Fuzzy measures and integrals in re-identification problems

Jordi Nin and Vicenç Torra

Institut d'Investigació en Intel·ligència Artificial - CSIC
Campus UAB s/n
08193 Bellaterra (Catalonia, Spain)
Email: vtorra@iiia.csic.es, jnin@iiia.csic.es

Abstract—In this paper we give an overview of our approach of using aggregation operators, and more specifically, fuzzy integrals for solving re-identification problems. We show that the use of Choquet integrals are suitable for some kind of problems.

Keywords: Record linkage, Fuzzy integrals, OWA operators, data mining, data cleaning, privacy preserving data mining.

1. Introduction

Re-identification algorithms when applied to databases permit to identify those objects that can be found in different files but that correspond to the same entity. Two particular family of algorithms can be distinguished.

- **Record linkage (or record matching) algorithms:** These algorithms intend to link records of one file with those records in another file that correspond to the same individual. The difficulties of the approach are due to the fact that the records might be described using different attributes, or, in the case of using the same attributes, there are errors in the data.
- **Schema matching or, more particularly, attribute matching:** They are algorithms to link attributes or schemata in databases. The typical problems these algorithms have to face is that the name of the attributes in different files do not coincide, or that a single attribute in one file corresponds to several ones in another file. In the most general case, schema matching needs to construct $n : m$ relationships.

In recent years we have shown that some particular fuzzy integrals can be used in some cases for re-identification. They have been applied to both record linkage and attribute matching.

In this paper we will describe how this problem is tackled using fuzzy integrals, the underlying assumptions of our model, and describe why this approach works. Some simple examples of application will be given.

The structure of the paper is as follows. In Section 2 we give some preliminaries. We focus on fuzzy measures and integrals and on the re-identification methods. Then, in Example 3 we present an example. The paper finishes with some conclusions.

2. Preliminaries

This section is devoted to give a short review of fuzzy measures and integrals and then, to some re-identification methods.

2.1. Fuzzy measures and integrals

Definition 1 *A fuzzy measure μ on a set X is a function $\mu : 2^X \longrightarrow \mathbb{R}^+$ with the following properties:*

1. $\mu(\emptyset) = 0$
2. $m(A) \leq m(B)$ *whenever* $A \subset B$ *and* $A, B \in 2^X$.

For the sake of simplicity, we assume $X = \{1, \ldots, N\}$.

Families of fuzzy measures have been defined in the literature. In this paper we focus on the so-called symmetric fuzzy measures. In a symmetric fuzzy measure, the measure of a set only depends on the cardinality of the set but not on the elements of the set.

Definition 2 *A fuzzy measure μ is symmetric when $\mu(A) = \mu(B)$ whenever $|A| = |B|$.*

Here, $|\cdot|$ represents the cardinality of a set.

Symmetric fuzzy measures can be represented by non-decreasing functions $f : [0,1] \rightarrow [0,1]$ such that $f(0) = 0$ and $f(1) = 1$ so $\mu(A) = f(|A|/|X|)$.

Definition 3 *[1] Given a fuzzy measure μ and a function f, the Choquet integral of f with respect to μ is defined using the i-th order statistics ($x_{(i)}$) as:*

$$CI_\mu(f) := \sum_{i=1}^{n} f((x_{(i)}) - f(x_{(i-1)}))\mu(\{(i) \cdots (n)\})$$

where we define $x^{(0)} := 0$.

The Choquet integral with respect to a symmetric fuzzy measure corresponds to the OWA operator [9].

Definition 4 *[4] Given a fuzzy measure μ and a function f (into $[0,1]$), the Sugeno integral of f with respect to μ is defined using the i-th order statistics ($x_{(i)}$) as:*

$$S_\mu(f) := \bigvee_{j=1}^{n} f(x_{(j)}) \wedge \mu(A_{(j)})$$

$A_{(i)} = \{x_{(i)}, \cdots, x_{(n)}\}, A_{(n+1)} = \emptyset$. *Here $\wedge$ denotes the minimum and $\vee$ denotes the maximum.*

2.2. Re-identification methods

As said in the introduction, standard record linkage methods are centered on the linkage of objects belonging to the same entities from two or more files when such files share a set of variables, or any other kind of information. In this case, the difficulties for a good performance of record linkage algorithms are due to the fact that files contain errors (*e.g.*, the income of an individual is not the same in both files or attributes are represented in different scales).

Two main approaches have been used for standard record linkage. See [6, 7, 8] for more details:

Probabilistic Record Linkage(PRLB): For each pair of records (a,b), we compute a conditional probability of having a correct link using a coincidence vector of variables. Then, we use this probability to classify each pair (a,b) as either a linked pair (LP) or a non-linked pair (NP).

Distance-based Record Linkage(DBRL): Records of file A are compared to records of file B with respect to a given distance function, and then each record in A is linked to the nearest record in B using such distance function.

3. Example: normal distributions

We describe below some experiments to test the performance of fuzzy integrals in a re-identification problem. The fuzzy integrals have been combined with fuzzy measures.

The experiments use normal distributions as entities, so a normal distribution is considered as an individual or attribute (either record linkage or attribute linkage).

The fuzzy integrals have been computed with respect to symmetric fuzzy measures generated from the following three parametric functions:

$$Q_1^\alpha(x) = x^\alpha \text{ for } \alpha = 1/5, 2/5, \cdots, \ldots, 10/5$$

$$Q_2^\alpha(x) = 1/(1 + e^{(\alpha - x)*10}) \text{ for } \alpha = \{0, 0.1, \ldots 0.9\}$$

$$Q_3^\alpha(x) = \begin{cases} 0 & \text{if } x \leq \alpha \\ 1 & \text{if } x > \alpha \end{cases} \quad \text{for } \alpha = \{0, 0.1, \ldots 0.9\}$$

3.1. Data generation

We have generated two different synthetic data files using a pseudo-random gaussian generator. The files were generated with twenty normal distribution (entities). The first file contain normal distributions with (μ, σ) from the sets $\mu = \{0, 1, 2, 3, 4\}$ and $\sigma = \{0.5, 1, 1.5, 2\}$. The second file contains normal distributions from (μ, σ) with μ in $\mu = \{0, 2, 4, 6, 8\}$ and σ in the same set $\sigma = \{0.5, 1, 1.5, 2\}$. For each distribution we have generated twenty-five elements (variables).

These files correspond to the original data files. These distributions are represented in Figure 1.

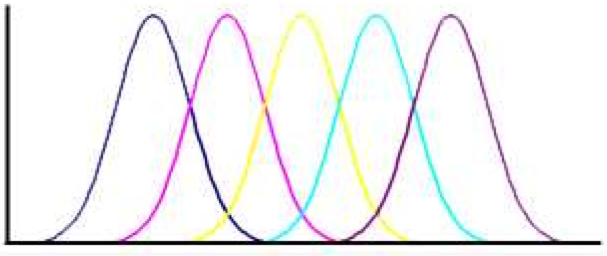

Figure 1: Normal distributions with different μ

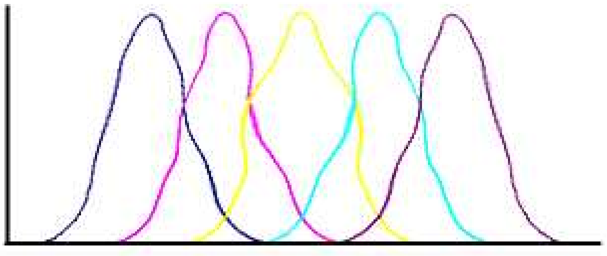

Figure 2: Normal distributions with adding linear noise

Afterwards, we have generated four synthetic data files for each original data file. These files have been generated with the same normal distributions but adding linear noise based on a normal distribution $N(0, 1)$. We have considered four levels of noise, $\epsilon = 0.1 * N(0, 1)$, $\epsilon = 0.25 * N(0, 1)$, $\epsilon = 0.5 * N(0, 1)$ and $\epsilon = N(0, 1)$. These files are to be used in the re-identification experiment. These distributions are displayed in Figure 2.

According to this description, we obtain two original files with twenty records (normal distributions) described with twenty-five attributes. Additionally, we get eight distorted files with an identical number of records belonging to the same normal distributions but modified with some noise. The number of attributes is the same, although in this case, they are *noisy* attributes.

3.2. Re-identification

We have used the fuzzy integrals for re-identification. In particular, we have used the Choquet integral with symmetric fuzzy measures. As said above, such operator is equivalent to the OWA operator.

The integral, with a particular measure, is applied to each record, obtaining a representative for such record. Different parameterizations lead to different representatives.

Then, once we have representatives for all records, re-identification is done comparing the representatives of each record in one file with the representatives of each record in the other file. The rationale is that when the representatives are similar, the original records should also be similar.

In our case, we have considered the fuzzy measures listed above with 10 different parameterizations. This leads to 10 representatives for each record (for each normal distribution, either original or distorted).

Taking this into account, new files are created with the representatives. That is, for each pair of files, we obtain a second pair of files. These new files have the same number of records as the original files (representatives are built for each original record) and 10 attributes (there is one attribute for each parameterization). As the representatives of both files have been obtained from the same parameterizations, we can say that the new files are described using the same attributes.

Therefore, as both file share attributes (the parameterizations), we can use now standard record linkage methods (*e.g.*, probability or distance-based) to link the files.

3.3. Results

We have obtained good results with all fuzzy measures. The results are given in Table 1. The table contains the best number of re-identifications obtained for each experiment with either distance-based or probabilistic record linkage.

The experiments show that the re-identification is possible, and that the larger the distortion, the worse the re-identification. These are expected results.

ϵ	μ	Case 1	Case 2
N(0,1)*0.1	$Q_1^\alpha(x)$	14	16
N(0,1)*0.25	$Q_1^\alpha(x)$	11	14
N(0,1)*0.5	$Q_1^\alpha(x)$	10	14
N(0,1)*1.0	$Q_1^\alpha(x)$	8	11
N(0,1)*0.1	$Q_2^\alpha(x)$	14	18
N(0,1)*0.25	$Q_2^\alpha(x)$	14	17
N(0,1)*0.5	$Q_2^\alpha(x)$	10	17
N(0,1)*1.0	$Q_2^\alpha(x)$	11	11
N(0,1)*0.1	$Q_3^\alpha(x)$	15	17
N(0,1)*0.25	$Q_3^\alpha(x)$	11	18
N(0,1)*0.5	$Q_3^\alpha(x)$	12	16
N(0,1)*1.0	$Q_3^\alpha(x)$	11	11

Table 1: Results of the re-identification. Case 1: Files with the following normal distributions $N(0, 0.5)..N(4, 2)$; Case 2: Files with the following normal distributions $N(0, 0.5)..N(8, 2)$

3.4. Alternative example

In the previous section we have described the results when the masked file are obtained with linear noise addition. But we can considered other ways to obtain a masked file.

We can consider the addition of other kind of noise like $N(\mu, \sigma) * \epsilon + (1 - \epsilon) * N(\mu', \sigma')$. In this case we obtain a new probability distribution like the one we can see in Figure 3.

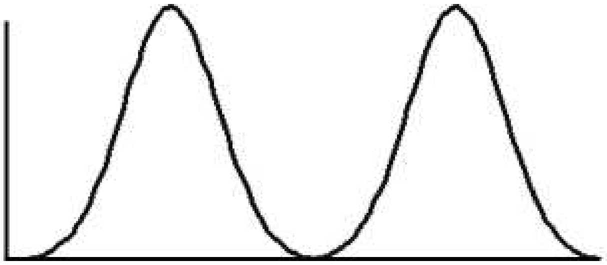

Figure 3: normals

If we apply the Choquet integral with a symmetric measure (OWA operator) to this family of distribution for small ϵ we will obtain similar results than when the distortion corresponds to the addition of linear noise. This is so because the OWA operator sorts data values from the smallest values to greatest ones and the resulting values after the ordering are not so dissimilar. For larger values of ϵ this would not be the case, the noisy distribution might be re-identified with a time-series generated from $N(\mu', \sigma')$.

4. Conclusions

In this paper, we have studied the use of owa operators for the re-identification individual problem, and focused in the particular case in which individuals are represented by normal distributions. We have proved that owa operators are a suitable tools for such re-identification as they have lead to good results with three different fuzzy measures.

Additional experiments have been done with real data. Some of the results are reported in [5] (attribute matching) and [2, 3] (record linkage).

Acknowledgments

This work was partly funded by the Spanish Ministry of Education and Science under project SEG2004-04352-C04-02 ”PROPRIETAS”.

References

[1] Choquet, G., (1953/54), Theory of capacities, Ann. Inst. Fourier, 5 131-295.

[2] Nin, J., Torra, V., (2005), Towards the use of OWA operators for record linkage, Proc. of the European Soc. on Fuzzy Logic and Technologies, CD-ROM.

[3] Nin, J., Torra, V., (2005), Empirical analysis of database privacy using twofold integrals, Lecture Notes in Artificial Intelligence 3801 1-8.

[4] Sugeno, M., (1974), Theory of fuzzy integrals and its applications, Ph. D. Dissertation, Tokyo Institute of Technology, Tokyo, Japan.

[5] Torra, V., (2004), OWA operators in data modeling and re-identification, IEEE Trans. on Fuzzy Systems, 12:5 652-660.

[6] Torra, V., Domingo-Ferrer, J., (2003), Record linkage methods for multidatabase data mining, in V. Torra (Ed.), Information Fusion in Data Mining, Springer, 101-132.

[7] Winkler, W. E., (2003), Data Cleaning Methods, Proc. SIGKDD 2003, Washington.

[8] Winkler, W. E., (2004), Re-identification methods for masked microdata, Privacy in Statistical Databases 2004, Lecture Notes in Computer Science 3050 216-230.

[9] Yager, R. R., (1988), On ordered weighted averaging aggregation operators in multi-criteria decision making, IEEE Trans. on Systems, Man and Cybernetics, 18 183-190.

[10] Yager, R. R., (1993), Families of OWA operators, Fuzzy Sets and Systems, 59 125-148.