Cluster Comput (2014) 17:1383–1399
DOI 10.1007/s10586-014-0393-9

Privacy preserving sub-feature selection based on fuzzy probabilities

Hemanta Kumar Bhuyan · Narendra Kumar Kamila

Received: 24 October 2013 / Revised: 10 April 2014 / Accepted: 14 July 2014 / Published online: 21 August 2014

Abstract The feature selection addresses the issue of developing accurate models for classification in data mining. The aggregated data collection from distributed environment for feature selection makes the problem of accessing the relevant inputs of individual data records. Preserving the privacy of individual data is often critical issue in distributed data mining. In this paper, it proposes the privacy preservation of individual data for both feature and sub-feature selection based on data mining techniques and fuzzy probabilities. For privacy purpose, each party maintains their privacy as the instruction of data miner with the help of fuzzy probabilities as alias values. The techniques have developed for own database of data miner in distributed network with fuzzy system and also evaluation of sub-feature value included for the processing of data mining task. The feature selection has been explained by existing data mining techniques i.e., gain ratio using fuzzy optimization. The estimation of gain ratio based on the relevant inputs for the feature selection has been evaluated within the expected upper and lower bound of fuzzy data set. It mainly focuses on sub-feature selection with privacy algorithm using fuzzy random variables among different parties in distributed environment. The sub-feature selection is uniquely identified for better class prediction. The algorithm provides the idea of selecting sub-feature using fuzzy probabilities with fuzzy frequency data from data miner's database. The experimental result shows performance of our findings based on real world data set.

H. K. Bhuyan (✉)
Department of Computer Science and Engineering,
Mahavir Institute of Engineering and Technology, Odisha, India
e-mail: hmb.bhuyan@gmail.com

N. K. Kamila
Department of Computer Science and Engineering,
C. V. Raman College of Engineering, Odisha, India

Keywords Distributed data mining · Fuzzy probabilities · Privacy · Feature selection

1 Introduction

Distributed data mining with privacy preservation has been developed by many prominent authors who have analyzed and made several strategies for data privacy. The role of feature selection has been made for several active research areas for industrial applications. Generally the feature selection techniques are used for clustering, classification etc. These techniques have been recognized as data mining techniques for relevant features. Many authors have proposed different algorithms and model for feature selection as [1–7] for their research work. But they have not considered the sub-features for their work. The sub-features play the vital role to predict the output (the target functions or class functions) correctly in many special tasks. The sub-feature doesn't have predicted capability can thereby be removed from consideration. Hence the existence of sub-feature and its role come to the limelight. Thus we have considered the sub-feature selection in our proposed work. The concept of sub-feature selection is being made by appropriate representation of fuzzy probabilities which are different from traditional techniques of feature selection. In order to maintain the privacy of sub-feature, some alias values have been taken into consideration. In this paper, the fuzzy probability has important role for all kind of processing tasks. Fuzzy random variables are used as fuzzy numbers which are vaguely defined as compare to real number and also associated for degree of acceptability. The true values are handled by each fuzzy random variable with membership function.

The information is shared by different parties under distributed network environment. The key challenge is to apply

fuzzy probabilities to multiparty collaborative distributed data mining to securely unify the perturbation used by different data providers, while each party still gets satisfactory privacy guarantee and the utility of the collected data is well preserved for making sub-feature selection model [8]. Generally there are some important factors that impact the quality of the sub-feature selection model such as frequency of sub-feature from each feature set, utility of fuzzy probability data for privacy, and data mining technique (i.e., gain ratio) etc. These factors are considered for developing the algorithms as well as model for sub-feature selection: fuzzy probability for sub-feature selection, estimation of upper bound and lower bound of gain ratio, and fuzzy privacy for sub-feature selection. The analytical and experimental results show that the fuzzy privacy algorithm for sub-feature selection is most efficient with effective result and privacy guarantee and also the estimation of gain ratio provide the feature selection within the expected interval.

This paper is organized as follows. In Sect. 2, it provides the background of the related work and preliminary study for proposed model. Section 3 defines the problem statement of the proposed work. In forth section, the fuzzy model for data processing has been illustrated. In Sect. 5, the estimation of both upper bounds and lower bounds of gain ratio with approximation solution based on fuzzy random variable are discussed whereas in sixth section, privacy preservation model for sub-feature selection is explained with algorithm. However in seventh section, the experimental details have been discussed and analyses with several dataset, parameters, proposed algorithms and privacy preservation for sub-feature selection for our proposed model. Section 8 ends with concluding remarks and open discussion for future work.

2 Background

In this section, the background is discussed with related works and some mathematically preliminaries describing the concepts for better understanding of the problem. Both parts derive the related concepts of privacy preservation in distributed data mining, feature selection under fuzzy environment.

2.1 Related work

The distributed data mining applications have been focused in several areas like large-scale distributed data mining, privacy preservation of data, peer to peer network systems etc., where each node makes exact solution for combined database [9]. The different distributed network algorithm like standard centralized algorithm for decision tree [10], sharing computation and information peer to peer network [11], centralized Bayesian networks [12], to discover criminal network [13], incentive compatible for distributed data mining [14] etc., have been derived on different database for several computational experiments. But in distributed environment, the role of participants is very important for computation and communication of individual data. Any party never wants to release their data without protection. Thus many researchers develop the different models and standard algorithms to protect the individual or organization data. The data privacy or privacy preservation of data has been developed by using standard and secure multiparty computation [15], game theory [16], K- anonymity and l-diversity approach in social network [17], privacy preserving data publishing [18], horizontal partitioned data [19] etc. To maintain high privacy by participating parties and coordinator of network, a model has been described for better computation [20].

Since feature has important role in distributed data mining, it needs approaches for feature selection. There are several approaches of feature selection like Decision Border feature [21], mutual information based on greedy selection [7], feature selection algorithm for large peer to peer networks [22], feature wrappers and filters [3] etc. Similarly different fuzzy techniques are used for feature selection such as fuzzy clustering technique [23], conventional search technique on fuzzy space [6], fuzzy support vector machine [4], fuzzy rough sets assisted attribute selection [5], fuzzy classification systems based on multi-objective evolutionary algorithm [24], construction fuzzy knowledge bases for feature selection [25], etc. Moreover additional methodologies such as higher order models for fuzzy random variable [26], upper and lower probabilities induced by fuzzy random variable [27], introduction of fuzzy rule based classifier [28], genetic fuzzy systems [29] and fuzzy linguistic models [30] are also used to strengthen different fuzzy techniques for feature selection. None of the authors of cited papers have discussed/proposed privacy preservation for fuzzy sub-feature selection. Our paper differs from [4,6,23] in many aspects. Firstly, we proposed, fuzzy sub-feature selection for better class prediction, secondly based on alias values, we have developed the algorithm to maintain the privacy. Thirdly, several solutions has been presented to evaluate the performance of selection. These results provide fundamental insights into the problem.

2.2 Preliminaries

In this section, it discusses basic concepts of fuzzy random variables, data mining technique (i.e., gain ratio) for feature selection and privacy preserving in distributed data mining for better understanding of the problem. The primary focus is on the related issues in the scenario of multiparty to release their perturbed data to data miner for

mining purpose. By convention, fuzzy random variables uses for privacy and also data evaluation for sub-feature selection.

2.2.1 Fuzzy random variables

As discussed by Huibert [31], the notion of a fuzzy random variable is discussed as follows. Assume (Ω,$\mathcal{F}$,$\mathcal{P}$) be a probability triple. Let U is a random variable defines on this triple. Assume that we perceive this random variable through a set of windows W_i i ϵ J, with J a finite or countable set, each representing an interval of the real line s.t. $W_i \cap W_j \neq \Phi$, for i $\neq$ j and $\bigcup_{i\epsilon j} W_i = R$ (perceiving the random variable through these windows means for each ω (omega), we can only establish $U_\omega \epsilon W_i$ for some i ϵ J). Let $\mathcal{F}_i$: R $\rightarrow$ [0, 1] be a character function defined on a set of windows W_i. Let S be the space of all piecewise continuous functions mapping R $\rightarrow$ [0,1]. Then define the perception of the random variable U as per above description and with mapping X: $\Omega \rightarrow$ S given by

$$\omega \xrightarrow{X} X_\omega \tag{1}$$

with $X_\omega = \mathcal{F}$ iff $U_\omega \epsilon W_i$(where W is perceiving data set) means it associates with each $\omega \epsilon \Omega$ is not a number, U_ω as an ordinary random variable, but a characteristic function X_ω, which is an element of S (where the mapping X : $\Omega \rightarrow$ S characterizes as a special type of fuzzy random variable). The random variable U is a fuzzy random variable which is a perception and is also called an *original* of the fuzzy random variable (FRV). Moreover for a given FRV, there may exist many originals. At this point a FRV is defined as a map ξ: $\Omega \rightarrow$ F, where F is the set of all fuzzy numbers (i.e., fuzzy random variables are random variables whose values are not real numbers but fuzzy numbers). Fuzzy numbers are numbers whose values are only vaguely defined. A fuzzy number may assume different real values, but it should be associated of degree of acceptability. The fuzzy random variable X is said to be discrete if Ω is a countable or finite set. When we deal with a single discrete fuzzy random variable X, we may take Ω = N, set of natural number and $\mathcal{F}$ the sigma algebra of subsets of N. We shall denote the probability P ({i}) = P_i, i ϵ N and $i \xrightarrow{X} X^i \ \forall$ i ϵ N. The fuzzy random variable is used to help feature selection for data mining task using gain ratio technique.

2.2.2 Evaluation of gain ratio for feature selection

Since feature selection issue is an important task in data mining [32,33], many authors have developed the different techniques for feature selection like entropy, gini index, gain ratio, mutual information etc. However we have considered the gain ratio technique in this paper for feature selection. Generally the gain ratio for feature selection is calculated as follows.

Let D is tuple set of partition related to class, P_i is probability of arbitrary tuple in D belongs to class C_i, x_i is the number of feature, D_j is the partition of feature data belongs to class C_i, info(D) is the information of D belong to class, $info_{xi}$ (D) is the information of x_i of D belong to class, $SplitInfo_{xi}$ is split information of x_i on D without class, only own feature data. Then the following calculation is required to find gain ratio for best feature with maximum value.

(1) Calculate Info(D) = $-\sum_{i=1}^{m} P_i \log_2(P_i)$
(2) Calculate $info_{xi}(D) = \sum_{j=1}^{n\ partition} \frac{|Dj|}{|D|}$ info(D_j)
(3) Calculate $gain_{xi}$ = Info(D) – $Info_{xi}$(D)
(4) $SplitInfo_{xi}(D) = \sum_{j=1}^{n\ partition} \frac{|Dj|}{|D|} \times \log_2 \frac{|Dj|}{|D|}$
(5) Gain $ratio_{xi} = \frac{Gain\ xi}{SplitInfo\ xi}$ where SpliInfo$_{xi} \neq 0$
(6) Find the best feature maximum gain ratio.

2.3 Privacy preservation on distributed data mining

Several perturbation techniques have been widely used for privacy preservation of individual data. Generally two types of perturbation techniques are used for privacy preservation of data, i.e., (1) multiplicative (2) additive perturbation. But we have considered additive perturbation for our work. The additive perturbation has been discussed by [15,34,35]. This approach makes data perturbation (Y) by adding some random noise data (Z) to the original data (X) as Y = X + Z where X, Y, Z are N-dimensional vectors,

where N is the number of attributes in X. The original data X follows the probability distribution. Hence mean vector and covariance matrix are μ_x, K_x. The noise Z is assumed to be independent of X and is a jointly Gaussian vector with zero mean and covariance matrix K_z. It is clearly to verify that the mean vector of Y is μ_x and its covariance matrix is $K_y = K_x + K_z$. It essential to choose K_z to be proportional to K_x i.e., $K_z = \sigma_z^2 K_x$ for some constant σ_z^2 denoting the perturbation magnitude [36].

3 Problem statement

In this section, we have considered a decentralized network and distributed data mining where the coordinator collects different data from each party indirectly and evaluates whole data for prediction of the classes. Although each party trusts on coordinator of the network system, still each party may maintain its privacy of individual data, but the coordinating data miner must maintain its privacy by adding some

amount of noise with data from each party i.e., the data miner collects only perturbed data for privacy purpose. By using additive perturbation to release the dataset, each party allows the coordinator to make statistical evaluation without releasing the exact values of individual data as discussed in Sect. 2.3. The assumption for system is that, the data miner always tries to construct an appropriate model on the basis of original data from given perturbed data. With respect to the original data X, the perturbed data Y represents how well the privacy is preserved for original data X.

Many traditional methods have been used for feature selection. Since feature data are used for different purpose, still it needs refinement of data for better classification. However there are some sensitive features (called sub-feature) of individual feature under feature set. They play major role leading to new class and their frequency may be less in feature data. We view all biological data values of feature as random variable. As the perception of biological data is always fuzzy, randomness is inevitable. Hence fuzzy random variable comes into the picture. Randomness occurs because it is not known which response may be expected from any given individual. Once response is available, there is still uncertainty about the precise meaning of the response. The latter uncertainty will be characterized by fuzziness. Thus the feature data are defined as fuzzy random variable and we have focused on privacy preservation for sub-feature selection. The following definitions have been considered for better understanding.

Definition 1 The feature (F_i) is said to be defined as sub-feature (S_j) if frequency of sub-feature value is more than zero i.e., $|S_j| > 0$. Each unique feature is recognized as sub-feature. A feature may have many sub-features.

Definition 2 A feature (F_i) is a set of sub-feature (S_j) that satisfying the following conditions

(1) $|F_i| = \sum_j |S_j|$
(2) $F_i = \bigcup_j S_j$
(3) $\bigcap S_j = \Phi$

where i is the i^{th} feature and j = number of sub-feature. Example: From UCI machine learning repository, following IRIS data set have been taken into consideration to explain the concepts being discussed in the definitions.

From the Table 1, features are sepal length, sepal width, petal length, petal width with the classes Iris-setosa, Iris-versicolor and Iris-virginica. In the feature sepal width, the sub-features are S_1(2.3,1), S_2 (2.7,1), S_3 (3.0,1), S_4 (3.1,1), S_5 (3.2,3), S_6 (3.3,2), S_7 (3.7,1) where arguments of the sub-features are feature values and its frequency in the data set. For feature sepal width F_{sw} = { S_1, S_2, S_3, S_4, S_5, S_6, S_7}. Hence $|F_{sw}| = \sum_{j=1}^{7} |S_j|$, $F_{sw} = \bigcup_j S_j$ and $\bigcap_j S_j = \Phi$.

Table 1 Consideration data of Iris

Sepal length	Sepal width	Petal length	Petal width	Class setosa/versicolor/ virginica
4.6	3.2	1.4	0.2	Iris-setosa
5.3	3.7	1.5	0.2	Iris-setosa
5.0	3.3	1.4	0.2	Iris-setosa
7.0	3.2	4.7	1.4	Iris-versicolor
6.4	3.2	4.5	1.5	Iris-versicolor
6.9	3.1	4.9	1.5	Iris-versicolor
5.5	2.3	4.0	1.3	Iris-versicolor
6.3	3.3	6.0	2.5	Iris-virginica
5.8	2.7	5.1	1.9	Iris-virginica
7.1	3.0	5.9	2.1	Iris-virginica

4 Fuzzy model for data processing

This section discusses both network design and data model task. Since the data are in distributed manner, it needs sharing among all parties for selection of best features and sub-features. Thus data sharing is important to solve local problems of individual party. Likely the privacy preservation of individual data is also essential.

Initially the model considers the collection of data from each party under decentralized manner. Since data are collected in different ranges of each feature from each party, it needs to make the global range of each feature. As data are in decentralized manner geographically, the data values of each feature will vary from place to place. For example, from UCI machine learning repository, it is found that women affected by breast cancer varies from rural to urban to metropolitan cities. Using alias techniques and fuzzy random variable, the model has been developed to maintain privacy preservation of data. The concept of fuzzy random variable has been discussed in Sect. 2.2.1. Fuzzy random variable (FRV) is defined as X = (T − a)/T, where 'a' is number of frequency of sub-feature data and T is the total dataset in the database. For example if a = 1 for a particular sub-feature and T = 150 datasets then the value of FRV X will be 0.993. Thus the coordinator collects the data value of sub-feature of which the value of fuzzy random variable is 0.993. In other words coordinator collects only sub-feature data as per the value of fuzzy random variables. The original data and its alias values of iris dataset have been presented in Tables 10 and 9 of Appendix 1. The data collection and processing of data at coordinator's end has been depicted in Figs. 1 and 2.

Based on Fig. 2, the coordinator makes easy to get original data from alias data and maintains its own database for processing is illustrated as bellows.

(a) Collection of the data as alias value

During flow of data from party to party, each sub-feature values is assigned as alias values for individual privacy. Each

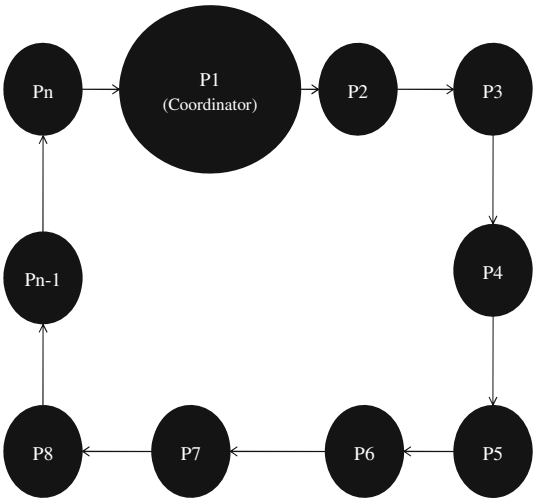

Fig. 1 Decentralized network with coordinator

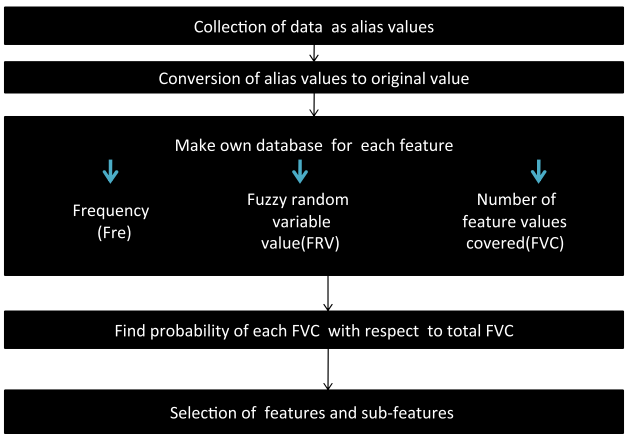

Fig. 2 Coordinator database for processing of task

party sends its own data in terms of alias values based on fuzzy random variable X to next party (as depicted in Fig. 1) under decentralized environment i.e., party P_1 to party P_2, …, party P_n. As alias data are moving from party to party, privacy is maintained. Finally alias data reach at coordinator's jurisdiction. In order to get alias data, each feature data are assigned by natural number. This natural number would be alias to original feature data. For example, suppose the feature data is 3.2, the alias data would be 1 (one). If data range is more, alias data range would be set accordingly which is shown in Table 9 presented in Appendix 1. This conversion process is within the knowledge of coordinator.

(b) Conversion of alias value to original data

Since the coordinator knows the conversion process, he can convert easily alias to actual data for further processing, if alias data reach at coordinator's control. Thus the coordinator can have new database. This has been presented in Table 10 of Appendix 1.

(c) Make own database of each feature

Now database is designed based on feature and sub-feature as shown in Table 8 of Appendix 1.

(d) Selection of feature and sub-feature

In this section, the data processing task is highly important for coordinator to select feature and sub-feature using fuzzy probability. We describe the general sub-feature selection algorithm presented in algorithm 1 using fuzzy random variable.

Algorithm-1 (Generalized sub-feature selection)

Input-: (i) Fuzzy random variable X , (ii) a = frequency number, (iii) T = Total data set, (iv) Z = collection of X based on their priority, (v) x = number of sub-feature value covered by frequency, (vi) Y = Total number of sub-feature value, (vii) Threshold value, (viii) d = fraction of criteria threshold value for fuzzy random variable.

Output-: Probability of sub-feature data as fractional criteria

FOR i = 1,2…..n // for number of frequency

Determine $X_i = \frac{T-a}{T}$

Z = Collect X_i with order.

Determine the number of sub-feature values covered by fuzzy random variable X_i.

END

FOR k=1,2…..m

Determine $\mu_k = \Pr(X_i = x) = \frac{x}{y}$

Determine D_j=

$$\frac{number\ of\ frequency\ covered\ by\ threshold\ value}{Total\ number\ of\ frequency}$$

$$R(d) = \begin{cases} X_i & if \quad D_j \leq d \\ 0 & otherwise \end{cases}$$

FOR d ϵ [0,1]

$\text{Prob}(X_i = x)(d) = \max_k R(d)\ \mu_k$

END

END

The algorithm 1 derives the general sub-feature selection from coordinator database. However the fuzzy privacy sub-feature selection algorithm is presented subsequently.

5 Fuzzy random variable for feature selection

In this section, the theoretical derivation of gain ratio is discussed using fuzzy random variable.

5.1 Gain ratio based on fuzzy random variable

Before discussing the gain ratio technique, let it considers the discussion of mutual information between discrete random variables which determine the statistical dependence between variables. The definition of mutual information between two random variables is described in [37]. For feature selection, the useful of mutual information is important to access the quality of discretization [38]. In contrast, the natural definition of mutual information between fuzzy variable and a crisp variable is a fuzzy number which is not numerical value. Since the mutual information is a part of gain ratio for feature selection, we can derive the gain ratio through mutual information using random variable and also fuzzy random variable.

The fuzzy random variable is a perception of which the random variable of the fuzzy random variable is called original and is regarded as family of random sets $(\psi_u)_{u\epsilon[0,1]}$ where each one of them associated to a confidence level 1- u. A random set is a mapping whose images are crisp sets. A random variable X is a selection of a random set Γ where the image of X is being the member in the image of the same outcome by Γ [39]. In other words if X be random variable and Γ be random set we can define as

$$X : \Omega \rightarrow R \tag{2}$$

$$\text{and} \quad \Gamma : \Omega \rightarrow \mathcal{P}(R) \tag{3}$$

where X is a selection of Γ (i.e., $X \epsilon A(\Gamma)$) and $X(e) \epsilon \Gamma(e)$ for all $e \epsilon \Omega$. Otherway Γ is also associative of random variables. A random set can be observed as a family of random variables. The gain ratio between a random variable X and random set Γ can be defined as the set of all values of X and Γ. Thus the gain ratio between random variable and random set is

$$GR(X, \Gamma) = \{GR(X, Z) | \, Z \epsilon A(\Gamma)\} \tag{4}$$

where X is a selection of Γ and $A(\Gamma)$ is a association of random variable.

The fuzzy random variable is being used as nested family of random sets as (Ψ_u) (where $u \epsilon (0,1)$) and each of them associated to certain confidence level. Thus we define the gain ratio between random variable X and fuzzy random variable Ψ as fuzzy set as defined by membership function

$$\widetilde{GR}(X, \Psi)(v) = \max\{u \mid v \epsilon GR(X, \Psi_u)\} \tag{5}$$

Similarly, assume that we are giving two paired standard random variable samples $X(X_1, X_2, \ldots X_N)$ and $Y(Y_1, Y_2, \ldots, Y_N)$ in which both universes of discourse are finite. Let $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots b_m$ are relative frequencies (probabilities) of the values of the samples of X, Y respectively and $c_1, c_2, \ldots c_s$ be the frequencies (probabilities) of the values of the joint samples $X \times Y$. Thus the gain ratio between variables X and Y is evaluated using following steps.

(1) Since a_i, b_i, and c_i are the relative frequencies (probabilities) of the arbitrary tuples in database D belonging to class C_i, the information needed to classify a tuple in D is

$$\text{Info}(X, Y)(D) = -\sum_{i=1}^{n} a_i \log a_i - \sum_{i=1}^{m} b_i \log b_i + \sum_{i=1}^{s} c_i \log c_i$$

(2) Let x_i be the number of feature and D_j is the partition of D for particular feature data belongs to class C_i then $info_{xi}(D)$ is the information of feature x_i of the database D for the above class. Thus

$$\text{Info}_{xi}(X, Y)(D) = \sum_{j=1}^{n\ partition} \frac{|D_j|}{|D|} \text{Info}(X, Y)(D)$$

(3) Hence the gain information from such a participating would be

$$\text{Gain}_{xi}(X, Y) = \text{Info}(X, Y)(D) - \text{Info}_{xi}(X, Y)(D_j).$$

Similarly split information can be derived analogously with info (D) as given [33]. Hence the gain ratio is defined as

$$\text{Gain ratio}_{xi}(X, Y) = \frac{\text{Gain}_{x_i}(X, Y)}{Splitinfo_{x_i}(X, Y)} \tag{6}$$

where $Splitinfo_{x_i}(X, Y) \neq 0$.

To estimate the gain ratio, let us consider two paired samples $X\{X_1, X_2, \ldots\ldots\ldots\ldots X_N\}$ and $\Psi\ \{\Psi_1, \Psi_2, \ldots\ldots\ldots\ldots \Psi_N\}$ of a crisp random variable X and fuzzy random variable Ψ. The estimation of gain ratio between X and Ψ can be derived by the fuzzy set as

$$\begin{aligned}\widehat{GR}((X_1, X_2, \ldots X_N), (\Psi_1, \Psi_2, \ldots \Psi_N))(v) \\ = \max\{u \mid v \epsilon \{GR((X_1, X_2, \ldots X_N), \\ (Z_1, Z_2, \ldots Z_N)) | \, (Z_1, Z_2, \ldots Z_N) \\ \epsilon A((\Psi_1, \Psi_2, \ldots \Psi_N))\, u\}\}\end{aligned} \tag{7}$$

Here the gain ratio determines the feature value based on its maximum membership.

5.2 Estimation of upper bounds and lower bounds of gain ratio

A fuzzy random variable is considered to find upper and lower bounds of gain ratio, otherwise probability distribution defined on class of random variables. Thus fuzzy random variable X is a mapping from Ω to R i.e.,

$$X : \Omega \rightarrow R$$

where Ω is feature space and R is fuzzy number.

The corresponding probability distribution P_F is defined on the class of random variables as

$$P_F(z) = \max\{u \mid z \in \Psi_u\} \tag{8}$$

where z is member of random sets Ψ_u.

By induction, it generates probability distribution on the values of the gain ratio as

$$P(GR(X, \Psi) = t) = \sum_{Z|GR(X,Z)=t} P_F(z) \tag{9}$$

Using the estimation of the bounds $P^u_F(z)$ and $P_{Fl}(z)$, we can estimate upper and lower bounds of P(GR(X,Ψ)). Finally we can also estimate the expected value of gain ratio with fuzzy optimization. Since the probability of sample of any fuzzy random variable Z is the product of all probabilities of Z_i, the model can be represented as

$$P_F(Z_1, Z_2, \ldots\ldots Z_m) = \prod_{i=1}^{m} P_F(Z_i) \tag{10}$$

Then the estimation of gain ratio is defined by above probability distribution as

$$P\left(GR\left(\left(\bigcup_{i=1}^{m} X_i, \bigcup_{i=1}^{m} \Psi_i\right) = t\right)\right) = \sum_{GR(\bigcup_{i=1}^{m} Xi, \bigcup_{i=1}^{m} Zi)=t} P_F(Z_1, Z_2, \ldots..Z_m) \tag{11}$$

The above probability provides a avenue to have a general formulation for fuzzy optimization with constraints and expected value as

$$\text{Max } E(GR) = \sum_{i=1}^{m} P_i * GR(X, Z_i) \tag{12}$$

$$\text{Subject to } \sum_{i=1}^{m} P_i = 1 \tag{13}$$

$$P_l \leq P_i \leq P^u \tag{14}$$

where P_i is the probability of each samples and (P_l, P^u) are the lower and upper bound probability. But this cannot find accurate solution practically. For the approximate solution, it can consider the above problem as

$$\text{Max } E(GR) = \frac{\sum_{i=1}^{m} P_i' * GR(X, Z_i)}{\sum_{i=1}^{m} P_i'} \tag{15}$$

$$\text{Subject to max } P_j'^l \leq P_i' \leq \max P_j'^u \tag{16}$$

where max P'^l select maximum value from all lower bound probability and max P'^u select maximum value from all upper bound probability. For approximation solution we consider two cases as follows.

(1) Case-1: Upper bound estimation

$$\text{Max } E^u(GR) = \frac{\sum_{i=1}^{m} q' * GR(X, Z_i)}{\sum_{i=1}^{m} q_i'} \tag{17}$$

$$\text{Subject to min}\{\max P_j'^l\} \leq q_i' \leq \max\{\max P_j'^u\} \tag{18}$$

(2) Case-2: Lower bound estimation

$$\text{Max } E_l(GR) = \frac{\sum_{i=1}^{m} q_i'' * GR(X, Z_i)}{\sum_{i=1}^{m} q_i''} \tag{19}$$

$$\text{Subject to min}\{\max P_j'^u\} \leq q_i'' \leq \max\{\max P_j'^l\} \tag{20}$$

Thus, from the above two cases, the approximate expected value with upper and lower bound estimation is

$$\text{Appx } E(GR) = \left[E_l(GR), E^u(GR)\right] \tag{21}$$

The several examples may be considered for feature selection within above expected interval. In next section, the sub-feature selection is derived based on fuzzy probability.

6 Privacy preservation model for fuzzy sub-feature selection

In this section, three criteria are considered for sub-feature selection such as less frequent (LF), medium frequent (MF) and very large frequent (VLF) feature values from the database. The criteria are assumed characterized as fuzzy numbers with membership function as sketched in Fig. 3. Thus using the above information fuzzy random variable is a mapping from feature elements to level of criteria i.e.,

$$X : \Omega \rightarrow L \tag{22}$$

where each $\omega \in \Omega$ represents sub-feature elements and X (ω) represents the label (L) is defined for criteria (i.e., LF, MF, VLF). Here the values of fuzzy random variables are fuzzy numbers vaguely. A fuzzy number may assume different real values with a degree of acceptability. This degree of acceptability can be handled accordingly to rules of fuzzy logic. The fractions of feature values as per the criteria are given in Table 2.

Since Ω is countable, the fuzzy random variable X is said to be discrete. As we are dealing with a single discrete fuzzy random variable X, we can take Ω = N, where N is set of natural numbers. Thus

Table 2 Fractional criteria

Criteria (frequency of feature values)	Fraction of criteria values
Less frequent	0.2
Medium frequent	0.5
Very large frequent	0.3

$$X: N \rightarrow P_n \text{ and } X: N \rightarrow Z_N \tag{23}$$

where P_n is probability and Z_N is the membership functions corresponds to criteria LF, MF, VLF. i.e., a single discrete fuzzy random variable X is essentially characterized by set of pairs (P_n, Z_N). From Table 2, the probabilities are P_1=0.2, P_2=0.5, P_3 = 0.3, P_n=0 for n > 3 (since ΣP_n=1), and membership functions are LF, MF, VLF as depicted in Fig. 3.

To continue this discussion, let us consider S, the set of piecewise continuous functions with mapping $R \rightarrow [0, 1]$. We then define X: $\Omega \rightarrow S$ as a special type of fuzzy random variable as per the perception described above which in term called as selected fuzzy random variable. That means there may exist many random variables for a given selected fuzzy random variable. Under this condition we generalize and define the discrete fuzzy random variable X as a mapping from $\Omega \rightarrow F_N$ where F_N is the set of fuzzy numbers. If $\omega \in \Omega$, we define image as ω in F_N as X_ω which satisfy the following conditions.

For each membership function $\mu \in [0,1]$, the two selected fuzzy random variables P_μ and Q_μ on the sub-feature values defined by

$$P_\mu(\omega) = \inf\{i \in R \mid X_\omega(i) \geq \mu\} \tag{24}$$

$$Q_\mu(\omega) = \sup\{i \in R \mid X_\omega(i) \geq \mu\} \tag{25}$$

are finite real valued random variables satisfying for all $\omega \in \Omega$

$$X_\omega(P_\mu(\omega))) \geq \mu,\ X_\omega(Q_\mu(\omega))) \geq \mu \tag{26}$$

The above conditions are finite support for selecting the sub-features because finite support is important to choose large enough for all purposes. In addition to this condition each random variable should be normal that means for each i $\in$ R, $X_\omega(i) = 1$. To find the selecting sub-feature, it considers the level sets corresponding to given fuzzy sets. An algorithm is being considered for determining the fuzzy expectation of a fuzzy random variable with membership function. Now the family of level sets F with $\mu \in [0,1]$ is as follows

$$F_\mu = \{x \in G \mid g(x) \geq \mu\} \tag{27}$$

where G is basic space with g: $G \rightarrow [0, 1]$. The membership function g is defined again as

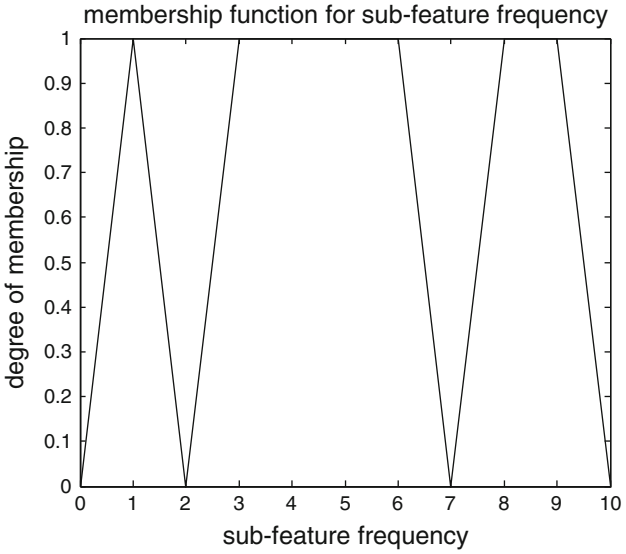

Fig. 3 Membership functions of less frequency (0,2), medium frequency(2–7), very large frequency (7–10)

$$g(x) = \sup\left\{\mu \in [0, 1] \mid x \in F_\mu\right\},\ x \in G \tag{28}$$

Thus it can define a fuzzy set (G, g) on the basic space G. Now the algorithms which are being presented only allow the evaluation of the level sets F_μ at the discrete values of $\mu \in [0, 1]$ which in turn allow an approximate evaluation of the membership function g with the help of above equation. The level sets of expectation E(X) of a discrete fuzzy random variable X are considered here for determination of sub-feature from the feature space.

As we have considered decentralized network model, each party provides their converted fuzzy data to the coordinator of which individual privacy is maintained. In other-words the coordinator collects the individual alias data with adding randomized data for privacy is discussed in algorithm 2. Then subsequently the coordinator evaluates whole data and selects the sub-feature within expected range (given in algorithm 3). The alias data as fuzzy frequency from each party are transferred from party to party in decentralized system for which each party can't able to know other party's data due to the reason that privacy is maintained at each party level.

The above two algorithms helps to select the sub-feature value for class with privacy among the participating parties in decentralized network.

7 Experimental details

This section discusses the application of proposed algorithms, data sets and different parameters being considered for implementation along with performance evaluation.

Algorithm-2: (fuzzy privacy sub-feature selection)

Input: (i) Ring network size = D, (ii) Number of features x_i, (iii) Range of each feature value, (iv) coordinator who initialize all task, (v) Total number of data = T, (vi) Number of sub-feature in a particular feature = Z, (vii) a = number of frequency, (viii) fuzzy frequency value Y.

Output: Select sub-feature from feature space.

Initialize R = random (Z) // random number of Z
FOR node, K = 1,2,….D // number of nodes

(participating parties)

FOR feature set i = 1,2,….n // for each feature

Initialize feature x_i
FOR sub-feature j = 1,2,……m

Initialize frequency of each sub-feature value x_{ij}

at coordinator as $Y_p = \frac{T-a}{T}$ and collect $Y_p = \bigcup x_{ij}$
Each x_{ij} assign with natural number Z_n

Send the frequency of Y_p with Z_n to next node with adding R.

Continue until last node and computational result send to coordinator.

Coordinator get each x_{ij} with frequency by subtracting R from last result and conversion of Z_n to x_{ij}

END

Call fuzzy randomized evaluation.

END

Send selected sub-feature to all nodes.

END

Algorithm-3: (fuzzy randomized evaluation)

Choose several membership functional value μ.
FOR ω ϵ Ω
For any sub-feature frequency number 'a' and total data set T.

Determine $X_\omega(a) = \frac{T-a}{T}$

END

FOR fixed μ value

IF $(X_\omega(a) \geq \mu)$ then

$P_\mu(\omega) = \inf\{X_\omega(a) \geq \mu\}$

$Q_\mu(\omega) = \sup\{X_\omega(a) \geq \mu\}$

Probabilities $Pi = \frac{Total\ sub-feature\ of\ each\ frequency}{Total\ data\ set}$

Determine $A = \sum_i P_i P_\mu(\omega)$
Determine $B = \sum_i P_i Q_\mu(\omega)$
Expected range $E_R = [A, B]$

END

For any sub-feature x value

IF $(x \in E_R)$

count the number of sub-features from feature space.

END
END
END

7.1 Description of datasets

Even though the proposed algorithm is primarily intended the privacy preserving sub-feature selection, it can also be used very well on conventional data set. In order to show this fact, we have evaluated of algorithm using IRIS plant data set [40] from University of California Irvine (UCI) machine learning repository. The data set with 150 data contains three different class flowers (Setosa, versicolor, virginica) for class. Each class consists of 50 data sets with four features out of which two features (petal, sepal) are important. Generally IRIS data set is the most popular and simple classification data set based on multi-variate characteristics of a plant species (length and thickness of its petal and sepal) divided into three distinct classes of 50 instances each. One class is linearly separable from each other. The four features are predicting features and one is goal feature. All predicting features are real values. The length and width of each feature are important to select sub-features with the help of feature values from four dimensional measurement spaces.

7.2 Environments and parameters

7.2.1 Environments

The proposed method is implemented on a personal computer with an Intel Pentium IV, 2.40 GHZ CPU, 1.00 GB RAM,

Table 3 Parameters used in proposed algorithm

S. n.	Symbol	Name and purpose of the parameters
1	μ	Membership function to make threshold value for sub-feature selection
2	T	Total number of data set
3	a	Frequency number
4	P_μ	Selected fuzzy random variable with minimum
5	Q_μ	Selected fuzzy random variable with maximum
6	ER	Expected range
7	P_i	Probabilities of sub-feature of each frequency with respect to total data set.

Microsoft Windows XP professional version 2002 operating system with Matlab 7.0.1 development environment. The data set have been processed under fuzzy environment for sub-feature selection.

7.2.2 Parameters

The predicted data of each feature from feature space is measured by fuzzy random variables with membership function. For evaluating the proposed algorithm, the interpretation of user defined parameters is illustrated in Table 3. Although the parameters are quite restricted but there is no such standard rule to assign systematic parameter values.

The brief description of parameter values are as follows. The frequency parameter 'a' and total dataset 'T' are used to measure the values of fuzzy random variable for each sub-feature as $X_i = \frac{T-a}{T}$. It is observed from computation (Table 8 from Appendix 1) that sixteen numbers of frequencies in IRIS data set are within the fuzzy frequency interval [0.993, 0.806]. The probability P_i of each fraction of criteria value for all features is determined by $(X_i * \mu_k)$ which is explained in algorithm-1. For example, the probability of feature "sepal length" having frequency one is measured by $0.993 * \frac{9}{35} = 0.993 * 0.257 = 0.255$ and $0.933 * \frac{1}{35} = 0.026$. Thus it is concluded that 10 numbers of fuzzy frequencies for feature "sepal length" are in the interval [0.255, 0.026]. Similarly 14 numbers of fuzzy frequencies for feature "sepal width" lie within the interval [0.215, 0.035]. Nine numbers of fuzzy frequencies for "petal length" are within [0.230, 0.041] and eleven numbers of fuzzy frequencies for "petal width" within [0.089, 0.036]. The detail description about experimental data using algorithm 1 is presented in Table 5.

Again two fuzzy random variables P_μ and Q_μ are important to determine the best sub-feature selection with certain expected range ER. The probability P_i determines the total sub-feature of each frequency with respect to total data set. For example P_i for frequency one of "sepal length" is $\frac{(1*9)}{150} =$ 0.06 and for frequency two of sepal length is $\frac{(2*2)}{150} = 0.026$. Since it considers only less frequent feature (i.e., only frequency one and two) then $P_\mu = 0.986$ and $Q_\mu = 0.993$. The sub-feature selection is restricted with expected range as $[\sum_i P_i P_\mu, \sum_i P_i Q_\mu]$. Thus the expected range of sub-feature selection of feature sepal length is [0.026 * 0.986, 0.06 * 0.993] = [0.0256, 0.0595]. The sub-feature selection within expected range for all features is shown in Table 4.

7.3 Results and analysis of proposed algorithms for sub-feature selection

The first part of experiment is analyzed based on coordinator database using algorithm-1. Fuzzy frequency variable is used to collect the sub-feature values from different parties which vary from one feature to another. The order of sub-feature values is arranged in the order of fuzzy frequency. The probability of sub-feature set covered by fuzzy frequency is determined by the number of corresponding sub-feature with respect to total number of sub-feature. The number of frequency covered by threshold value is used for fraction of criteria, otherwise it considered as zero which will never predict this sub-feature. The probability of each fraction of criteria is exhibited in Table 5. The probability is different for each frequency corresponding to available sub-features. The first two probabilities are considered for sub-feature selection due to fact that the sub-features have less frequency. If we consider the sub-features having frequency more than two, the sub-feature values would be available in more than one class for which the selection is a challenging task. The value of fuzzy random variable of individual frequency of IRIS database is shown in Table 5. The increase values of fuzzy random variable with decrease values of frequencies are shown in Fig. 6 of Appendix 2.

The maximum and minimum probabilities of sub-feature data of each feature are shown in Table 6. As we have considered, the sub-feature having frequency one and two for sub-feature selection, this range (max–min) is not helpful. The reason behind it is: sub-feature having less frequency can lead to a unique class. Hence Table 7 is used to solve our purpose.

In the second part of the experiment, the applications of fuzzy random variable for sub-feature selection are elaborated using Iris data set. We have considered, six parties are participated in a peer to peer network. Each party holds 25 data sets and four feature sets for experiments. Each party provides their own feature data to data miner using two ways of maintaining privacy (i.e., alias data and secure multiparty computation). There is no exact available sub-feature data range comparing to total sub-feature data range at each party, because each party maintains only their own local data. For example, first party holds the sub-feature data range (4.3 – 5.8) whereas second party holds range (4.4 – 5.5) and so on.

Table 4 Expected range of sub-feature selection in each feature of IRIS dataset

Purpose	SL	SW	PL	PW
Evaluation	[0.026*0.986, 0.06*0.993]	[0.0133*0.986,0.0333*0.993]	[0.0666*0.993, 0.1466*0.986]	[0.0133*0.986, 0.0133*0.993]
Result	[0.0256, 0.0595]	[0.0131, 0.0331]	[0.0662, 0.1446]	[0.0131, 0.0132]

Table 5 Probability of each fraction of criteria for all features

S. n.	FRVV	SL	SW	PL	PW
1	0.993	0.255	0.215	0.230	0.089
2	0.986	0.056	0.042	0.251	0.044
3	0.98	0.111	0.127	0.158	0.177
4	0.973	0.138	0.083	0.112	0.043
5	0.966	0.055	0.041	0.066	0.131
6	0.96	0.136	0.124	0.022	0.086
7	0.953	0.081		0.043	0.129
8	0.946	0.054	0.040	0.043	0.085
9	0.94	0.053	0.040		
10	0.933	0.026	0.040		
11	0.926		0.039		
12	0.92		0.039		0.082
13	0.913		0.039	0.041	0.041
14	0.906		0.038		
15	0.826		0.035		
16	0.806				0.036

Table 6 Max and min probability value of sub-feature value

S. n.	SL	SW	PL	PW
Max	0.255	0.215	0.251	0.177
Min	0.026	0.035	0.021	0.036

Table 7 Best sub-feature based on values of fuzzy random variable

S. n.	FRVV	SL	SW	PL	PW
1	0.993	0.255	0.215	0.230	0.089
2	0.986	0.056	0.042	0.251	0.044

But the global range for feature "sepal length" is (4.3 – 7.9) which are not exactly equal as compared to each party's data range. After the collection of data from each party, the data miner is observed that total data set is 150 and the number of sub-feature data for each feature (sepal length, sepal width, petal length and petal width) are {35, 23, 43, 22} respectively. These sub-feature data sets are used for selection of sub-feature predicting to new class.

As per the probability of fractional criteria shown in Table 5, the sub-feature data from each feature corresponding to their probability is depicted in Fig. 4. In this figure, it is observed that the probability of available sub-feature of each feature having same number of frequency is different due to the reasons that the frequencies of each sub-feature from each feature are different. It is further observed that as frequency increases, the corresponding probability decreases. It shows that the sub-features having less probability are involved in many classes; don't lead to a particular class whereas sub-features having highest probabilities lead to new class. As we have considered only two selected fractional criteria, the probabilities of sub-features of each feature based on values of fuzzy random variables (i.e., 0.993, 0.986) are exhibited in Fig. 4e relating to features SL, SW, PL and PW. It is interesting to note that if we choose any sub-features in between graphs having the values of fuzzy random variable 0.993 and 0.986 for all features; it will lead to new class. Thus the selected sub-feature for all features are depicted in Fig. 5 for frequency one and two being considered as best of sub-features.

7.4 Results and analysis for privacy preservation

The coordinator collects sub-feature or feature data from each party using two ways of maintaining privacy (i.e., alias data technique and secure multiparty computation technique). As per fuzzy frequency, each party provides their available data as alias data to coordinator. The fuzzy frequencies are computed using the fuzzy technique $\frac{T-a}{T}$ where T for total data set and 'a' is the number of frequency. During the collection of sub-feature data from different parties, the data size is changed in the current party after collecting sub-feature data due to the reason that the same sub-feature data are not available in all parties. For example, the coordinator (as first party) sends the sub-feature data set {4.3,4.4,4.7,5.8} with fuzzy frequency value 0.993 and { 4.9, 5.0, 5.7} with fuzzy frequency value 0.986 to second party. When it reaches at second party, the sub-feature data set would be { 4.3, 4.4, 4.5, 5.3, 5.8} and { 4.7, 5.5} with fuzzy frequency 0.993 and 0.986 respectively with own sub-feature data which will send to third party.

As it is a secure multiparty computation problem, the coordinator as first party collects the data range for all features and makes global range accordingly prior to reaching original data. Subsequently the coordinator makes alias range of data range by assigning natural number starting from 1. Now the first party sends sub-feature data in the form {1, 2, 5,

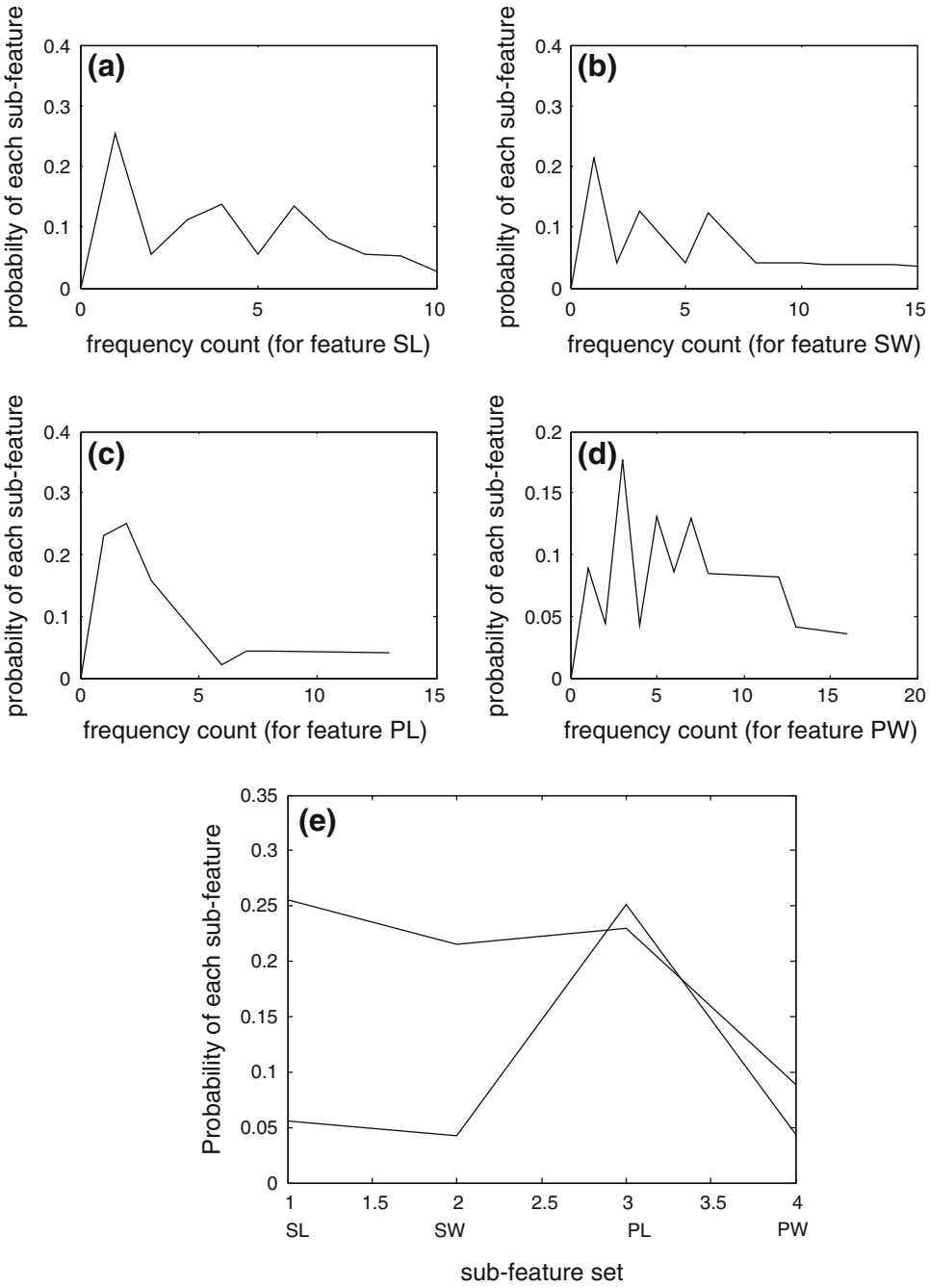

Fig. 4 Feature values versus corresponding values of fuzzy random variable **(a–d)** and selected fractional criteria with corresponding probabilities **(e)**

16} and {7, 8, 15} with 0.993 and 0.986 as fuzzy frequency value respectively to second party. Similarly second party sends his alias data {1,2,3,11,16} and {5,13} with fuzzy frequency value 0.993 and 0.986 to third party without knowing the original data of first party and so on until it reaches at first party. Finally all alias data reach at first party under decentralized environment. Hence it is observed, fuzzy frequency value and alias data both help to preserve the privacy of each participating party's data during computation. The whole process is being implemented using algorithm 2 and algorithm 3. Here the privacy is maintained two times. Since each party maintains the privacy of individual data set, it is necessary to measure "how much privacy is preserved" for own data set which is beyond the scope of this paper.

8 Conclusion

This paper explores the use of fuzzy probability and perturbation technique to select sub-feature maintaining privacy

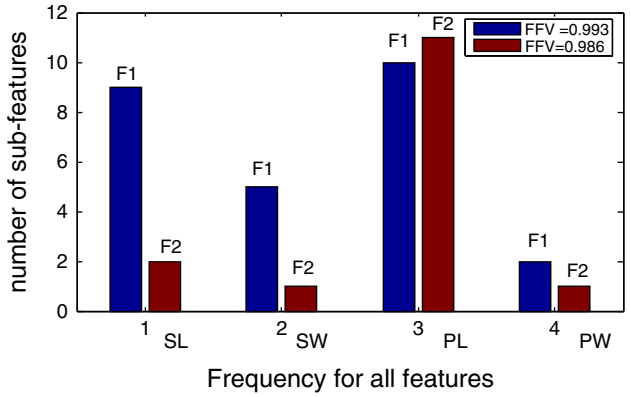

Fig. 5 Selected number of sub-features having frequency one and two for all features (F1-frequency one and F2- frequency two)

preservation in distributed data mining environment. From results and analysis, it is concluded that when frequency count is more, the probability of each sub-feature is less. It proves that for more frequency count involvement of sub-feature in whole classes is true for which it doesn't lead a new class whereas less frequency count can help sub-feature to a unique class. Moreover, the approach of fuzzy random variable confined the expected range on which the selection of sub-feature from feature database is made easy. At the same time privacy of original data are still well maintained without divulging the exact data values of each party during secure multiparty computation because of perturbation and fuzzification. Under this distributed data mining environment, data values of individual party become secure doubly. The experimental results demonstrate that the notion of fuzzy random variable for sub-feature selection can be successfully applied to different kinds of data mining task including clustering, decision making, gain ratio etc. This technique offers another interesting direction to extend the unique association rule to predict a new class. Even though the IRIS data set is used for this experiment, but it is scalable to consider medical data set which are more sensitive than IRIS data.

Appendix 1

See Tables 8, 9, 10

Table 8 Database for each feature

S. n.	NF	VFRV	Number of sub-feature value covered				Probability of each covered frequency w.r.t. total frequency			
			SL	SW	PL	PW	SL	SW	PL	PW
1	1	0.993	9	5	10	2	0.257	0.217	0.232	0.090
2	2	0.986	2	1	11	1	0.057	0.043	0.255	0.045
3	3	0.98	4	3	7	4	0.114	0.130	0.162	0.181
4	4	0.973	5	2	5	1	0.142	0.086	0.116	0.045
5	5	0.966	2	1	3	3	0.057	0.043	0.069	0.136
6	6	0.96	5	3	1	2	0.142	0.130	0.023	0.090
7	7	0.953	3		2	3	0.085		0.046	0.136
8	8	0.946	2	1	2	2	0.057	0.043	0.046	0.090
9	9	0.94	2	1			0.057	0.043		
10	10	0.933	1	1			0.028	0.043		
11	11	0.926		1				0.043		
12	12	0.92		1		2		0.043		0.090
13	13	0.913		1	2	1		0.043	0.046	0.045
14	14	0.906		1				0.043		
15	26	0.826		1				0.043		
16	29	0.806				1				0.045

Where *NF* number of frequency, *VFRV* value of fuzzy random variable

Table 9 Coordinator collects alias data as natural numbers

S. n.	SL		SW		PL		PW	
	Fuzzy frequency value	Alias data	Fuzzy frequency value	Alias data	Fuzzy frequency value	Alias data	Fuzzy frequency value	Alias data
1	0.993	{1,3,11,28,29, 31,32,33,35}	0.993	{1,20,21, 22,23}	0.993	{1,2,11,17, 18,19,43, 44,45,48}	0.993	{5,6}
2	0.986	{5,24}	0.986	{19}	0.986	{3,10,14, 16,24,33,34, 35,40,41,46}	0.986	{14}
3	0.98	{2,17,26,30}	0.98	{2,4,17}	0.98	{20,22,27,36, 38,39,42}	0.98	{8,19, 21,22}
4	0.973	{4,10,20, 27,34}	0.973	{3,16}	0.973	{8,23,25, 29,31}	0.973	{13}
5	0.966	{6,23}	0.966	{6}	0.966	{21,28,30}	0.966	{1,9,16}
6	0.96	{7,12,14, 18,19}	0.96	{13,15,18}	0.96	{37}	0.96	{17,18}
7	0.953	{13,16,22}			0.953	{4,7}	0.953	{3,4,7}
8	0.946	{15,25}	0.946	{5}	0.946	{32, 26}	0.946	{11,20}
9	0.94	{9,21}	0.94	{7}				
10	0.933	{8}	0.933	{9}				
11			0.926	{11}				
12			0.92	{14}			0.92	{12,15}
13			0.913	{12}	0.913	{5,6}	0.913	{10}
14			0.906	{8}				
15			0.826	{10}				
16							0.806	{2}

NF number of frequency, *OD* original data

Table 10 Conversion of alias value to original value

S. n.	SL		SW		PL		PW	
	NF	OD	NF	OD	NF	OD	NF	OD
1	1	{4.3,4.5,5.3,7.0, 7.1,7.3,7.4,7.6,7.9}	1	{2.0,4.0,4.1, 4.2,4.4}	1	{1.0,1.1,3.0,3.6,3.7 3.8,6.3,6.4,6.6,6.9}	1	{0.5,0.6}
2	2	{4.7,6.6}	2	{3.9}	2	{1.2,1.9,3.3,3.5,4.3 5.2,5.3,5.4,5.9,6.0,6.7}	2	{1.7}
3	3	{4.4,5.9,6.8,7.2}	3	{2.2,2.4,3.7}	3	{3.9,4.1,4.6,5.5,5.7, 5.8,6.1}	3	{1.1,2.2, 2.4,2.5}
4	4	{4.6,5.2,6.2,6.9,7.7}	4	{2.3,3.6}	4	{1.7,4.2,4.4,4.8,5.0}	4	{1.6}
5	5	{4.8,6.5}	5	{2.6}	5	{4.0,4.7,4.9}	5	{0.1,1.2,1.9}
6	6	{4.9,5.4,5.6,6.0,6.1}	6	{3.3,3.5,3.8}	6	{5.6}	6	{2.0,2.1}
7	7	{5.5,5.8,6.4}	7		7	{1.3,1.6}	7	{0.3,0.4,1.0}
8	8	{5.7,6.7}	8	{2.5}	8	{5.1, 4.5}	8	{1.4,2.3}
9	9	{5.1,6.3}	9	{2.7}			9	
10	10	{5.0}	10	{2.9}	10		10	
11			11	{3.1}	11		11	
12			12	{3.4}	12		12	{1.5,1.8}
13			13	{3.2}	13	{1.4,1.5}	13	{1.3}
14			14	{2.8}	14			
15			26	{3.0}	26			
16			29		29		29	{0.2}

Appendix 2

See Fig. 6

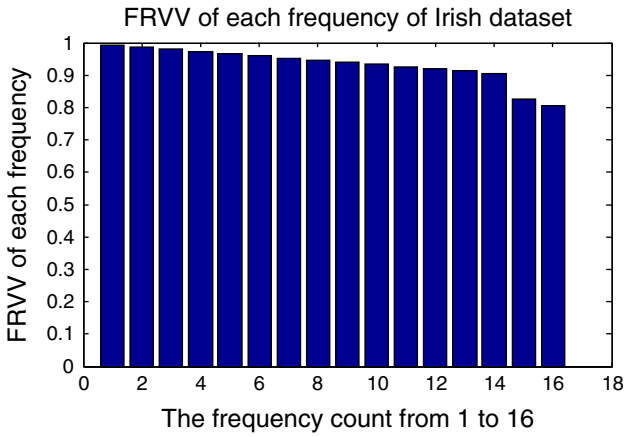

Fig. 6 Values of fuzzy random variable for different frequency of IRIS data set

References

1. Rogati, M., Yang, Y.: High -performing feature selection for text classification. In: CIKM'02, ACM, McLean, 4–9 Nov (2002)
2. Azizi, A., Pourreza, H. R.: Efficient IRIS recognition through improvement of feature extraction and subset selection. Int. J. Comput. Sci. Infor. Sec. (IJCIS). **2**, (1), (2009)
3. Uncu, O., Turksen, I.B.: A novel feature selection approach: combining feature wrappers and filters. Infor. Sci. **177**(2), 449–466 (2007)
4. Xia, H., Hu, B.Q.: Feature selection using fuzzy support vector machines. Fuzzy Optim. Decis. Mak. **5**(2), 187–192 (2006)
5. Jensen, R., Shen, Q.: Fuzzy-rough sets assisted attribute selection. IEEE Trans. Fuzzy Syst. **15**(1), 73–89 (2007)
6. Rezaee, M. R., Goedhart, B., Lelieveldt, B. P. F., Reiber, J. H. C.: Fuzzy feature selection. Pattern Recognit. **32**, 2011–2019 (1999)
7. Battiti, R.: Using mutual information for selecting features in supervised neural net learning. IEEE Trans. Neural Netw. **5**(4), 537–550 (1994)
8. Bhuyan, H. K., Kamila, N. K., Mishra, M., Jena, S. S., Bhuyan, G.: Sub-feature selection with privacy in decentralized network based on fuzzy environment. In: Proceedings of CNC 2013, Chennai, India, pp. 19–26. LNICST, Chennai, 22–23 Feb (2013)
9. Wolf, R., Schuster, A.: Association rule mining in peer-to-peer systems. IEEE Trans. Syst. Man Cybern. Part B **34**(6), 2426–2438 (2004)
10. Bhaduri, K., Wolff, R., Gianella C., Kargupta, H.: Distributed Decision tree induction in peer-to-peer systems. Stat. Anal. Data Min. J. **1**(2), 85–103, (2008)
11. Das, K., Bhaduri, K., Liu, K., Kargupta, H.: Distributed identification of Top-l inner products elements and it's application in a peer-to-peer network. TKDE **20**(4), 475–488 (2008)
12. Chen, R., Sivkumar, K., Kargupta, H.: Collective mining of Baysian networks from distributed heterogeneous data. Knowl. Inf. Syst. **6**(2), 164–187 (2004)
13. Al-Zaidy, R., Fung, B.C.M., Youssef, A.M., Fortin, F.: Mining criminal networks from unstructured text documents. Digit. Investig. **8**(3—-4), 147–160 (2012)
14. Nix, R., Kantarcioglu, M.: Incentive compatible privacy-preserving distributed classification. IEEE Trans. Dependable Secure Comput. **9**(4), 451–462 (2012)
15. Clifton, C., Kantarcioglu, M., Lin, X., Vaidya, J., Zhu, M.: Tools for privacy preserving distributed data mining. SIGKDD Explor. **4**(2), 28–34 (2003)
16. Kargupta, H., Das, K., Liu, K.: Multiparty, privacy preserving distributed data mining using game theoretic framework. In: Proceedings of PKDD'07, pp. 523–531. Warsaw (2007)
17. Zhou, B., Pei, J.: The k-anonymity and l-diversity approaches for privacy preservation in social networks against neighborhood attacks. Knowl. Inf. Syst. **28**(1), 47–77 (2011)
18. Fung, B.C.M., Wang, K., Chen, R., Yu, P.S.: Privacy preserving data publishing: a survey of recent developments. ACM Comput. Surv. **42**(4), 14 (2010)
19. Kaleli, C., Polat, H.: Privacy-preserving SOM-based recommendations on horizontally distributed data. Knowl.-Based Syst. **33**, 124–135 (2012)
20. Bhuyan, H. K., Kamila N. K., Dash, S. K.: An approach for privacy preservation of distributed data in peer-to-peer network using multiparty computation. Int. J. Comput. Sci. Issues (IJCSI). **8**(4), 2 (2011)
21. Diamantini, C., Gemelli, A., Potena, D.: Feature ranking based on decision border. In: International conference on pattern recognition, IEEE Computer Society (2010)
22. Das, K., Bhaduri, K., Kargupta, H.: A local asynchronous distributed privacy preserving feature selection algorithm for large peer to peer networks. Knowl. Inf. Syst. **24**(3), 341–367 (2014)
23. Sun, H. J., Sun, M., Mei, Z.: Feature selection via fuzzy clustering. In: Proceedings of International Conference on Machine Learning and Cybernetics, pp. 1400–1405. (2006)
24. Zhang, Y., Wu, X.B., Xiang, Z.R., Hu, W.L.: Design of high dimensional fuzzy classification systems based on multi-objective evolutionary algorithm. J. Syst. Simul. **19**(1), 210–215 (2007)
25. Xiong, N., Funk, P.: Construction of fuzzy knowledge bases incorporating feature selection. Soft Comput. **10**(9), 796–804 (2006)
26. Couso, I., L. Sánchez, L.: Higher order models for fuzzy random variables. Fuzzy Sets Syst. **159**, 237–258 (2008)
27. Couso, I., Sánchez, L.: Upper and lower probabilities induced by a fuzzy random variable. Fuzzy Sets Syst. **165**, 1–23 (2011)
28. Jesus, M.J.D., Hoffmann, F., Junco, L., S'anchez, L.: Induction of fuzzy rule based classifiers with evolutionary boosting algorithms. IEEE Trans. Fuzzy Sets Syst. **12**(3), 296–308 (2004)
29. S'anchez, L., Couso, I., Casillas, J.: Modelling vague data with genetic fuzzy systems under a combination of crisp and imprecise criteria. In: Proceedings of IEEE MCDM, Honolulu (2007)
30. S'anchez, L., Otero, J., Villar. J. R.: Learning fuzzy linguistic models from low quality data by genetic algorithms. In: FUZZ-IEEE, London. (2007)
31. Kwakernaak, H.: Fuzzy random variable-I. Definition and Theorem. Inf. Sci. **15**, 1–29 (1978)
32. Tan, P., Steinbach, M., Kumar, V.: Introduction to Data Mining. Addision-Wesley, Redwood (2006)
33. Han, J., Kamber, M.: Data Mining Concepts and Techniques, 2nd edn. Elsevier, Morgan Kaufmann Publishers, San Francisco (2006)
34. Agrawal, R., Srikant, R.: Privacy preserving data mining. In Proceedings of ACM SIGMOD International Conference on Management of Data, pp. 439–450. Dallas (2000)
35. Huang, Z., Du, W., Chen, B.: Deriving private information from randomized data. In: Proceedings of ACM SIGMOD International Conference on Management of Data, pp. 37–48. Baltimroe (2005)
36. Li, Y., Chen, M., Li, Q., Zhang, W.: Enabling multilevel trust in privacy preserving data mining. IEEE Trans. Knowl. Data Eng. **24**(9), 1598–1612 (2012)

37. Sanchez, L., Suarez, M.R., Couso, I.: A fuzzy definition of mutual information with application to the design of genetic fuzzy classifiers. In: International Conference on Machine Intelligence, pp. 5–7. Tozeur (2005)
38. Bacardit, J.: Pittsburgh generic based machine learning in the data mining era: representations, generalization, and run time. Ph.D. Thesis. La Salle-Univ. Ramon Llull (2005)
39. Sanchez, L., Suarez, M.R., Villar, J.R., Couso, I.: Some results about Mutual information based feature selection and fuzzy Discretization of vague data. In: IEEE, Fuzzy Systems Conference, FUZZ-IEEE 2007, pp 1–6. London, 23–26 July (2007)
40. Asuncion, A., Newman, D.: UCI machine learning repository, (2007)

Hemanta Kumar Bhuyan is currently a PhD candidatae in the Department of Computer Science and Engineering at Sikhya 'O' Anusandhan (SOA) University, Odisha, India. He received his M.Tech degree in Computer Science and Engineering from Utkal University, Odisha, India in 2005. He is currently working as an Assistant Professor in the department of computer science & engineering at Mahavir Institute of Engineering and Technology, Odisha, India. His research interests include privacy preserving data mining, distributed data mining, feature selection.

Narendra Kumar Kamila is a professor of computer science and engineering at C V Raman College of Engineering, Bhubaneswar under Biju Patnaik University of Technology Rourkela, India. He received the master degree from Indian Institute of Technology, Kharagpur and doctorate degree from Utkal University, India in the year 2000. Later he had visited USA for his post doctoral work at University of Arkansas in 2005. His research interest includes artificial intelligence, data privacy, image processing, and wireless sensor networking. He has several publications in international, national journals and conference proceedings. His professional activities include teaching computer science, besides he organizes many conferences, workshops and faculty development programs funded by All India Council for Technical Education, Govt. of India. However Dr. Kamila is a DSC member of Biju Patnaik University of Technology, Dr. Kamila has been rendering his best services as editorial board member to *American Journal of Intelligent System, American Journal of Advances in Networks, American journal of Networks and Communications, Reviewer of International journal of Intelligent Information System (USA), Reviewer of International journal of Automation Control and Intelligent Systems (USA), Reviewer of Elsevier Publication, Reviewer of AMSE, modelling simulation (France),* editor-in-chief of *International Journal of Advanced Computer Engineering and Communication Technology, former editor-in-chief of International Journal of Communication Network and Security (IJCNS)* and editor-in-chief of *many international conference proceedings.*