 | 0.0270 | 16.6538 |
| 1.0 | 0.0229 | 0.0322 | 0.0318 | 0.0245 | 18.5534 |
| 1.2 | 0.0943 | 0.0796 | 0.0393 | 0.0840 | 24.5021 |
| 1.4 | 0.0314 | 0.0257 | 0.0414 | 0.0560 | 28.6009 |
| 1.6 | 0.0356 | 0.0448 | 0.0491 | 0.0603 | 33.7005 |
| 1.8 | 0.0969 | 0.0735 | 0.0585 | 0.0934 | 35.6461 |
| 2.0 | 0.1622 | 0.0367 | 0.0737 | 0.0654 | 37.5090 |
| | 0.7679 | 0.7403 | 0.9780 | 0.8923 | 1.0000 |

Table 1. (a) Columns 2-5 give the $\alpha$-cut based distance computed for several files (protected with noise addition with different values of noise, first column) when the clustering algorithm selected is one of the fuzzy clustering methods; (b) the last column includes the averaged probabilistic information loss measure (aPIL); (c) last row corresponds to the correlation of the cluster-specific measures with respect to the aPIL.

Table 1 presents the results of such distances computed on the results of different fuzzy clustering algorithms for a data file with 1080 records and 13 variables. Each row corresponds to a different level of protection (protection using noise addition with a parameter $p = 0.1, 0.2, \ldots, 2.0$ – the first column indicates the degree of protection). The data file (named `census`), which is public and is described in detail in [26], has been used by several researchers in several experiments [5], [25]. The Table also includes the correlation of such measures with respect to the average PIL [13], a generic information loss.

Later on, in order to overcome the difficulties of the previous approach and, at the same time, avoiding the transformation from a fuzzy partition to a set of crisps sets, we proposed two different distances for fuzzy partitions. This result, presented in [21], permitted to analyse two fuzzy clustering methods: fuzzy $c$-means and fuzzy $c$-means with tolerance [6].

The two distances proposed were based, respectively, on the cluster centers and the membership functions. We define them below.

- **Distance based on cluster centers.** The distance is solely based on the cluster representatives of each cluster. That is, their centroids. First, a mapping between the clusters is obtained so that the clusters of each clustering result are *aligned*

(the *nearest* cluster center is assigned in the alignment). Then, the Euclidean distance between a center and its associated one is computed. The overall distance is the summation of the distances between the pairs of clusters. We will denote this distance by $d_1$.

- **Distance based on membership functions.** The distance is based on the membership functions. The computation uses the mapping established before, and then computes for each record, the distance between its membership values to the clusters obtained for the original file and the membership values to the clusters obtained for the protected file. We will denote this distance by $d_2$.

The range of the two distances are rather different. The maximum values we have obtained for $d_1$ and $d_2$ after all our experiments using different clustering algorithms, parameterizations and noisy data, we have got a maximum of 130 for $d_1$ and 5500 for $d_2$.

The application of these distances to real data presents an additional problem. Clustering algorithms ensure convergence to a local optima, but not to a global one. Due to this, different executions of the method might result into different clusters.

Local convergence of clustering algorithms is not a big problem in some applications of unsupervised machine learning. The different fuzzy partitions obtained in different executions can represent different knowledge, and might correspond to different points of view. Nevertheless, in our case, when we are interested in measuring the information loss, this is a big problem.

Note that due to the local optima, different executions of the same algorithm with the same data might result into different clusters. Therefore, we might have that $clust_{par}(X) \neq clust_{par}(X)$ for different executions of $clust$ with parameter $par$ on the same data set $X$. Moreover, we might have that the difference between $clust_{par}(X)$ and $clust_{par}(X')$ is very large not because $X$ and $X'$ are different but because we are just in rather different local optima.

In [21], for each data file $X$, each cluster algorithm $clust$ and each parameterization $par$, we have considered several executions of $clust_{par}(X)$ computing for each of them its objective function. Then, we have selected the fuzzy partition with the lowest membership function. Such fuzzy partition is the one used latter for comparison.

Up to 20 executions have been done in [21] for each $< X, clust, p >$. Nevertheless, we still got several local optima as we got a few results with $clust_{par}(X) \neq clust_{par}(X')$ when $X' = X + noise$ with $noise = 0$.

| | $d_1$ | $d_2$ | $O.F.$ |
|---|---|---|---|
| 0.0 | 3.21 | 40.73 | 2826.0 |
| 0.1 | 3.21 | 40.67 | 2827.0 |
| 0.2 | 3.17 | 40.86 | 2829.0 |
| 0.4 | 0.32 | 0.92 | 2859.0 |
| 0.6 | 3.28 | 42.09 | 2844.0 |
| 0.8 | 3.48 | 43.48 | 2862.0 |
| 1.0 | 3.55 | 48.87 | 2886.0 |
| 1.2 | 2.24 | 55.56 | 2908.0 |
| 1.4 | 1.44 | 18.35 | 2935.0 |
| 1.6 | 2.27 | 36.83 | 2978.0 |
| 1.8 | 2.71 | 45.59 | 3006.0 |
| 2.0 | 4.24 | 96.87 | 3028.0 |
| | 0.0125 | 0.4073 | |

Table 2. Distances $d_1$ and $d_2$ between the clusters originated from the original and the protected file for different values of noise and using the fuzzy $c$-means (FCM) as the clustering algorithm. Executions with the number of clusters set to 10 (i.e., $c = 10$). The values achieved for the objective function are also included for each protected file (last columne). The optimal value found for the original file was 2851. The last row corresponds to the correlation with aPIL.

Table 2 shows the results of the distance between the original file and the protected one when data is clustered using fuzzy $c$-means on the whole file (all 13 variables) and the number of clusters is 10 (i.e., $c = 10$). The two distances defined above $d_1$ and $d_2$ are used. Nevertheless, the results show that the distance is not monotonic with respect to the noise added. This is due to the different local minima found.

In Table 3 we present similar results, but in this case only 2 of the variables are considered in the clustering. As in this case we get a better convergence, we have monotonicity of the distance with respect to the noise. Two cases are presented, one with the number of clusters equal to 10 (i.e., $c = 10$) and the other with the number of clusters equal to 20 (i.e., $c = 20$).

## 4. Conclusions and future work

In this paper we have studied cluster-specific information loss measures. We have reviewed a few approaches for computing the differences between clusters and we have shown the difficulties such methods pose. In particular, we have explained that fuzzy clustering algorithms converge into an optimum that might be a local optimum. This causes some inconveniences when comparing the results of the same clustering algorithm on both the original and the protected file.

| | $c = 10$ | | | $c = 20$ | | |
|---|---|---|---|---|---|---|
| | $d_1$ | $d_2$ | $O.F.$ | $d_1$ | $d_2$ | $O.F.$ |
| 0.0 | 5E-9 | 1E-15 | 225.26 | 2.86 | 208.90 | 107.19 |
| 0.1 | 0.10 | 0.92 | 225.67 | 3.03 | 157.10 | 107.20 |
| 0.2 | 0.08 | 1.74 | 225.02 | 0.69 | 13.46 | 107.21 |
| 0.4 | 0.21 | 8.45 | 224.63 | 1.80 | 113.00 | 106.97 |
| 0.6 | 0.49 | 25.27 | 225.45 | 2.15 | 73.73 | 106.67 |
| 0.8 | 3.16 | 217.38 | 224.85 | 3.22 | 214.29 | 108.47 |
| 1.0 | 1.29 | 73.13 | 226.53 | 2.80 | 224.25 | 108.66 |
| 1.2 | 3.80 | 252.37 | 225.21 | 3.96 | 259.46 | 109.11 |
| 1.4 | 0.66 | 80.99 | 227.00 | 4.45 | 318.17 | 109.61 |
| 1.6 | 3.13 | 257.35 | 228.43 | 2.92 | 337.55 | 112.14 |
| 1.8 | 3.20 | 315.55 | 230.97 | 5.11 | 454.07 | 111.77 |
| 2.0 | 3.25 | 313.78 | 231.82 | 5.31 | 510.52 | 110.00 |
| | 0.78 | 0.87 | | 0.75 | 0.85 | |

Table 3. Distances $d_1$ and $d_2$ between the clusters originated from the original and the protected file for different values of noise and using the fuzzy $c$-means (FCM) as the clustering algorithm; and values for the objective function ($O.F.$). Results correspond to the best result after 20 executions. Clustering was based on the first 2 variables of the file and 10 clusters (left) and 20 clusters (right). The optimal value found for the original file was 225.26 for the case of 10 clusters (left) and 107.06 for the case of 20 clusters (right). The last row corresponds to the correlation with aPIL.

To solve the difficulties presented here, we have considered the use of intuitionistic fuzzy sets for expressing the results of the execution of fuzzy clustering. Intuitionistic fuzzy sets are used when there is some uncertainty on the membership function.

Formally, this uncertainty is represented with a pair of functions $\mu$ and $\nu$. $\mu$ corresponds to the membership function and $\nu$ permits to express the uncertainty.

Then, using intuitionistic fuzzy sets, we might consider the definition of an intuitionistic fuzzy partition, that permits us to summarize the 20 fuzzy partitions obtained from the 20 executions of the fuzzy clustering algorithms. Initial steps on the definition of intuitionistic fuzzy partitions have been presented in [22].

As a future work, we need to check whether the approach presented in [22] is suitable for our purposes here. That is, if it is suitable to define cluster-specific information loss measures. At present, only formal and theoretical results on the suitability of the approach have been obtained. We have proven [22] the convergence of our definition to a fuzzy partition when the number of executions is large. Convergence results can be proven for a few fuzzy clustering methods. E.g., fuzzy $c$-means and fuzzy $c$-means with entropy [15]. Finally, as future work, we have to evaluate the cluster-specific information loss for some families of protection methods.

## Acknowledgments

Partial support by the Spanish MEC (projects ARES – CONSOLIDER INGENIO 2010 CSD2007-00004 – and eAEGIS – TSI2007-65406-C03-02) is acknowledged.

## References

[1] Bezdek, J. C. (1981) Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York, 1981.

[2] Davé, R. N. (1991) Characterization and detection of noise in clustering, Pattern Recognition Letters, 12 657-664.

[3] Domingo-Ferrer, J., Mateo-Sanz, J. M., Torra, V. (2001) Comparing SDC methods for microdata on the basis of information loss and disclosure risk, Pre-proceedings of ETK-NTTS'2001, (Eurostat, ISBN 92-894-1176-5), Vol. 2, 807-826, Creta, Greece.

[4] Domingo-Ferrer, J., Torra, V. (2001) Disclosure Control Methods and Information Loss for Microdata, Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies, Elsevier Science, Pages 91-110, 2001.

[5] Domingo-Ferrer, J., Torra, V. (2001) A Quantitative Comparison of Disclosure Control Methods for Microdata, Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies, Elsevier Science, Pages 111-133, 2001.

[6] Hasegawa, Y., Endo, Y., Hamasuna, Y., Miyamoto, S. (2007) Fuzzy $c$-means for data with tolerance defined as hyper-rectangle, Proc. MDAI 2007, Lecture Notes in Artificial Intelligence 4617 237-248.

[7] Höppner, F., Klawonn, F., Kruse, R., Runkler, T., (1999), Fuzzy cluster analysis, Wiley.

[8] Hubert, J., Arabie, P. (1985) Comparing partitions, Journal of Classification 2:1 193-218.

[9] Krishnapuram, R., Keller, J. M. (1993) A possibilistic approach to clustering, IEEE Trans. on Fuzzy Systems, 1 98-110.

[10] Ladra, S., Torra, S., On the comparison of generic information loss measures and cluster-specific ones, submitted.

[11] Ladra, S., Torra, V., Fuzzy clustering in data privacy: measuring information loss for synthetic data, submitted.

[12] Lopez de Mantaras, R. (1991) A Distance-Based Attribute Selection. Measure for Decision Tree Induction, Machine Learning, 6, 81-92.

[13] Mateo-Sanz, J. M., Domingo-Ferrer, J. Sebé, F. (2005) Probabilistic information loss measures in confidentiality protection of continuous microdata, Data Mining and Knowledge Discovery, 11:2 181-193.

[14] Miyamoto, S., (1999), Introduction to fuzzy clustering, (in Japanese), Ed. Morikita, Tokyo.

[15] Miyamoto, S., Mukaidono, M. (1997) Fuzzy $c$ - means as a regularization and maximum entropy approach, Proc. of the 7th IFSA Conference, Vol.II 86-92.

[16] 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.

[17] Pal, N. R., Pal, K., Bezdek, J. C. (1997) A Mixed c-Means Clustering Model, Proc. of the 6th IEEE Int. Conf. on Fuzzy Systems, Barcelona, Spain, 11-21.

[18] Raghavan, V. V., Ip, M. Y. L. (1982) Techniques for measuring the stability of clustering: a comparative study, Proc. of the 5th annual ACM conference on Research and development in information retrieval, 209-237.

[19] Rand, W. M. (1971) Objective criteria for the evaluation of clustering methods, J. of the American Statistical Association, 66 846-850.

[20] Sneath, P. H. A., Sokal, R. R (1973) Numerical Taxonomy. Freeman, San Francisco.

[21] Torra, V., Endo, Y., Miyamoto, S., On the comparison of some fuzzy clustering methods for privacy preserving data mining: towards the development of specific information loss measures, submitted.

[22] Torra, V., Miyamoto, S., Intuitionistic Fuzzy Partitions for the Comparison of Fuzzy Clusters, submitted.

[23] Trottini, M. (2003) Decision models for data disclosure limitation, PhD Dissertation, Carnegie Mellon University, http://www.niss.org/dgii/TR/Thesis-Trottini-final.pdf.

[24] Willenborg, L., de Waal, T. (2001) *Elements of Statistical Disclosure Control*, Lecture Notes in Statistics, Springer-Verlag.

[25] Yancey, W. E., Winkler, W. E., Creecy, R. H. (2002) Disclosure risk assessment in perturbative microdata protection, Inference Control in Statistical Databases 2002, Lecture Notes in Computer Science 2316 135-152.

[26] CASC: Computational Aspects of Statistical Confidentiality, EU Project, http://neon.vb.cbs.nl/casc/ (Test Sets)