Journal of Theoretical and Applied Information Technology
30th November 2022. Vol.100. No 22

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

MODELLING THE RELATIONSHIP BETWEEN PRIVACY AND LOSS OF DATA UTILITY WHEN ANALYSING THE DATA

CHOKRI NOUAR[1], AHMED DRISSI[2]

[1]Mohammed V University. Department of Mathematics, Rabat, Morocco

[2]AbdelMalek Essaadi University. National School for Applied Sciences, Tangier, Morocco

E-mail: [1]chokri.nouar@gmail.com, [2]idrissi2006@yahoo.fr

ABSTRACT

Big data has become a primary resource for decision-makers. Data analysis makes it possible to extract new data which can be very useful. Meanwhile, the data owner has to protect the privacy of people and establishments. Therefore, private information can neither be all shared nor extractable from any analysis. That is why the owner of the data must negotiate with the explorer according to the objectives of each of the two. In many cases objectives can contradict each other. The challenge is to make a compromise between protecting privacy and retaining the usefulness of the data processed. This article strives to propose a model of anonymization based on the theory of fuzzy logic. In the process of anonymization, we assign, in the first step, a degree of identification and sensitivity to each data, then the qualitative data is encoded before we finally anonymize depending on the type of data. Our model has been proved to be useful and reliable for both the owner and the explorer of the data. Therefore, it helps protect the data and ensures its privacy.

Keywords: *Data analysis security, Privacy, Fuzzy logic, Yager, Anonymization.*

1. INTRODUCTION

The basic idea of the privacy-preserving data mining is to modify data in order to efficiently execute data mining algorithms without compromising the security of sensitive information contained in the data [1],[2]. The basic idea of the privacy-preserving data mining is to modify data in order to efficiently execute data mining algorithms without compromising the security of sensitive information contained in the data [1],[2].

Data privacy-preserving has been considered a priority in data mining. In this respect, researchers have used numerous methodologies to address the privacy-preserving problem. A method based on fuzzy optimization was proposed by Hemanta in [3]. Anonymization techniques such as subtree and full-tree generalization are also employed to solve data privacy-preserving as stated in [4].

In addition, a methodology based on the role of the user is developed by [5]. In this methodology, we can identify four types of users involved in data mining applications:

1. The user who has certain data desired for an exploration task (The data provider),
2. The user who collects data from suppliers and then publishes it to the data extractor (The data collector),
3. The user who performs data mining tasks (The data miner)
4. The user who makes decisions based on the results of data mining in order to achieve certain objectives (The decision maker).

A user represents either a person or an organization and can play several roles at the same time. Hence, the types of users can be summarised into two categories: the owner who seeks to protect the data and the explorer who seeks to obtain more data useful for analysis and subsequently extracts credible results.

In the next section, we discuss the relationship between sensitive information, that identify their owners, and the possibilities to hide them while retaining the likelihood of exploring this data with a certain degree of credibility. The third section is devoted to detailing our modelling approach, we recall the theory of fuzzy logic, precisely the fuzzy conjunction and the fuzzy disjunction, and we measure the importance of information. In the fourth section, the findings of the study and a discussion of

the results will be presented before we finally draw our conclusions.

2. PROTECTION OF PRIVACY

The proposal of protection strategies for big data security and privacy becomes a priority for scientists [6]. To be published, data protection requires a categorization according to the degree of identification and its sensitivity, and then we mask them by a process of anonymization with a certain degree.

2.1 Data Categorization

Direct disclosure of data to the explorer leads to a loss of confidentiality. Raw sensitive data must not be used directly for extraction, but they should rather be hidden because they identify their owners. The first step to protect information is to classify them according to the degree of identification of their owner. These information can be classified into four categories:

- Identifier (ID): Attributes that can directly and uniquely identify a person such as names, telephone number, …
- Sensitive attributes (SA): Attributes that an individual wants to hide such as illness, salary, political affiliation, …
- Quasi-identifier (QID): Attributes that can be linked to external data to re-identify individual records, such as gender, age and postal code.
- Other attributes: Attributes other than ID, QID and SA

Before being published to other people, the data table is anonymized, that is, the identifiers will be removed and the quasi-identifiers will be modified. Therefore, the identity of the individual and the values of sensitive attributes can be masked from the adversaries.

2.2 The Anonymization Operations

The anonymization of the data table depends mainly on the degree of confidentiality that we want to preserve in the anonymized data. An anonymization algorithm based on chaos and perturbation is proposed in [7] of which five types of anonymization operations are cited:

- The Generalization: This operation replaces certain values with a parent value in the taxonomy of an attribute.
- The deletion: This operation replaces some values with a special value (for example "*" artesi)
- The Anatomization: This operation does not modify the quasi-identifier or the sensitive attribute, but disassociates the relationship between the two.
- The permutation: This operation disassociates the relationship between the QID and the SA by partitioning a set of data records into groups and mixing their sensitive values within each group.
- The disturbance: this operation replaces the original data values with some synthetic data values, so that the statistical information calculated from disturbed data does not differ significantly from the statistical information calculated from the original data.

3. THE PROPOSED MODEL

In this section, we will describe our model which is based on fuzzy logic to protect the sensitive data and its exploited results. The privacy-preserving data mining covers two types of protections: the protection of sensitive data itself and the protection of sensitive exploited results. Each user seeks to protect their personal interests in terms of preservation of privacy or usefulness of the data. The interests of different users are correlated so that the interactions between different users can be modelled. First, we recall some concepts of fuzzy logic, including operations on fuzzy sets and Triangular norm and Triangular co-norm. We use them for modelling the importance of information.

3.1 Overview of Fuzzy Logic

Definition 1 [8]

Let X be a reference set, a fuzzy subset A of X is characterized by a membership function $\mu_A: X \rightarrow [0,1]$ which associates to each element x of X a real $\mu_A(x)$ of the interval$[0,1]$. $\mu_A(x)$ denotes the degree of belonging from x to A (the degree of truth of the proposition " x belongs to A ", where the

degree of compatibility of x with the state described by A).

Definition 2 [9]

A triangular norm (t-norm) is a function $T: [0,1] \times [0,1] \rightarrow [0,1]$ checking commutativity, associativity, monotony and the neutral element $(T(x,1) = x, \forall x \in [0,1])$. Any t-norm can be used to define the intersection of fuzzy subsets $\left(\forall x \in X\ \mu_{A \cap B}(x) = T(\mu_A(x), \mu_B(x))\right)$.

Example: $T = \min(\mu_A(x), \mu_B(x))$

Definition 3 [9]

A triangular co-norm (t-conorm) is a function $C: [0,1] \times [0,1] \rightarrow [0,1]$ checking commutativity, associativity, monotony and the neutral element $(C(x,0) = x, \forall x \in [0,1])$. Any t-conorm can be used to define the union of fuzzy subsets $\left(\forall x \in X\ \ \mu_{A \cup B}(x) = C(\mu_A(x), \mu_B(x))\right)$.

Example: $C = \max(\mu_A(x), \mu_B(x))$.

3.2 Assessment of The Importance of Information

The owner of the data is interested in the degree of identification (α) and the degree of sensitivity (β) of each information. For this, the owner seeks to modify the data in such a way as to protect private data and only discloses the part which does not threaten his privacy. The data explorer is interested in the usefulness of each information in the analysis. It determines the threshold that the modification of the information must not exceed to obtain the objectives of the analysis.

In this regard, we can take the owner's phone number as an example. We can assign α=1 because the phone number identifies its owner. For sensitivity, we can attribute β=0 if the number is professional and β=0.03 if it is personal (he does not want to broadcast his number to 97% of the public).

S represents the set of the sensitive information, I denotes the set of the identification information and x and y imply two information. $\beta = \mu_S(x)$ is the degree of truth of "x is a sensitive information", $\alpha = \mu_I(x)$ is the degree of truth of "x is an identification information". The owner of the data assigns a parameter $\beta \in [0,1]$ where $\beta = \mu_S(x)$ to each information x to estimate its degree of sensitivity, and assigns a parameter $\alpha \in [0,1]$ where $\alpha = \mu_I(x)$ to each information x to estimate its degree of identification. Moreover, $\mu_{S \cap I}(x)$ is the degree of truth of "x is a sensitive information and x is an identification information" and $\mu_{S \cup I}(x)$ is the degree of truth of "x is sensitive information or x is an identification information".

The degree of truth of $\{x\} \subset S$ is $\mu_S(x)$ then the degree of truth of $\{x, y\} \subset S$ is $C(\mu_S(x), \mu_S(y))$. Similarly, the degree of truth of $\{x\} \subset I$ is $\mu_I(x)$ then the degree of truth of $\{x, y\} \subset I$ is $C(\mu_I(x), \mu_I(y))$.

The combination of a number of information $x_1, x_2, \ldots x_n$ increases the degree of identification $\mu_I(x_1, x_2, \ldots x_n) \geq \mu_I(x_i), \forall i \in \{1 \ldots n\}$. The explorer can extract additional information which can have a degree of identification $\mu_I(x_{suppl}) \geq \mu_I(x_1, x_2, \ldots x_n) \geq \mu_I(x_i), \forall i \in \{1 \ldots n\}$. The combination of a number of information $x_1, x_2, \ldots x_n$ increases the degree of sensitivity $\mu_S(x_1, x_2, \ldots x_n) \geq \mu_S(x_i), \forall i \in \{1 \ldots n\}$. The explorer can extract additional information which can be sensitive with a degree of sensitivity $\mu_S(x_{suppl}) \geq \mu_S(x_1, x_2, \ldots x_n) \geq \mu_S(x_i), \forall i \in \{1 \ldots n\}$.

Let two pieces of information $x_1(\alpha_1, \beta_1)$ and $x_2(\alpha_2, \beta_2)$ with α_i is the degree of identification and β_i is the degree of sensitivity of x_i, $i = 1,2$. The owner of x_1 and x_2 is identified by α_1 or α_2 from which we can assign a degree of identification by $C(\alpha_1, \alpha_2)$ with C being the fuzzy disjunction.

Let us define the sensitivity of the two information, that is to say the owner wants only a proportion β_1 of the public to know this information x_1, and a proportion β_2 to know the information x_2. The proportion of the public that should know both information at the same time is $T(\beta_1, \beta_2)$ with T being the fuzzy conjunction. For our study, we choose the so-called t-norm and t-conorm of Yager.

$$T(x,y) = \max\left(1 - ((1-x)^p + (1-y)^p)^{\frac{1}{p}}, 0\right)$$

and

$$C(x,y) = \min\left((x^p + y^p)^{\frac{1}{p}}, 1\right).$$

Our choice is due to the fact that $x + y = 1$ does not generally imply that $C(x,y) = 1$, which is relevant to the degrees of identification. $\mu_S(\{x,y\}) = C(\mu_S(x), \mu_S(y))$ is the degree of sensitivity of $\{x,y\}$, and $\mu_I(\{x,y\}) = C(\mu_I(x), \mu_I(y))$ is the degree of identification of $\{x,y\}$.

We assign m as the degree of importance of the information x, we calculate $C(\mu_S(x), \mu_I(x))$ and $T(\mu_S(x), \mu_I(x))$, and the maximum among which is

m. Similarly, we calculate $C(\mu_{SUI}(x), \mu_{SUI}(y))$ and $T(\mu_{SUI}(x), \mu_{SUI}(y))$ and the maximum among which is the degree of importance of the information $\{x, y\}$.

The degree of importance of information measures a combination of sensitivity and identification of such information. If the degree of importance of information m is big enough, that is to say there is more chance that the owner of the data will be identified, and, therefore, their privacy will be threatened. The explorer of the data wishes to have it as large as possible. using the value of the degree of importance of the information m, as an essential parameter, we construct the modification function f_m.

3.3 The Construction of The Modification Function

The data variables can generally be quantitative (continuous or discrete) or qualitative (ordinal or nominal). In the continuous case, quantitative variables are represented by real intervals (or real numbers), while the qualitative ones are coded through integers. The modification can either completely or partially hide information or disassociate the links between the different information. Hiding the information is of various levels: (Suppression, generalization, disturbance, ...). The levels of disassociations of the links between the information are: (Anonymization, Permutation, …).

The modification function f_m includes the functions of generalization, restriction, deletion or disassociations (anonymization, permutation). It is always possible to hide the variables if $m = 1$ (in this case, we replace all the variables with "*" artesi) or to keep them as they are if $m = 0$ (no modifications). When, $0 < m < 1$ we make the modification depending on the type of variables.

3.3.1 The continuous variables

When variables are continuous, we assume that the information is represented in the form of real intervals$[a, b]$ (especially points $\{a\} = [a, a]$. The modification function f_m is defined between two intervals $f_m([a, b]) = [a', b']$.

If $[a, b] \subset [a', b']$, we say that f_m is a generalization. If $[a', b'] \subset [a, b]$, we say that f_m is a disturbance (restriction). If $[a', b'] = \{$"a symbol"$\}$, we say that f_m is a deletion.

The modification will be a generalization or restriction by a proportion m, if the variables are real intervals.

$$f_m([a,b]) = \begin{cases} \left[a - m\frac{b-a}{2}, b + m\frac{b-a}{2}\right] & \text{generalization} \\ \left[a + m\frac{b-a}{2}, b - m\frac{b-a}{2}\right] & \text{restriction} \\ ******** & \text{deletion} \end{cases}$$

if the information is a real number

$$f_m(a) = [a - ma, a + ma]$$

3.3.2 The discrete or qualitative variables

When variables are discrete or qualitative, we can neither generalize nor reduce, but we can restrict or permute. For this, we construct the modification function as a perturbation or a permutation of information. Perturbation is used if the variables are quantitative discrete or qualitative ordinal, and permutation is utilized when the variables are qualitative nominal.

- **perturbation function**

$$f_m(x_i) = p + \left(x_i + (-1)^i E(x_i \times m)\right) \bmod(q)$$

with $i \in [0; N]$, q is the max and p is the min of discrete variables x_i.

- **Permutation function**

$$\sigma_m(i) = (i + E(N \times m))\ mod(N)$$

with $i \in [0;\ N]$ and E(x) is the integer part of x, N is the number of variables (number of the lines), x_i is the i^{th} variable in the quantitative case and the value of the variable in the qualitative ordinal case, and i is the position of the variable in the qualitative nominal case.

The permutation function, completely, disassociates the relationship between the modified variables and the other data variables, then there are, on the one hand, indicators which should not be extracted because they will be erroneous. On the other hand, there remain other indicators which keeps the same importance for the explorer. We will treat a concrete example in the simulation paragraph.

We note that the intervals $\left[a - m\frac{b-a}{2}, b + m\frac{b-a}{2}\right]$, $\left[a + m\frac{b-a}{2}, b - m\frac{b-a}{2}\right]$, $[a - ma, a + ma]$ and $[a, b]$ have the same centre, which implies that

the arithmetic mean, variance and covariance with variables of the same type remain unchanged. The other indicators change depending on m.

4. APPLICATION AND RESULTS

The data owner assigns degrees of identification x and sensitivity y to each variable of the data. In the first step, we calculate the degree of importance using a new method which is based on fuzzy logic, namely the Yager t-norm and t-conorm (with degree p=2).

$$m = \mathrm{Max}(\mathrm{T}(x, y), \mathrm{C}(x, y))$$

In the second step, we convert the qualitative nominal variable into corresponding numbers. In the next step, we will organize the data as follows: the first line contains the names of the variables, the second contains the types of the variables (string, float, list) and the 3th contains the degrees of importance. After organizing the data, we apply our model of modification according to degree of importance of each variable.

Sample

In table 1, we present the initial data, for which we want to study the statistical indicators of central tendencies and dispersion without revealing some sensitive information.

Firstly, we determine the degrees of sensitivity and identifications of each attribute as shown in table 2.

We convert the qualitative nominal variable as follows: Athletic: Beginner =1, Advanced =2, Professional =3. and the academic levels: Tertiary =1, High school=2, Primary =3, Pre-school=4.

Table 3 shows the data after converting the qualitative data and calculating the degree of importance. The data will, then, be ready to be modified by our model.

Table 1: the initial data

name	Monthly salary	Athletic	Academic levels	Number of children	City	City temperature
Ahmed	1500	Beginner	Tertiary	2	Rabat	[2, 33]
Bob	500	Professional	Elementary	3	Fes	[-5, 45]
Ali	1700	Beginner	Tertiary	1	Tangier	[-3, 34]
Alice	600	Advanced	Primary	5	Ifran	[-7, 32]
Guerov	1000	Advanced	High school	2	Casa	[3, 39]
Jilali	1000	Professional	High school	3	Midelt	[-10, 46]
Eve	500	Professional	Pre-school	5	Agadir	[5, 39]
Driss	700	Beginner	High school	4	Marrakesh	[2, 49]
Omar	700	Professional	Primary	4	Dakhla	[4, 41]
Mick	800	Advanced	Pre-school	3	Ouazzane	[-2, 40]

Table 2: the degree of identification and sensitivity

	name	Monthly salary	Athletic	Academic levels	Number of children	City	City temperature
Degree of identification	1	0.2	0.2	0.3	0.3	0.5	0.2
Degree of sensitivity	0.1	0.6	0.9	0.2	0.1	0.3	0.7

Table 3: The initial coded data

	name	Monthly salary	Athletic	Academic levels	Number of children	City	City temperature
Type of data	Str	List	Float	Float	Float	Str	List
m = Degree of importance	1.000	0.632	0.922	0.361	0.316	0.583	0.728
	Ahmed	1500	1	1	2	Rabat	[2, 33]
	Bob	500	3	3	3	Fes	[-5, 45]
	Ali	1700	1	1	1	Tangier	[-3, 34]
	Alice	600	2	3	5	Ifran	[-7, 32]
	Guerov	1000	2	2	2	Casa	[3, 39]
	Jilali	1000	3	2	3	Midelt	[-10, 46]
	Eve	500	3	4	5	Agadir	[5, 39]
	Driss	700	1	2	4	Marrakesh	[2, 49]
	Omar	700	3	3	4	Dakhla	[4, 41]
	Mick	800	2	4	3	Ouazzane	[-2, 40]

For the anonymization of the data, we use the following three algorithms: (Figures 1, 2 and 3)

Algorithm 1 Generalization

Input String str = "$[a, b]$" and float m
Output List of variables
$a \leftarrow$ ' '
$b \leftarrow$ ' '
if type(str) == "String" **then**
  $i \leftarrow 1$
  while $str[i] \neq$ "," **do**
    $a \leftarrow a + str[i]$
    $i \leftarrow i + 1$
  end while
  for $j \leftarrow i + 1$ **to** len(str)-1 **do**
    if str[i]== " " **then**
      $b \leftarrow b + str[i]$
    end if
  end for
  $a \leftarrow float(a)$
  $b \leftarrow float(b)$
  $x \leftarrow a - \frac{m*(b-a)}{2}$
  $y \leftarrow b + \frac{m*(b-a)}{2}$
else
  $a \leftarrow float(str)$
  $x \leftarrow a - m * a$
  $y \leftarrow a + m * a$
end if
return $[x, y]$

Figure 1: Algorithm of Generalization

Algorithm 1 is used to generalize continuous variables which can be in the form of intervals or integers. It is noteworthy that algorithm 2 is used for the restriction in the same case. We use generalization if m > 0.5 and restriction if m < 0.5.

Algorithm 2 Restriction

Input String str = "$[a, b]$" and float m
Output List of variables
$a \leftarrow$ ' '
$b \leftarrow$ ' '
if type(str) == "String" **then**
  $i \leftarrow 1$
  while $str[i] \neq$ "," **do**
    $a \leftarrow a + str[i]$
    $i \leftarrow i + 1$
  end while
  for $j \leftarrow i + 1$ **to** len(str)-1 **do**
    if str[i]== " " **then**
      $b \leftarrow b + str[i]$
    end if
  end for
  $a \leftarrow float(a)$
  $b \leftarrow float(b)$
  $x \leftarrow a + \frac{m*(b-a)}{2}$
  $y \leftarrow b - \frac{m*(b-a)}{2}$
else
  $a \leftarrow float(str)$
  $x \leftarrow a - m * a$
  $y \leftarrow a + m * a$
end if
return $[x, y]$

Figure 2: Algorithm of Restriction

Algorithm 3 is used, mainly, to modify the data.

Table 4 shows the data after being modified using the algorithms 1, 2 and 3

Algorithm 3 Modification

```
Input Data excel
Output Data excel
T ← read − excel("MF.xlsx")
N ←number of lines
M ←number of columns
p ←MIN of column
q ←MAX of column
for j ← 0 to M do
    m = MF[1,j]
    if MF[0,j] == "list" then
        if m > 0.5 then
            for i ← 2 to N do
                T[i, j] ← Generalization(MF[i, j], m)
            end for
        else
            for i ← 2 to N do
                T[i, j] ← Restriction(MF[i, j], m)
            end for
        end if
    else
        if MF[0,j] == "float" then
            for i ← 2 to N do
                T[i, j] ← p + (MF[i, j] + floor(MF[i, j] ∗ m) ∗ (−1)^i) mod(q)
            end for
        else
            k ← (i + floor(N ∗ m)) mod(N − 2)
            for i ← 2 to N do
                tmp ← T[i, j]
                T[i, j] ← T[k + 2, j]
                T[k + 2, j] ← tmp
            end for
        end if
    end if
end for
return [x, y]
```

Figure 1: Algorithm of Generalization

Table 4: the modified data

name	Monthly salary	Athletic	Academic levels	Number of children	City	city temperature
****	[551, 2449]	2	2	3	Midelt	[-9.284, 44.284]
****	[184, 816]	3	1	4	Agadir	[-23.2, 63.2]
****	[625, 2775]	2	2	2	Marrakesh	[-16.468, 47.468]
****	[221, 979]	1	1	2	Dakhla	[-21.196, 46.196]
****	[368, 1632]	2	3	3	Ouazzane	[-10.104, 52.104]
****	[368, 1632]	3	3	4	Rabat	[-30.384, 66.384]
****	[184, 816]	2	4	5	Fes	[-7.376, 51.376]
****	[257, 1143]	2	3	1	Tangier	[-15.108, 66.108]
****	[257, 1143]	2	3	4	Ifran	[-9.468, 54.468]
****	[294, 1306]	1	2	4	Casa	[-17.288, 55.288]

The next tables, 5, 6 7, 8, 9 and 10, give a comparison of the statistical indicators of the data before and after the modification.

Table 5: The Means arithmetic

Columns	Monthly salary	Athletic	Academic levels	Number of children	City temperature
Means before	900.0	2.1	2.5	3.2	19.35
Means after	900.0	2	2.4	3.2	19.35

Table 6: The Range

Columns	Monthly salary	Athletic	Academic levels	Number of children	City temperature
Range before	1000	2	3	4	50
Range after	2591	2	3	4	96

Table 7: The variance

Columns	Monthly salary	Athletic	Academic levels	Number of children	City temperature
Variance before	152000	0.69	1.05	1.56	12.91
Variance after	152000	0.4	0.84	1.36	12.91

We notice that the arithmetic Means and the Variance are almost the same, and the Ranges are the same only in the discrete variables.

Table 8: Covariance

Columns	Athletic and Academic levels	Academic levels and Number of children	Number of children and Athletic
Covariance before	0.55	0.9	0.48
Covariance after	0.1	0.32	0.2

In table 8, we notice that the covariances are decreasing, but are always positive.

Table 9: Correlation

Columns	Athletic and Academic levels	Academic levels and Number of children	Number of children and Athletic
Correlation before	0.65	0.70	0.46
Correlation after	0.17	0.30	0.27

The correlations between columns became weak after the modification, so we can say that the variables become almost independent.

Table 10: Correlation before and after

Columns	Athletic	Academic levels	Number of children
Correlation before and after	0.38	0.11	0.18

The modified data are almost independent from the initial data

5. DISCUSSION

Several studies have been proposed to preserve data and improve its usefulness, especially when analyzing the data while preserving its privacy. Most of this research is based on clustering techniques and algorithms to apply anonymization models on private data. In [10], for instance, the authors proposed a hybrid strategy that combines self-organizing maps with conventional privacy-based clustering algorithms OKA & KMA. In [11], the study introduces an evaluation and comparison between different privacy-preserving techniques in different fields including big data.

In this paper we assigned a coefficient m that represents the degree of importance of the information. This coefficient is calculated through the degrees of identification and sensitivity using a new method which is based on fuzzy logic, namely the Yager t-norm and t-conorm. The results in [12] reveal that the use of parameters in the algorithms plays a significant role in data anonymization. In this regard, our algorithms depend mainly on the coefficient m to anonymize the data.

In light of the results and the previously published work, the anonymization of the data in this study has several advantages: First, it supports fast and easy calculations. Second, it leads to good anonymization of the data. Finally, it helps in the preservation of statistical rates.

Our algorithm is based on the property of permutations in the case of text data, which negatively affects the statistical averages conditioned by this text data. However, to solve problem, the data owner can either delete the data, in case it is identifying, or keep it, in case it is non-identifying.

6. CONCLUSION

In this study, we propose a fuzzy-logic based model to assess the importance of the information, according to their identification and sensitivity, by assigning a degree of importance, and subsequently constructing a modification function. Our proposal will be useful for the data owner if this latter keeps the degree of modification secret, but the degree of modification can itself be considered as an information leak vulnerable to a possible attack.

REFERENCES:

[1]. S.S. Eliabeth and S. Sarju, "Bigdata Anonymization Using One Dimensional and Multidimensional Map Reduce Framework on Cloud", *International Journal of Database Theory and Application*, Vol. 8, No. 6, 2015, pp. 253-262.

[2]. L. Sweeney, "k-anonymity: A model for protecting privacy", *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, Vol. 10, No. 5, 2002, pp. 557-570.

[3]. K.B. Hemanta, K.K Narendra, and K.P. Narendra, "Individual privacy in data mining using fuzzy optimization", *Engineering Optimization*, Vol. 54, No. 8, 2022, pp. 1-6.

[4]. S.U. Bazai, J. Jang-Jaccard and H. Alavizadeh, "Scalable, High-Performance, and Generalized Subtree Data Anonymization Approach for Apache Spark", *Electronics*, Vol. 10, No. 5, 2021, pp. 589.

[5]. L. Xu, C Jiang, J. Wang, J. Yuan and Y. Ren, "Information security in big data: privacy and data mining", *Ieee Access*, Vol. 2, 2014, pp. 1149-1176.

[6]. D. Zhang, "Big data security and privacy protection", *Proceedings of 8th International Conference on Management and Computer Science*, 2018, pp. 275-278.

[7]. C. Eyupoglu, M. Aydin, A. Zaim and A. Sertbas, "An efficient big data anonymization algorithm based on chaos and perturbation techniques", *Entropy*, Vol. 20, No. 5, 2018, pp. 373.

[8]. L.A. Zadeh, "Fuzzy sets", *Information and control*, Vol. 8, No. 3, 1965, pp. 338-353.

[9]. J. Dombi, "A general class of fuzzy operators, the DeMorgan class of fuzzy operators and fuzziness measures induced by fuzzy operators", *Fuzzy sets and systems,* Vol. 8, No. 2, 1982, pp. 149-163.

[10]. K. MOHAMMED, A. AYESH, and E. BOITEN, "Complementing Privacy and Utility Trade-Off with Self-Organising Maps". *Cryptography*, 2021, vol. 5, no 3, p. 20.

[11].A.Tamer, M.H. Khafagy, M.H. Farrag "Big data challenges : preserving techniques for privacy violations" *Journal of Theoretical and Applied Information Technology,* Vol. 100, No. 8, 2022, pp. 2505-2517.

[12]. C. Ni, Li Shan Cang, P. Gope et al." Data anonymization evaluation for big data and IoT environment" Information Sciences, Vol. 605, 2022, pp. 381-392.