Engineering Optimization

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/geno20

Individual privacy in data mining using fuzzy optimization

Hemanta Kumar Bhuyan, Narendra Kumar Kamila & Subhendu Kumar Pani

To cite this article: Hemanta Kumar Bhuyan, Narendra Kumar Kamila & Subhendu Kumar Pani (2021): Individual privacy in data mining using fuzzy optimization, Engineering Optimization, DOI: 10.1080/0305215X.2021.1922897

To link to this article: https://doi.org/10.1080/0305215X.2021.1922897

Published online: 20 May 2021.

ENGINEERING OPTIMIZATION
https://doi.org/10.1080/0305215X.2021.1922897

Taylor & Francis
Taylor & Francis Group

Individual privacy in data mining using fuzzy optimization

Hemanta Kumar Bhuyan[a], Narendra Kumar Kamila[b] and Subhendu Kumar Pani[c]

[a]Department of Information Technology, Vignan's Foundation of Science, Technology and Research (VFSTR), Deemed to be University, Andhra Pradesh, India; [b]Department of Computer Science and Engineering, BRMIIT, Bhubaneswar, Odisha, India; [c]Krupajal Computer Academy, BPUT, Bhubaneswar, Odisha, India

ABSTRACT

This article proposes the individual data privacy during collaborative computation in data mining method using an optimization model. The privacy problem is solved using different methodologies. The solution for individual privacy is considered as a multi-objective optimization model. Practically, the requirement for privacy varies from user to user. Therefore, it generates inherent vagueness for individual privacy. In this article, the vagueness is considered and the privacy problem is solved by a fuzzy optimization method. The fuzzy multi-objective optimization model is proposed to be used as a supplementary privacy method to address individual privacy issues. The fuzzy constraints are generated to solve the models on the basis of the privacy requirements of users. The fuzzy set domain for the optimization problem is used to fulfil the individual privacy requirements in a computing environment. The proposed solution allows data owners to choose their own privacy level on demand, with maximum flexibility.

ARTICLE HISTORY
Received 24 August 2020
Accepted 21 April 2021

KEYWORDS
Multi-objective optimization; fuzzy optimization; privacy; fuzzy constraints; data mining

Notation	Explanation
$f_1, \ldots, f_M$	scalar objectives with $f_i : R^m \rightarrow R$,
P_j and Q_k	mapping $R^m \rightarrow R$ for both constraint functions, $g, h \in R$
x_i^l and x_i^u	variables in a precisely bounded solution
$U_i(x)$	respective dimensions of the utility function
U_{HP}	maximization of privacy
U_{CC}	minimization of computational cost
C_{HP}, C_{CC}	measurement costs
a_i^{HP}, a_i^{CC}	weights of each individual utility component
C_1 and C_2	confidence limits
(α)	amount of privacy satisfying the level of significance
$[C_1, C_2]$	confidence interval
Z	expected to lie in the interval
$1 - \alpha$	confidence coefficients depending on the desired precision
μ	mean
σ^2	variance
v	large, random-size sample from a large database
m, n	set of integer numbers
$x_{i,}^u$	feasible solutions
w_1 and w_2	relative weights that a data miner uses for data privacy and cost

CONTACT Hemanta Kumar Bhuyan hmb.bhuyan@gmail.com

x	simplicity and prototype for optimization problem
s_1 and s_2	weights for scalar optimization problem
$x*$	critical solution vectors
g_1 and g_2	invertible functions; the variation of weights lies between $[g_1^{-1}(x_1^l), g_1^{-1}(x_1^u)]$ and $[g_2^{-1}(x_2^l), g_2^{-1}(x_2^u)]$
$\leq^{DM}$	data mining in conditional relationship
$p_{i(t)}$	degree of decision maker's requirements on the ith objective
$P_i(x)$	good compromise to solutions for the ith objective
$v_i(t)$	fuzzy function
m_i and M_i	independent optimal value of the ith objective
(x')	optimal compromise solutions
$P_k(\alpha)$	α-cut of the kth constraint
$g_k(\alpha)$	α-cut for fuzzy constraints
d_k and $(1-\alpha)$	maximum deviation from the original constraints, according to the demands of participating users
P_f	Pareto function
X	non-dominated set
Y	optimized value
P	entire search space
P'	Pareto-optimal set
M	objective function
K	public key
v_i	information shared through site
S	sum (v_i)

1. Introduction and related works

Privacy-preserving data mining (PPDM) is an emerging field in data-mining research, where cost and privacy are issues for secure computation in distributed data-mining systems (Kumar and Mohbey 2019). In such systems, the data owner and data miner are under well understanding for the data-mining task during the release of private data by the data owner (Mendes and Vilela 2017). In many cases, they maintain their monolithic (common/unique) privacy where the data owner does not choose their own privacy during distributed computation (Shajin and Rajesh 2020). Under these circumstances, several options are available to the data owner regarding the privacy of their own private data (Rajesh and Shajin 2020). Therefore, this type of data-mining problem can be considered as an optimization problem to solve individual privacy issues (Zhao *et al.* 2018). Sin *et al.* (2020) analyze General Model for Privacy-Preserving Data Mining which helped to privacy model of data during mining. Purohit and Bhargava (2017) presented a data-mining approach to extract valuable data from wide multiple-information sources. Jahan *et al.* (2018) presented a data-mining strategy which examines the patterns recognized in data, irrespective of individual secret information, and demonstrated that the suggested model was better in securing secret information. Chen, Panahi, and Pourghasemi (2017) suggested an innovative ensemble data-mining strategy depending on a geographical information system, which included an adaptive neuro-fuzzy inference system, differential evolution and particle swarm optimization for landslide spatial modelling. Lekshmy and Rahiman (2020) applied the ant bee colony algorithm for optimum key generation as well as encryption of huge amounts of data. Langari *et al.* (2020) presented a combined approach depending on the k-member fuzzy clustering and firefly algorithm. To obtain individual privacy in data-mining applications, it is necessary to propose a multi-objective optimization model (Transpire Online 2020). Each party will try to maintain the optimal data privacy, with a cost for preserving privacy. Here, the optimization model has been developed based on (1) maximizing privacy and (2) minimizing the total computational cost (Pellungrini *et al.* 2017).

In multi-objective optimization models, identical objective functions with dissimilar constraints are considered for each data owner according to their own interests (Bakhtavar *et al*. 2020). A detailed description of this model is presented in Section 3 for better understanding of the proposed work.

To the best of the authors' knowledge, fuzziness of data in real-world applications is dealt with for the first time in the present model. Here, a theoretical framework of fuzzy sets is considered to find the solution to extensive problems (Babaee Tirkolaee *et al*. 2020). Fuzzy constraints have been taken into consideration for generating the optimal solution (Stojiljković 2017). The two cases of optimal solution are: (1) the membership function that derives an independent optimal value on the objectives and (2) the membership function that derives fuzzy constraints using the α-cut to generate a well-optimized solution (Lamata, Pelta, and Verdegay 2018). Bhuyan *et al.* (2019) used optimal model for sub-feature selection and analysis of cost.

Different solutions have been obtained for several data-mining problems using privacy techniques (Christen *et al*. 2018). Bhuyan *et al.* (2011), have used own model of privacy on distributed data. These approaches treat PPDM as a special case of secure multi-party computation both for the preservation of individuals' privacy and to prevent the outflow of any information except for the final result. Similarly, Kamila, Jena, and Bhuyan (2016) used Pareto-based multi-objective optimization for classification in data mining. The certain interval of featured data was also considered to distinguish the differentiated class from the existing class in data mining (Bhuyan and Reddy 2018). The privacy of data was preserved with different techniques in quantifying differential privacy and privacy preservation in encrypted graphs (Sharma, Powers, and Chen 2018). Bhuyan *et al.* (2012) developed privacy preservation for Sub-feature Selection in data mining.

PPDM is defined as a model as well as a concept of data mining that ensures that data are not revealed to unauthorized users when receiving essential information from the data warehouse. Data mining is a strategy that is used to deal with large data counts. The data diminution method reduces the difficulty in dealing with large amounts of data. Bhuyan and Huque (2018) have developed sub-feature selection model, where sub-feature data can be maintained for privacy under this model. The main goal is to improve the storage efficiency and decrease the data storage and costs. A major challenge in execution is the ability to integrate conflicting or unnecessary data from various sources. To overcome this problem, the fuzzy optimization algorithm is used. Cao *et al.* (2019) used quantifying differential privacy in continuous data with correlation model.

The remainder of this article is organized as follows. Section 2 presents the problem statement for the optimization model based on decision variable constraints. Section 3 illustrates the multi-objective optimization framework based on the measurement of privacy and cost optimization. Section 4 defines the challenging optimal model with the help of fuzzy constraints. Section 5 explains the Pareto-optimal set for multi-objective optimization. Section 6 presents the experimental procedure and results of the proposed optimal model. Finally, Section 7 concludes the article.

2. Problem statement

Privacy is always an important issue in PPDM, for both data miner and data owner. The data miner sometimes struggles to provide a solution through which the data owner can choose their own privacy settings during the release of data. Hence, privacy measurement plays a major role in choosing the appropriate level of privacy for each data owner. Under these circumstances, the optimization problem provides a good solution, in which the data owner can choose the maximum privacy for their data, with minimum computational cost to pay to the service provider.

The multi-objective optimization model is developed based on multiple, possible diverging objectives. Mathematically, it is expressed as:

$$\text{Max}\, f(x) = [f_1(x), \ldots, f_M(x)]^T \tag{1}$$

Under

$$P_j(x) \leq g, \forall j = 1 \ldots .P$$
$$Q_k(x) = h, \quad \forall k = 1 \ldots q$$
$$x_i^l \leq x_i \leq x_{i,}^u \quad \forall i = 1 \ldots m$$

where $f_1, \ldots, f_M$ indicates M scalar objectives with $f_i : R^m \rightarrow R$, P_j and Q_k denote mapping $R^m \rightarrow R$ for both constraint functions, g, $h \in R$, and variables are precisely bounded between x_i^l and x_i^u. The solution is considered as a vector $x' = \{x_{1,}x_{2,} \ldots, x_m\} \in R^m$ for the above multi-objective optimization problem.

The following objectives are created by the data miner to fulfil the utility function: (1) maximize the privacy U_{HP}; and (2) minimize the computational cost U_{CC}. The following decision variable constraints are required to develop the utility function for the model.

2.1. Decision variable constraints

(1) I^{HP} = length of interval for required privacy, *i.e.* high privacy (HP) or low privacy (LP).
(2) I^{CC} = computational cost, *i.e.* high priority-based objective with low computational cost.

In this model, each data owner chooses their own strategies and decides on deterministic actions to maximize their utility scores. According to the respective dimensions of utility function $U_i(x)$, Equation (1) can be rewritten as

$$\text{Max}U_i(x) = \left[\sum C_{HP}U_{HP,} \sum C_{CC}U_{CC}\right] \tag{2}$$

where the usual notations have been defined successively along the required constraints. The weighted linear combination of the above dimensions of the utility function for multi-objective optimization can be considered mathematically as

$$\text{Max}U_i(x) = a_i^{HP} \sum C_{HP}U_{HP} + a_i^{CC} \sum C_{CC}U_{CC} \tag{3}$$

subject to

$$\sum I_i^{HP}(x) \leq g$$
$$\sum I_i^{CC}(x) \leq h$$
$$\sum a_i^{HP} + \sum a_i^{CC} = 1$$
$$x_i^l \leq x \leq x_{i,}^u \text{ for all } i = 1, 2, \ldots . n$$

where U_{HP}, U_{CC} are utility factors of privacy and computational cost; C_{HP}, C_{CC} are the measurement cost (*i.e.* how much measured privacy and cost) of each individual utility factor; and a_i^{HP}, a_i^{CC} are the weights of each individual utility component.

3. Multi-objective optimization framework

The multi-objective optimization framework is established in this section. This is a mathematical approach to privacy protection in the process of data mining from distributed databases. Here, n parties are involved in general transaction implementation. The parties calculate the data value in secure

mode in a given manner: let n sites be included in communication through a distributed environment. Moreover, a common public key K is shared. Every site contains objects to share with others.

$$x = sum(v_i) \tag{4}$$

where v_i indicates data that are shared through the site along residual $n-1$ sites. Here, site A specifies the initialization site, which also maintains the secret key. Site A determines the value S by modular arithmetic as:

$$S = ((K + vA) \bmod N/K) \tag{5}$$

where N denotes the range of site v_i, and A can send the determined value to the subsequent site B. Likewise, site B can determine the subsequent value in a similar form:

$$S = \left(\left(K + \sum_{j-1}^{n} v_i\right) \bmod N\right)/k \tag{6}$$

where site A knows the secret key K, it is simply recognized as sum S. This process is utilized in an easy tool to secure data access with transfer while mining across distributed sites.

3.1. Measurement of privacy

Here, the estimation of privacy of each data owner is considered using confidence intervals. The confidence interval is used for evaluating the amount of privacy satisfying the level of significance (α).

Definition 1: An interval denotes a confidence interval if the estimation of this interval satisfies the level of significance (α).

For example: Let the interval $[C_1, C_2]$ be generated by two constants C_1 and C_2 at the level of significance where the original data lie. This means that

$$P(C_1 < Z < C_2) = 1 - \alpha \tag{7}$$

where Z is the standard normal distribution which is expected to lie in the confidence interval and P is the probability distribution. Here, the interval $[C_1, C_2]$ is said to be the confidence interval, where the unknown value of parameter Z is expected to lie within the interval. The amount of privacy on the confidence limits using statistical data is discussed as follows.

A large, random-size sample v is considered from a large database along with mean μ and variance σ^2; the sample mean is $\bar{x} \sim N(\mu, \sigma^2/n)$, *i.e.* $Z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$. Using the Gaussian distribution, the confidence interval is determined as

$$P\left(\bar{x} - C\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + C\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha \text{ for interval } [-C, C].$$

where the interval $\left[\bar{x} - C\frac{\sigma}{\sqrt{n}}, \bar{x} + C\frac{\sigma}{\sqrt{n}}\right]$ is called the confidence interval, using the mean for individual data in the database. Individual data are used to identify the individual feature data in the database. In particular, for a significance value of 1.96 and confidence coefficient of 0.95, the confidence interval

can be determined as

$$P(-1.96 \leq Z \leq 1.96 = 0.95)$$
$$\Rightarrow P\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

where the 95% confidence limits for large data mean are $\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$; σ is assumed to be known and the interval $\left[\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right]$ is the 95% confidence interval for estimating μ. The above statistical data are utilized to measure the amount of privacy within certain confidence limits.

3.2. Privacy and cost optimization

The optimal cost for high privacy is always solicited for the proposed optimization model because the computational cost is estimated by executing the algorithms for a system. Here, $x \in R^m$ indicates a multi-dimensional input vector to define the optimum privacy and cost identified by the computational costs for different levels of privacy. Based on the above objectives, the optimization problem is stated as

$$\text{Max} f(x) = [f_{HP}(x), f_{CC}(x)]^T \tag{8}$$

subject to

$$x_i^l \leq x_i \leq x_{i,}^u \text{ for all } i = 1 \ldots m$$

where $x \in R^m$, and each x_i is bounded between x_i^l and $x_{i,}^u$ for feasible solutions. The same optimization problem can be reformulated into a scalar optimal problem:

$$\text{Max } F = w^T f(x) = [w_1 f_{HP}(x) + w_2 f_{CC}(x)] \tag{9}$$

under

$$x_i^l \leq x_i \leq x_i^u \quad \forall i = 1 \ldots m$$
$$w_1 + w_2 = 1$$
$$w_1, w_2 \geq 0$$

where w_1 and w_2 denote the relative weights that the data miner uses for data privacy and cost, respectively. Using scalarization techniques, the above-mentioned optimization problem is converted into a scalar objective function with identified weights. Assuming that $x = (x_1, x_2) \in R^2$, which is considered for simplicity and as a prototype for the optimization problem:

$$F_p = s_1 f_1(x_1, x_2) + s_2 f_2(x_1, x_2) \tag{10}$$

where s_1 and s_2 are two weights for the scalar optimization problem. The derivative of the above equation must be computed and set to zero. The solutions to the resulting equations provide critical solution vectors $x* = (x_1^*, x_2^*)$ using s. The first order derivatives are considered for solutions that provide a mapping from the vector space to the weight space. Considering two possibly nonlinear functions, $g_1 : R^m \rightarrow R$ and $g_2 : R^m \rightarrow R$, that map from the weight to the objective variables, it can be stated that:

$$x_1^* = g_1(s_1, s_2) \tag{11}$$

$$x_2^* = g_2(s_1, s_2) \tag{12}$$

By $x_i^l \leq x_i \leq x_i^u$, the range of each s based on input objective variables can be generated. Here, the functions g_1 and g_2 are invertible; then, the variation of weights lies between $[g_1^{-1}(x_1^l), g_1^{-1}(x_1^u)]$ and $[g_2^{-1}(x_2^l), g_2^{-1}(x_2^u)]$, which provides the list of solutions in the Pareto-optimal set.

4. Fuzzy constraints for multi-objective optimization

The data from different stakeholders have different characteristics, such as random, vague and fuzzy, which are used in the optimization problem. The demands of stakeholders are also fuzzy. In this article, the proposed model deals with the fuzzy multi-objective optimization issue with fuzzy constraints. The complex fuzzy optimization problem is stated as:

$$\text{Max}Y = cx \tag{13}$$

subject to

$$(Ax)_i \leq b_i \text{ for all } i = 1, 2, \ldots .m$$
$$x_j \geq 0, x_j \in N, j = 1, 2, \ldots .n$$

where m, n represents a set of integer numbers, $c \in R^n$, $A = \sum_j a_{ij,}$ and $a_{ij,} b_i \in R$.

Let the constraints defining the issue have a fuzzy nature where the decision maker/data miner is willing to permit any change ($\leq^{DM}$) over restricted constraints:

$$\text{Max}Y = cx \tag{14}$$

subject to

$$(Ax)_i \leq DM_{bi,} \quad i = 1, 2, \ldots, m$$
$$x_j \geq 0, x_j \in N, j = 1, 2, \ldots, n$$

where $\leq^{DM}$ refers to the data-mining conditional relationship.

However, some typical approaches are required to solve multi-objective decision-making (MODM) problems with supportive and conflicting objectives. It is very complicated to choose the optimal decision with an increasing number of objectives.

$$\text{Max}\{< C^i, x \geq Y_i\} \tag{15}$$

under

$$x \in X = \{x \in R^n | Ax = b, x \geq 0, b \in R^m |\}$$

An innovative method is developed for solving MODM problems depending on the interdependences among the multiple objectives. Hence, the following two cases are considered.

Case 1

The optimization problem is defined on the multi-objective function:

$$\text{Max}\{f_1(x), f_2(x), \ldots .., f_k(x)\}, x \in X \tag{16}$$

where $f_j : R^n \rightarrow R$ is the objective functions, $x \in R^n$ is a variable, then $X \subset R^n$.

Let a function $p_{i(t)} : R \rightarrow [0, 1]$, where $p_{i(t)}$ determines the degree of the decision maker's requirements on the ith objective of value t. The membership of x degree in the fuzzy set using $p(x)$

is:

$$p_i(x) = p_i(Y(x)) \tag{17}$$

where $P_i(x)$ is considered as a good compromise solution for the ith objective. So, it is quite reasonable to search for the solution of the following auxiliary problem:

$$\begin{aligned} &\max\{P_1(x), \ldots P_K(x)\}, \\ &x \in X \end{aligned} \tag{18}$$

where $P_i(x) \in [0, 1]$. A single-objective problem:

$$\begin{aligned} &\max T\{P_1(x), \ldots P_K(x)\}, \\ &x \in X \end{aligned} \tag{19}$$

Now, the membership functions of the proposed functions can be considered as:

$$p_i(t) = \begin{cases} 1 & \text{if} \quad t \geq M_i \\ v_i(t) & \text{if} \quad m_i \leq t \leq M_i \\ 0 & \text{if} \quad t \leq m_i \end{cases} \tag{20}$$

where $m_i = \min\{Y_i(x)|x \in X\}$ and $M_i = \max\{Y_i(x)|x \in X\}$ with independent minimal and maximal values of the ith objective, and $v_i(t)$ is a fuzzy function. For the linear membership functions, P_i is defined as

$$p_i(x) = \begin{cases} 1 & \text{if} \quad Y_i(x) \geq M_i \\ 1 - \dfrac{M_i - Y_i(x)}{M_i - m_i} & \text{if} \quad m_i \leq Y_i(x) \leq M_i \\ 0 & \text{if} \quad Y_i(x) \leq m_i \end{cases} \tag{21}$$

Several objective functions are used to determine the optimal solutions, which are close to the independent minima and maxima. For optimal compromise solutions, (x') is efficient if: (1) x' is unique, (2) T is strict, and (3) $0 < P_i(x') < 1$, for $i = 1, 2, \ldots .k$, and also with strictly increasing function on $[m_i, M_i]$.

The application of fuzzy constraints makes a powerful tool available to determine the solution of the complex fuzzy issue. The constraints are permitted to change the constraints for the user's satisfaction. The constraints are reformulated as

$$m_i \leq \sum I_i^{HP}(x) \leq g_i + M_i$$

$$m_i \leq \sum I_i^{CC}(x) \leq h_i + M_i$$

where m_i and M_i are independent optimal values of the ith objective. This also makes the minimum and maximum deviation from the original range according to the linear membership function.

Case 2

In the second case, the constraints are defined with a fuzzy nature for the violation of restrictions. The constraints of the problem are derived with membership functions by the decision maker to avoid breaching the achievement of the constraints. The membership functions are considered with fuzzy constraints, as follows:

$$\mu_k : R^n \rightarrow (0, 1], k \in I$$

where the degree is 1 for the actual constraint. When the violation increases, the degree is reduced to 0, accordingly. Finally, the degree is implied to be 0 in all cases of non-violations. Under this condition,

the membership functions are stated as follows:

$$\mu_k(b_k) = \begin{cases} 1 & (Ax)_k \le b_k \\ 1 - \dfrac{(Ax)_k - b_k}{d_k} & b_{k\le}(Ax)_k \le b_k + d_k \\ 0 & (Ax)_k > b_k + d_k \end{cases} \tag{22}$$

For the above problem, each fuzzy constraint is defined as

$$P_k = \{x \in R^n | (Ax_k) \le DMb_k x \ge 0, x \in N\} \tag{23}$$

where $x \in P$ and $P = \cap_{k\in I} P_k$.

$$\text{Max}\{Y = cx | x \in X\}$$

It is obvious that for all $\alpha \in (0, 1]$, the α-cut of the fuzzy constraint set can be obtained, which is called the classical set:

$$P(\alpha) = \{x \in R^n | \mu_p(x) \ge \alpha\} \tag{24}$$

Hence, the α-cut of the kth constraint is denoted by $P_k(\alpha)$ for all $\alpha \in (0, 1]$:

$$M(\alpha) = \{x \in R^n | cx = \max cz, z \in P(\alpha)$$

The fuzzy set defined by the following membership function is the fuzzy solution for all $\alpha \in (0, 1]$:

$$M(\alpha) = \begin{cases} \sup\{\alpha : x \in M(\alpha)\} & x \in U_\alpha M(\alpha) \\ 0 & \text{otherwise} \end{cases} \tag{25}$$

$$P(\alpha) = \cap_{k\in I}\{x \in R^n | (Ax)_k \le g_k(\alpha), x \ge 0, x \in R^n\} \tag{26}$$

where $g_k(\alpha) = b_k + d_i(1 - \alpha)$ is the α-cut for fuzzy constraints:

$$\begin{aligned} &\text{Max} Y = cx \\ &s.t(Ax)_k \le b_k + d_k(1 - \alpha), k \in I, \alpha \in [0, 1] \end{aligned} \tag{27}$$

The above methodology can manage the fuzzy constraints to overcome the problem to provide a better solution. The complete multi-objective fuzzy optimization problem using α-cut fuzzy constraints is follows:

$$\text{Max} U_i(x) = a_i^{HP} \sum C_{HP}(U_{HP}) + a_i^{CC} \sum C_{cc}(U_{cc}) \tag{28}$$

subject to

$$\begin{aligned} &\sum I_i^{HP}(x) \le g_i + (1 - \alpha)d_i \\ &\sum I_i^{CC}(x) \le h_i + (1 - \alpha)d_i \\ &\sum a_i^{HP} + \sum a_i^{CC} = 1 \\ &x_i^l \le x_i \le x_i^u, \quad \forall i = 1 \ldots n \end{aligned}$$

where d_k and $(1 - \alpha)$ refer to the maximum deviation according to the demand of participating users from the original constraints. The privacy of each party depends on d_k and $(1 - \alpha)$ as per their deviation. Similarly, utility factor U_{CC} relates to the computational cost, which is determined by the execution of different data-mining algorithms. The two terms a_i^{HP} and a_i^{CC} implicate weight factors, which satisfy $a_i^{HP} + a_i^{CC} = 1$. Here, $a_i^{HP} = 1 - a_i^{CC}$ or $a_i^{CC} = 1 - a_i^{HP}$ is the appropriate design of the optimal test.

5. Pareto-optimal set for multi-objective optimization

This is an ideal methodology of the optimization problem for finding optimal solutions. As the problem contains multiple objectives, it provides a set of optimum solutions (Pareto-optimal solutions), as opposed to a single optimum solution. The following definitions are required for solving the above problem.

Definition 2 (**Pareto function**): This is a function from a non-dominated set of data to the optimized value. Mathematically, it can be written as

$$P_{f:}X \rightarrow Y$$

where P_f is the Pareto function, X is the non-dominated set, and Y is the optimized value. The Pareto function can be assessed by different data sets.

5.1. Non-dominated set and Pareto-optimal set

Let M objective functions exist in the multi-objective problems. With the purpose of covering the maximization and minimization objective functions, the operator $\triangleleft$ is used between the two solutions g and h as $g \triangleleft h$.

The given definition includes mixed complexity in both maximization and minimization objective functions.

Definition 3: The x^1 solution dominates the x^2 solution, if two conditions are met:

(a) The x^1 solution is better than x^2 in all objectives of $f_j(x^1) \not\triangleleft f_j(x^2)$ for all $j = 1, 2 \ldots M$.
(b) The solution x^1 is purely greater than x^2 for at least one objective, or $f_j(x^1) \triangleleft f_j(x^2)$ for at least one $j \in \{1, 2, \ldots, M\}$.

The solution x^1 is not dominates the solution of x^2 to violate the conditions. Conversely, it can specify the following:

- x^2 dominates by x^1
- x^1 dominates by x^2
- x^1 is non-inferior to x^2.

The dominance property is analysed for the proposed optimization model, where two objectives (privacy and computational cost) are focused on, with different solutions. An optimization problem with two objectives, *i.e.* maximization and minimization (maximum privacy and minimum computational cost), with five different solutions, is considered here, as shown in Figure 1. Here, the dominant nature decides the better solution of the objectives. From the observation solution 1 is considered as more effectual than solution 2 in the first and second objective functions. According to the dominant nature, it may be considered that solution 1 dominates solution 2. Finally, solution 5 dominates solution 1.

5.2. Pareto optimality

Figure 1 shows that solution 5 is greater than solution 3 for the first objective, and *vice versa* for the second objective. These two solutions do not satisfy the first condition. Similarly, for both objectives, the both solutions need not be dominated to each other. Based on this situation, solutions 3 and 5 are the non-dominated set.

To compare a finite set of solutions, the performance is measured on the dominated and non-dominated sets of solutions. If the comparison is not satisfied in dominant nature, then the set is known as a non-dominated set for a given set of solutions. For example, solutions 3 and 5 are said

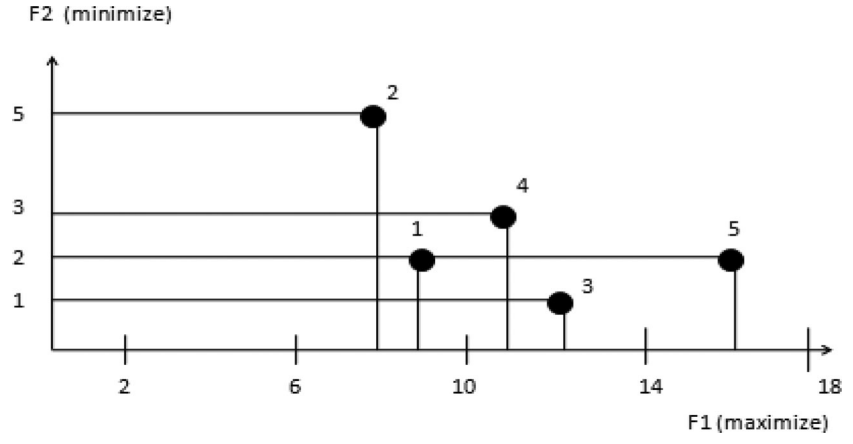

Figure 1. Dominance among five solutions with two objectives.

to be a non-dominated set from the above five solutions. The definition of a set of non-dominated solution is:

Definition 4(a) (non-dominated set): From any given set of solutions P, the set of solution P' is said to be the non-dominated set, if P' solutions are not dominated by any member of set P.

Definition 4(b) (Pareto-optimal set): If set P is the total search space or $P = S$, the resultant non-dominated set P' is named the Pareto-optimal set.

Sometimes, the dominance is weak for a few solutions, so modification of Definition 1 is needed for a strong dominance relation, as follows.

Definition 5: An x^1 solution robustly dominates the x^2 solution (or $x^1 \prec x^2$), if the x^1 solution is purely greater than the x^2 solution in every objective of the given problem.

Referring to Figure 1, it is observed that solution 5 does not strongly dominate solution 1, here identified that the solution 5 weakly dominates solution 1. However, solution 3 strongly dominates solution 1, because solution 3 is more proficient than solution 1 in both objectives. If solution x^1 robustly dominates solution x^2, solution x^1 does not weakly dominate solution x^2. The strong dominance operator has the same properties as the dominance relation.

The above definition of strong dominance can be used to define a weakly non-dominated set.

Definition 6: (weakly non-dominated set): In a set of solutions P, P' represents the weakly non-dominated set, if the solutions are not strongly dominated by any other member of set P.

For a given set of solutions, the cardinality of the non-dominated set is obtained using Definition 4.

Based on the above arguments, it can be seen that the Pareto-optimal set is identified from the non-dominated set. But this is not always the case. The non-dominated solutions obtained by the optimization approach may not be a proper Pareto-optimal set. The proposed method is sensitive to the values of its main controlling parameter.

6. Computational experiments

Computational experiments are analysed in this section to find the solution to the multi-objective optimization problem under the uncertainty demands of the data owner, based on both privacy and computational cost. The privacy and computational cost are implemented using confidence intervals and three classifiers, namely, the multi-layer perceptron (MLP), naïve Bayes (NB) and classification and regression tree (CART), on a real-world data set. The real-world data set [*i.e.* the Adult Data Set from the University of California, Irvine (UCI) machine learning repository] is analysed in the experiments. A total of 16,281 instances is used for the experiment.

Table 1. Execution time.

Method	Execution time
IPDM-FO (proposed)	0.5 s
IPDM-NN	2.7 s
IPDM-SMC	3.9 s

Note: IPDM-FO = individual privacy in data mining and fuzzy optimization; IPDM-NN = individual privacy in data mining and neural network; IPDM-SMC = individual privacy in data mining and secure multi-party computation.

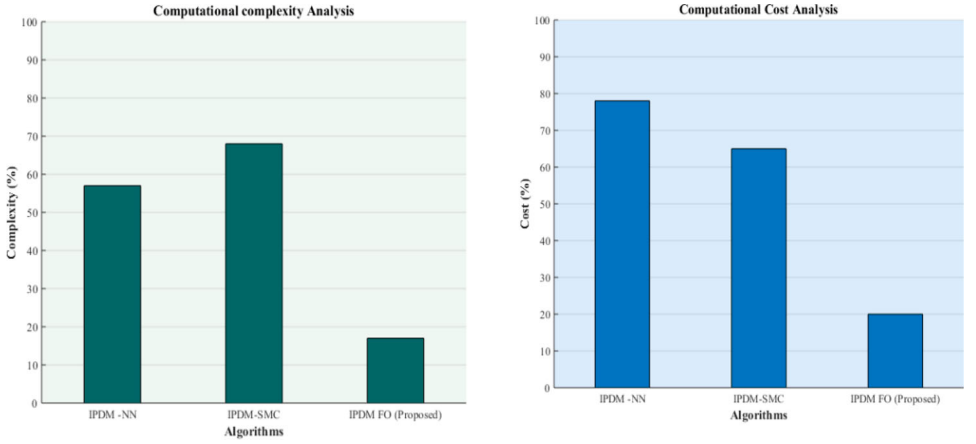

Figure 2. Computational complexity and computational cost analysis. IPDM-NN = individual privacy in data mining and neural network; IPDM-SMC = individual privacy in data mining and secure multi-party computation; IPDM-FO = individual privacy in data mining and fuzzy optimization.

6.1. Computational complexity

The computational complexity of the fuzzy optimization algorithm is calculated by

$$o(t(d * n + cof * n)) \tag{29}$$

where t is the number of iterations, d is the number of variables (dimension), n is the number of solutions, and cof is the cost of the objective function.

Table 1 shows the execution time of the proposed and existing methods. From the table, it can be seen that the execution time of the proposed method is low compared to the other two methods.

Figure 2(a) shows that the computational complexity of the proposed individual privacy in data mining and fuzzy optimization (IPDM-FO) algorithm is 68.96%, which is 73.91% lower than the existing algorithms, *i.e.* individual privacy in data mining and neural network (IPDM-NN) and existing individual privacy in data mining and secure multi-party computation (IPDM-SMC).

Figure 2(b) shows that the computational cost of the proposed IPDM-FO algorithm is 74.68%, which is 69.23% lower than the existing algorithms (IPDM-NN and IPDM-SMC).

6.2. Environment

The theoretical methods were tested on a personal computer with Intel™ core 2 Duo CPU, 2.92 GHz, 2.00 GB RAM, 32-bit Windows 7 OS with MATLAB® 7.0.1 development environment. The WEKA data-mining tool was used to run three classifiers (NB classifier, MLP and CART) on the data set. The experimental results are based on the amount of privacy measurement and computational cost.

Table 2. Computational cost of different types of perturbed data.

	Computational cost of perturbed data (using noise y)				
	Original data	Perturbed data using uniform distribution on (0,1)	Perturbed data using uniform distribution on (−1,1)	Difference between perturbed data based on distribution (0,1) and original data	Difference between perturbed data based on distribution (−1, 1) and original data
NB	0.13	0.08	0.06	−0.05	−0.07
MLP	34.58	38.59	36.67	4.01	2.09
CART	5.06	13.45	17.61	8.39	12.55
Computational cost of perturbed data (using noise $y/16$)					
NB	0.13	0.08	0.06	−0.05	−0.07
MLP	34.58	38.59	36.13	4.01	1.55
CART	5.06	14.17	14.48	9.11	9.42
Computational cost of perturbed data (using noise $16y$)					
NB	0.13	0.08	0.19	−0.05	0.6
MLP	34.58	37.11	36.99	2.53	0.91
CART	5.06	15.83	14.94	10.77	9.86
Computational cost of perturbed data (using noise $y/2$)					
NB	0.13	0.08	0.06	−0.05	−0.07
MLP	34.58	36.19	39.31	1.61	4.73
CART	5.06	13.64	15.11	8.58	10.05
Computational cost of perturbed data (using noise $2y$)					
NB	0.13	0.8	0.09	−0.05	−0.04
MLP	34.58	36.72	38.95	2.23	4.37
CART	5.06	14.14	16.31	9.08	11.25

Note: NB = naïve Bayes; MLP = multi-layer perceptron; CART = classification and regression tree.

6.3. Experimental results

The experiment was conducted based on both real and randomly generated data for both privacy and computational cost. The computational cost was determined by different classifiers on different perturbed data, as shown in Table 2. In Table 2, five different perturbed data were generated by $\{y/16, y/2, y, 2y, 16y\}$ using two uniform noise distributions, $[0,\ 1]$ and $[-1,\ 1]$. The detailed description of the generated perturbed data is as follows. Let the original data x lie between $[x - y]$ and $[x + y]$, where y is uniform noise distribution and the length of privacy is $2y$, *i.e.* twice the noise distribution. If x lies between $\left[x - \frac{y}{2}, x + \frac{y}{2}\right]$, then the length of privacy is y. If $x \in \left[x - \frac{y}{n}, x + \frac{y}{n}\right]$, then the length of privacy is $2y/n$. When $n \to \infty$, then $y/n \to 0$ and $\left[x - \frac{y}{n}, x + \frac{y}{n}\right]$ is very close to x. In this case, there is no requirement for adding or subtracting noise with original data. So, a certain limit of noise distribution with multiplication of the real value can be considered as privacy preservation; otherwise, the noise distribution is worthless. Hence, a decrease in the noise distribution value indicates a loss of privacy.

Multiplying any natural number by the noise distribution increases the privacy, with an increase in the interval length; for example, for $x \in [x - 2y, x + 2y]$, the length of privacy would be $4y$, and so on. Similarly, for $x \in [x - ny, x + ny]$ the length of privacy is $2ny$. When $n \to \infty$, then $2ny \to \infty$, and the privacy length is very large. Under this circumstance, it is difficult to find the original data.

In other words, data robustness is increased. According to the above discussion, the different perturbed data are used for the evaluation of computational cost, as shown in Table 2.

The experiments are carried out on the basis of the owner's demand. Here, the optimization problem is solved by the owner's required privacy level. The proposed optimal model is utilized with two components: privacy and cost. These components are derived as follows.

(a) Privacy (U_{HP}): This part defines how much privacy data are maintained by the data owner, as presented in Table 3. Similarly, C_{HP} is the measurement cost for preserved privacy determined

Table 3. Interval length and privacy.

	$y/16$	$y/8$	$y/4$	$y/2$	y	$2y$	$4y$	$8y$	$16y$
Interval length (−1,1)	0.000101125	0.00020225	0.000404	0.000808	0.001616	0.003232	0.006464	0.012928	0.025856
Privacy	2.50E−06	0.000005	0.00001	0.00002	0.00004	0.00008	0.00016	0.00032	0.00064
Interval length (0,1)	0.062506125	0.125012	0.250024	0.500048	1.000096	2.000192	4.000384	8.000768	16.00154
Privacy	0.001547475	0.003094	0.006188	0.012376	0.024752	0.049504	0.099008	0.198016	0.396032

Table 4. α-Cut value for noise data.

α-Cut value	0.2	0.4	0.6	0.8	1
Different noise data	$y/16$	$y/2$	y	$2y$	$16y$

within the interval length. Here, privacy is measured by (Interval length)/(Original data). Table 4 displays the results of $C_{\mathrm{HP}} * U_{\mathrm{HP}}$.

(b) Cost (U_{CC}): The U_{CC} is specified as computational cost on a priority basis, *i.e.* high priority for low computational cost. Uniform distributions between (−1, 1) and (0, 1) for different perturbed data are generated. The computational cost is measured by fuzzy sets. The priority of computational cost among classifiers based on perturbed data is recognized as a fuzzy set. The membership function is described as follows:

$$\mu_x = \begin{cases} 0.3 & \text{if HCC} \\ 0.6 & \text{if MCC} \\ 1.0 & \text{if LCC} \end{cases}$$

where HCC is high computational cost, MCC is medium computational cost, and LCC is low computational cost. Both original data and different types of perturbed data are considered for this experiment, and the utility factor for computational cost (U_{CC}) is considered on the basis of priority.

The evaluation of the optimization problem depends on fuzzy constraints related to privacy and cost. However, α-cuts produce different deviations of constraints compared to the original constraints according to the demands of the data owner. The α-cut values ($\alpha = 0.2, 0.4, 0.6, 0.8$ and 1.0) are considered for the optimization problem, corresponding to different levels of privacy. The value of α-cut varies according to the demand of data owner (Table 4).

The computational cost can also vary as privacy varies. Thus, to find an appropriate solution, the decision maker/data miner needs to make right decision to choose the amount of privacy in such a way that computational cost can be balanced. A high computational cost carries attention of less/low priority and *vice versa*; thus, the priority of cost is measured by 1/(computational cost), *i.e.* high priority = 1/(less/low cost) or less/low priority = 1/(high cost).

The experimental results for the optimality test are presented in Table 5. The different experiments were carried out based on α-cut values according to the demands of the data owner.

The fuzzy multi-objective optimization issue is solved based on privacy and cost, with uncertain demands of the data owner by means of fuzzy sets.

Here, the two types of perturbed data are considered based on uniform distributions between (0, 1) and between (−1, 1). The performance of the NB classifier is always better than the MLP and CART classifiers (Figure 3). The performance of each classifier on different weight factors for the optimality test at $\alpha = 1.0$ is shown in Figure 4. For all different weight factors, the performance of the NB classifier is better than the other classifiers in the optimality test. However, the performance of all classifiers varies as the weight factors vary.

For the Pareto-optimal test, the solution for the proposed model is best at $\alpha = 1.0$. As per the above discussion, based on the uniform distribution between (0, 1), the best optimal solution is obtained.

Table 5. Optimality test based on different α-cut values.

Weight factors (a_i^{HP}, a_i^{CC})	Uniform distribution between (0, 1)	Computational cost	Uniform distribution between (−1, 1)	Computational cost
Optimal test value for $\alpha = 0.2$				
(0.3, 0.7)	0.0055 (for MLP)	38.59	0.0058 (for MLP)	36.13
	0.0294 (for CART)	14.17	0.029 (for CART)	14.48
	8.75 (for NB)	0.08	11.6667 (for NB)	0.06
(0.7, 0.3)	0.0024 (for MLP)	38.59	0.0025 (for MLP)	36.13
	0.0128 (for CART)	14.17	0.0124 (for CART)	14.48
	3.7501 (for NB)	0.08	5 (for NB)	0.06
(0.5, 0.5)	0.0039 (for MLP)	38.59	0.0042 (for MLP)	36.13
	0.0212 (for CART)	14.17	0.0207 (for CART)	14.48
	6.25 (for NB)	0.08	8.3333 (for NB)	0.06
Optimal test value for $\alpha = 0.6$				
(0.3, 0.7)	0.0129	38.59	0.0057	36.67
	0.0386	13.45	0.0238	17.61
	8.7574	0.08	11.6667	0.06
(0.7, 0.3)	0.0197	38.59	0.0024	36.67
	0.0308	13.45	0.0102	17.61
	3.7674	0.08	5	0.06
(0.5, 0.5)	0.0163	38.59	0.004	36.67
	0.0347	13.45	0.017	17.61
	6.2624	0.08	8.3333	0.06
Optimal test value for $\alpha = 1.0$				
(0.3, 0.7)	1.9067	37.11	0.0057	36.99
	1.9276	15.83	0.028	14.94
	10.6511	0.08	3.6842	0.19
(0.7, 0.3)	4.4383	37.11	0.0025	36.99
	4.4472	15.83	0.012	14.94
	8.1859	0.08	1.5789	0.19
(0.5, 0.5)	3.1725	37.11	0.004	36.99
	3.1874	15.83	0.02	14.94
	9.4185	0.08	2.6316	0.19

Note: MLP = multi-layer perceptron; CART = classification and regression tree; NB = naïve Bayes.

Following the theory, the corresponding Pareto-optimal solutions including the Pareto-optimal set are evaluated based on the best optimal solution being obtained for a uniform distribution on (0, 1) at $\alpha = 1.0$. The Pareto-optimal solution (*i.e.* functional value of the utility function) is {1.9067, 1.9276, 10.6511, 4.4383, 4.4472, 8.1859, 3.1725, 3.1874, 9.4185} at $\alpha = 1.0$. The Pareto-optimal sets for privacy and computational cost are {0.396032} and {12.5, 0.0631712, 0.0269469}, which satisfy the optimization model.

6.4. Statistical analysis of the results

The performance of the proposed IPDM-FO method is analysed by comparing various metrics, *i.e.* the mean, variance and standard deviation, with the existing approaches (IPDM-NN and IPDM-SMC).

Figure 5 shows a comparison of the results attained by the different methods in terms of the mean values. Figure 6 compares the results attained by the different methods in terms of variance. Figure 7 displays the results attained by the different methods in terms of the standard deviation.

6.5. Discussion

The PPDM model is delineated according to the data phase, namely collection, publishing, distribution and output of data. The estimation of this model is addressed by analysing metrics to estimate

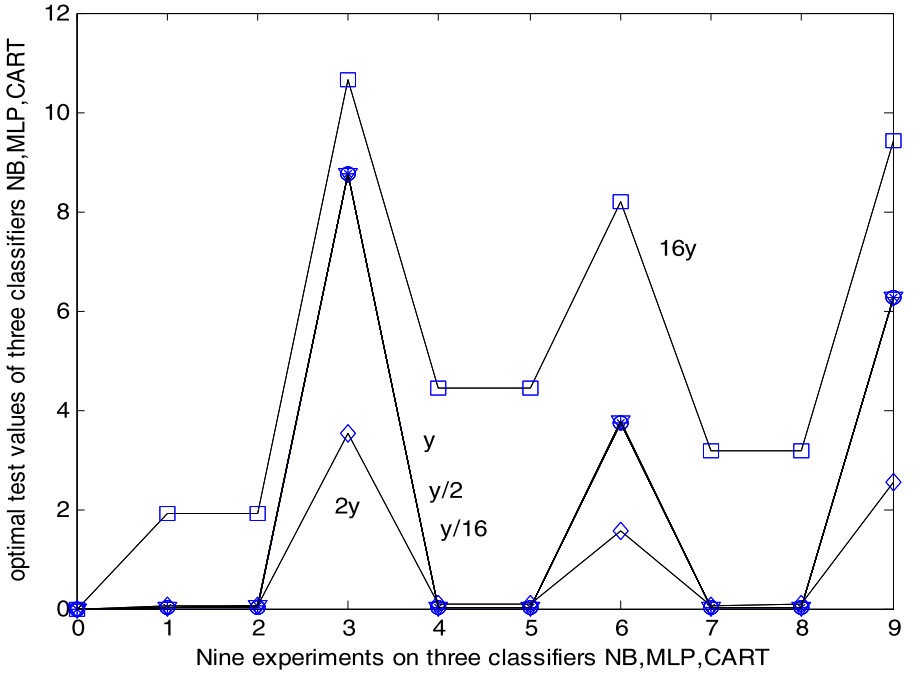

Figure 3. Optimal test for different classifiers. NB = naïve Bayes; MLP = multi-layer perceptron; CART = classification and regression tree.

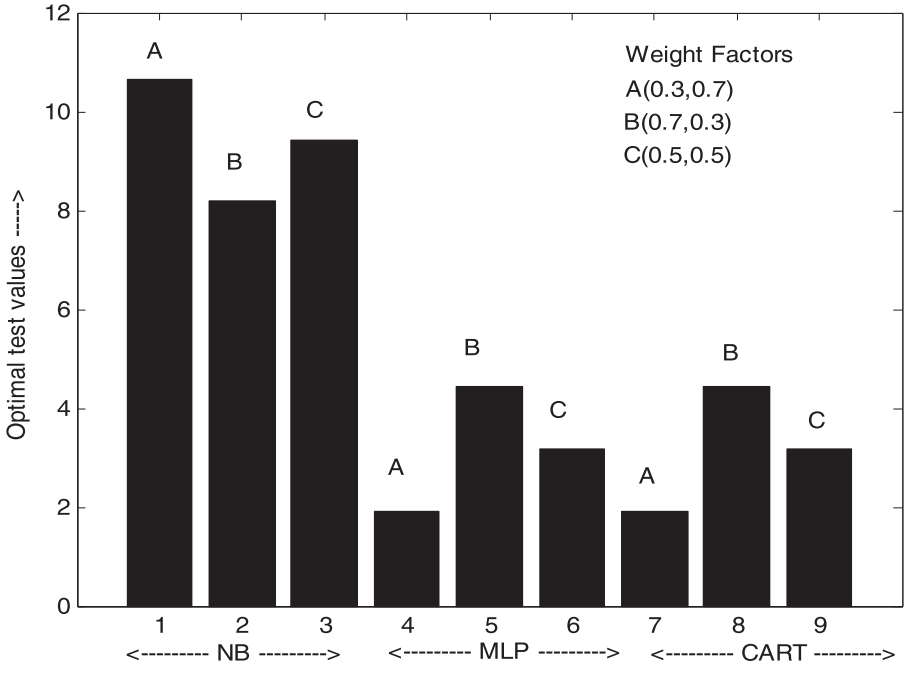

Figure 4. Optimal test value on weight factors. NB = naïve Bayes; MLP = multi-layer perceptron; CART = classification and regression tree.

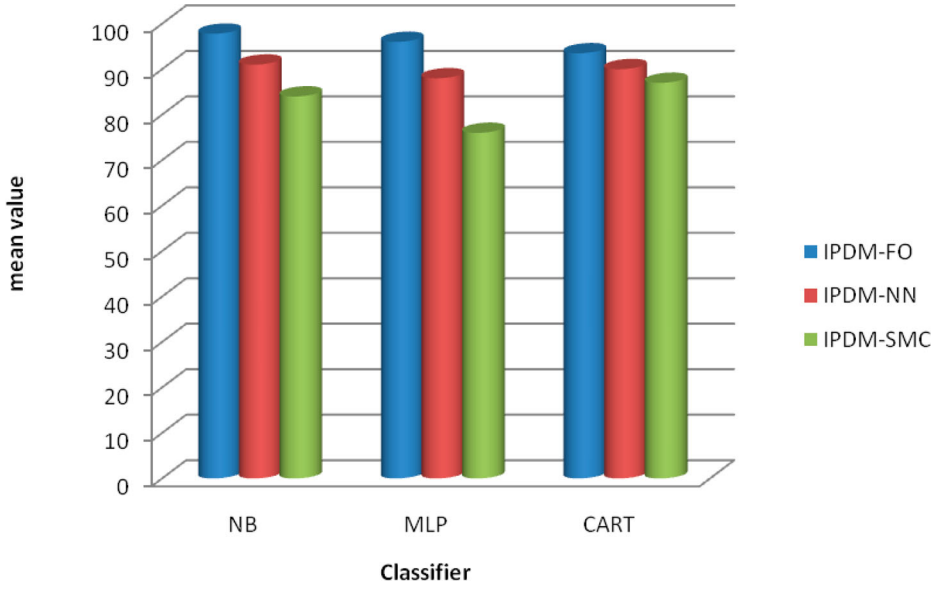

Figure 5. Comparison of the mean using the proposed and existing approaches. IPDM-FO = individual privacy in data mining and fuzzy optimization; IPDM-NN = individual privacy in data mining and neural network; IPDM-SMC = individual privacy in data mining and secure multi-party computation.

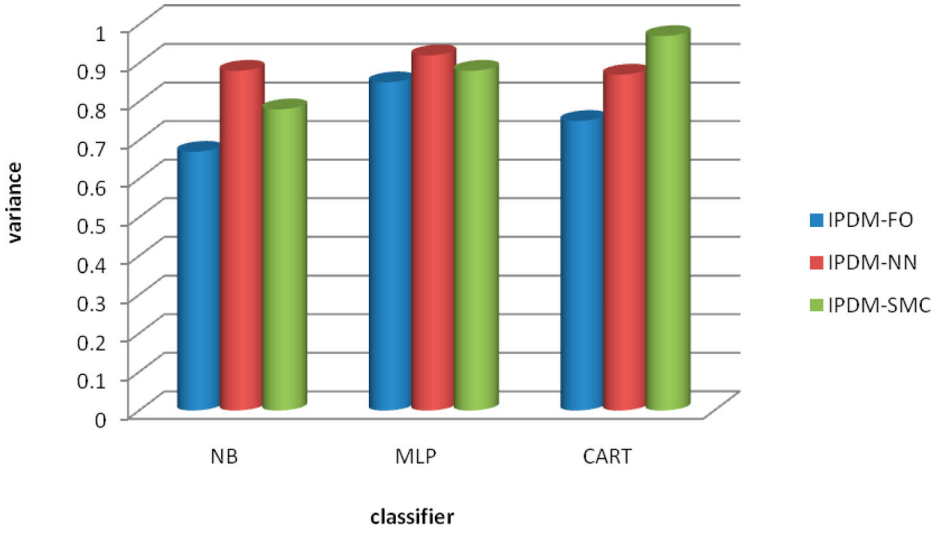

Figure 6. Comparison of variance using the proposed and existing approaches. IPDM-FO = individual privacy in data mining and fuzzy optimization; IPDM-NN = individual privacy in data mining and neural network; IPDM-SMC = individual privacy in data mining and secure multi-party computation.

the privacy and quality levels of data, and the complexity of the proposed PPDM model. This model can anonymize the data in such a way and hide a specific individual data to an unauthorized person. Hence, sensitive data are anonymized, so that background attacks can be eliminated while transforming the data. Even if a cyber-attack occurs in the system, storing anonymized data can prevent the retrieval of personal sensitive information. The PPDM model provides a safe way to store, exchange and publish information. In the case studies, the PPDM model is applied to assess several parameters, namely: (1) performance, determined depending on the time needed to attain the privacy

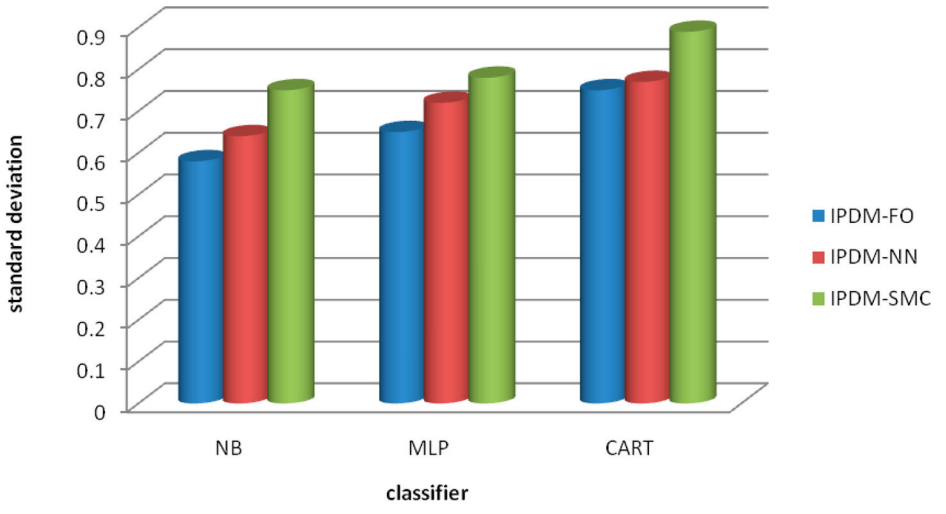

Figure 7. Comparison of standard deviation using the proposed and existing approaches. IPDM-FO = individual privacy in data mining and fuzzy optimization; IPDM-NN = individual privacy in data mining and neural network; IPDM-SMC = individual privacy in data mining and secure multi-party computation.

components; (2) data utility, which is fundamentally the calculation of information loss or loss of data processing in delivering outcomes, created by the lack of a PPDM model; (3) uncertainty level, *i.e.* the level of uncertainty with which the sensitive information that is hidden can still be predicted; and (4) resistance, *i.e.* the level of tolerance displayed by the PPDM model against various data-mining approaches. The aforementioned criterion requires to be scaled for optimum evaluation of the privacy preserving method. But, the two significant criteria are quantification of privacy and information loss. The quantification of privacy, or the privacy metric, is a measure that implies how closely the attribute's original value can be assessed. If it is assessed with greater confidence, the privacy is reduced and *vice versa*. Inaccuracy in estimating the original data set is called loss of information, which can lead to failure of the purpose of data mining. Therefore, a balance must to be achieved between privacy and information loss. Comparison of the case studies shows that the proposed method attains a good balance among disclosure, utility and costs.

7. Conclusion

In this article, a multi-objective fuzzy optimization model was proposed for individual privacy with fuzzy demands in data mining. As well as the optimality, the performance was verified using real-time data sets from the UCI machine learning repository. The results of this work can be used in real-time applications to include personal safety, security and privacy. These data can be vulnerable in terms of people's safety and a company's reputation. This method has some limitations, such as more time being needed to implement the data during the privacy method for the digital construction data and there being a possibility of loss of information during implementation. Thus, future work on this method will aim to reduce the execution time and retrieve the lost information.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Babaee Tirkolaee, E., I. Mahdavi, M. M. SeyyedEsfahani, and G. W. Weber. 2020. "A Hybrid Augmented Ant Colony Optimization for the Multi-trip Capacitated Arc Routing Problem Under Fuzzy Demands for Urban Solid Waste Management." *Waste Management & Research* 38 (2): 156–172.

Bakhtavar, E., T. Prabatha, H. Karunathilake, R. Sadiq, and K. Hewage. 2020. "Assessment of Renewable Energy-Based Strategies for Net-Zero Energy Communities: A Planning Model Using Multi-objective Goal Programming." *Journal of Cleaner Production* 272: 122886.

Bhuyan, H. K., and M. S. Huque. 2018. "Sub-feature Selection Based Classification." *2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI)*, 210–216. Tirunelveli, India: IEEE.

Bhuyan, H. K., N. K. Kamila, and S. K. Dash. 2011. "An Approach for Privacy Preservation of Distributed Data in Peer-to-Peer Network Using Multiparty Computation." *International Journal of Computer Science Issues (IJCSI)* 8 (4): 424.

Bhuyan, H. K., L. R. Kumar, and K. R. Reddy. 2019. "Optimization Model for Sub-feature Selection in Data Mining." *2019 International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 1212–1216. Tirunelveli, India: IEEE.

Bhuyan, H. K., M. Mohanty, and S. R. Das. 2012. "Privacy Preserving for Feature Selection in Data Mining Using Centralized Network." *International Journal of Computer Science Issues (IJCSI)* 9 (3): 434.

Bhuyan, H. K., and C. M. Reddy. 2018. "Sub-feature Selection for Novel Classification." *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, 477–482. Coimbatore, India: IEEE.

Cao, Y., M. Yoshikawa, Y. Xiao, and L. Xiong. 2019. "Errata on "Quantifying Differential Privacy in Continuous Data Release Under Temporal Correlations"." *IEEE Computer Architecture Letters* 31 (11): 2234–2234.

Chen, W., M. Panahi, and H. R. Pourghasemi. 2017. "Performance Evaluation of GIS-Based New Ensemble Data Mining Techniques of Adaptive Neuro-Fuzzy Inference System (ANFIS) with Genetic Algorithm (GA), Differential Evolution (DE), and Particle Swarm Optimization (PSO) for Landslide Spatial Modelling." *Catena* 157: 310–324.

Christen, P., T. Ranbaduge, D. Vatsalan, and R. Schnell. 2018. "Precise and Fast Cryptanalysis for Bloom Filter Based Privacy-Preserving Record Linkage." *IEEE Transactions on Knowledge and Data Engineering* 31 (11): 2164–2177.

Jahan, T., K. Pavani, G. Narsimha, and C. G. Rao. 2018. "A Data Perturbation Method to Preserve Privacy Using Fuzzy Rules." In Bhateja V., Tavares J., Rani B., Prasad V., Raju K. (eds) *Proceedings of the Second International Conference on Computational Intelligence and Informatics*, 9–16. Singapore: Springer.

Kamila, N. K., L. Jena, and H. K. Bhuyan. 2016. "Pareto-Based Multi-objective Optimization for Classification in Data Mining." *Cluster Computing* 19 (4): 1723–1745.

Kumar, S., and K. K. Mohbey. 2019. "A Review on Big Data Based Parallel and Distributed Approaches of Pattern Mining." *Journal of King Saud University-Computer and Information Sciences*. 1–24.

Lamata, M. T., D. Pelta, and J. L. Verdegay. 2018. "Optimisation Problems as Decision Problems: The Case of Fuzzy Optimisation Problems." *Information Sciences* 460-461: 377–388.

Langari, R. K., S. Sardar, S. A. A. Mousavi, and R. Radfar. 2020. "Combined Fuzzy Clustering and Firefly Algorithm for Privacy Preserving in Social Networks." *Expert Systems with Applications* 141: 112968.

Lekshmy, P. L., and M. A. Rahiman. 2020. "A Sanitization Approach for Privacy Preserving Data Mining on Social Distributed Environment." *Journal of Ambient Intelligence and Humanized Computing* 11 (7): 2761–2777.

Mendes, R., and J. P. Vilela. 2017. "Privacy-Preserving Data Mining: Methods, Metrics, and Applications." *IEEE Access* 5: 10562–10582.

Pellungrini, R., L. Pappalardo, F. Pratesi, and A. Monreale. 2017. "A Data Mining Approach to Assess Privacy Risk in Human Mobility Data." *ACM Transactions on Intelligent Systems and Technology (TIST)* 9 (3): 1–27.

Purohit, R., and D. Bhargava. 2017. "An Illustration to Secured Way of Data Mining Using Privacy Preserving Data Mining." *Journal of Statistics and Management Systems* 20 (4): 637–645.

Rajesh, P., and F. H. Shajin. 2020. "A Multi-objective Hybrid Algorithm for Planning Electrical Distribution System." *International Information and Engineering Technology Association*. 11: 66–82.

Shajin, F. H., and P. Rajesh. 2020. "Trusted Secure Geographic Routing Protocol: Outsider Attack Detection in Mobile Ad Hoc Networks by Adopting Trusted Secure Geographic Routing Protocol." *International Journal of Pervasive Computing and Communications*. https://doi.org/10.1108/IJPCC-09-2020-0136

Sharma, S., J. Powers, and K. Chen. 2018. "PrivateGraph: Privacy-Preserving Spectral Analysis of Encrypted Graphs in the Cloud." *IEEE Transactions on Knowledge and Data Engineering* 31 (5): 981–995.

Sin, G., Jianneng Cao Teo, and Vincent C. S. Lee. 2020. "DAG: A General Model for Privacy-Preserving Data Mining." *IEEE Transactions on Knowledge and Data Engineering* 32 (1): 40–53.

Stojiljković, M. M. 2017. "Bi-level Multi-objective Fuzzy Design Optimization of Energy Supply Systems Aided by Problem-Specific Heuristics." *Energy* 137: 1231–1251.

Transpire Online. 2020. "Horse Optimization Algorithm (HOA): Representative in Engineering Problem Application on Classification of the Smart Grid Stability." *Transpire Online* 2020. Accessed December, 2020. https://transpireonline.blog/tag/feature-selection-based-on-an-improved-cat-swarm-optimization-algorithm-for-big-data-classification/.

Zhao, L., L. Ni, S. Hu, Y. Chen, P. Zhou, F. Xiao, and L. Wu. 2018. "Inprivate Digging: Enabling Tree-Based Distributed Data Mining With Differential Privacy." In *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*, 2087–2095. Honolulu, HI, USA: IEEE.