Differential Private Random Forest

Abhijit Patil and Sanjay Singh
Department of Information & Communication Technology
Manipal Institute of Technology, Manipal University, Manipal-576104, INDIA
abhijeet24patil@gmail.com, sanjay.singh@manipal.edu

***Abstract*—Organizations be it private or public often collect personal information about an individual who are their customers or clients. The personal information of an individual is private and sensitive which has to be secured from data mining algorithm which an adversary may apply to get access to the private information. In this paper we have consider the problem of securing these private and sensitive information when used in random forest classifier in the framework of differential privacy. We have incorporated the concept of differential privacy to the classical random forest algorithm. Experimental results shows that quality functions such as information gain, max operator and gini index gives almost equal accuracy regardless of their sensitivity towards the noise. Also the accuracy of the classical random forest and the differential private random forest is almost equal for different size of datasets. The proposed algorithm works for datasets with categorical as well as continuous attributes.**

I. Introduction

Privacy of individual's personal information has become an important issue in the digital world. An increasing amount of personal information is aggregated and stored in data repositories, and mined to extract useful knowledge. It is important to develop mechanisms to avoid disclosure of personal information while extracting the required knowledge from it.

Private companies, government entities and institutions such as hospitals collect vast amount of personal information about individuals who are their customers, clients or patients. Individual's personal information is private and sensitive which has to be secured from adversaries or general public. The problem of securing statistical databases from revealing the personal records from collected data has been the subject matter of research for very long time. There are many methods available such as input perturbation, output perturbation, objective perturbation [1] and exponential mechanism [2] to protect the sensitive data from any possible attack.

Differential privacy measures privacy risk by a parameter ϵ that bounds the log-likelihood ratio of the output of a (private) algorithm under two databases differing in a single individual's data [1].

By repeated execution of these algorithms, adversary analyzes the results and may infer some private information from an individual's record. Differential privacy is applied on many algorithm, such as linear regression, logistic regression and k-means clustering etc.

In our work, we wanted to investigate whether it is possible to incorporate differential privacy in any decision tree based algorithm which gives good accuracy in addition to being computationally efficient. In our work, we consider the problem of providing differential privacy using random forest [3], which has better accuracy compared to other decision tree based algorithms. Random forest runs efficiently on large data sets without deletion and it also gives estimates of what variables are important in the classification. In addition to that, major advantage of using random forest for providing differential privacy is that there is no need of pruning the tree which is a common requirement in most of the decision tree based algorithms, hence it is more computationally efficient. In this paper we study, not only the working of random forest in differential privacy framework, but also the effect of different quality functions, such as information gain [4], max operator [5], and gini index, on the accuracy and their sensitivity towards the noise with appropriate results. Whenever an adversary analyzes the results of differential private random forest, he may not infer any private information. This process is shown in Fig. 1.

Rest of the paper is organized as follows. Section II, discusses several related work on differential privacy. Section III, briefs required theoretical background on differential privacy, exponential mechanism and random forest. Section IV, explains the working of proposed differential private random forest algorithm. Section V gives experimental results and section VI discusses those results. Finally, section VII concludes the paper.

II. Related Work

There are several recent works on the use of differential privacy for practical applications. Machanavajjhala et al.[6] applied a variant of differential privacy to create synthetic datasets from U.S. Census Bureau, with the goal of using it for statistical analysis of commuting patterns in mapping applications. Chaudhuri and Monteleoni [7] proposed differentially-private algorithms for logistic regression. These algorithms ensure differential privacy by adding noise to the outcome of the logistic regression model or by solving logistic regression for a noisy version of the target function. However, in categorical or discrete attribute dataset, adding noise is not possible.

Many algorithms have been proposed to preserve privacy, but only few have considered classification algorithm in the differential privacy framework [8]. Noman Mohhamed et al. [9] applied differential private methods to release differential private dataset for data mining. They have used exponential mechanism[2] in generalization of attributes, however, there

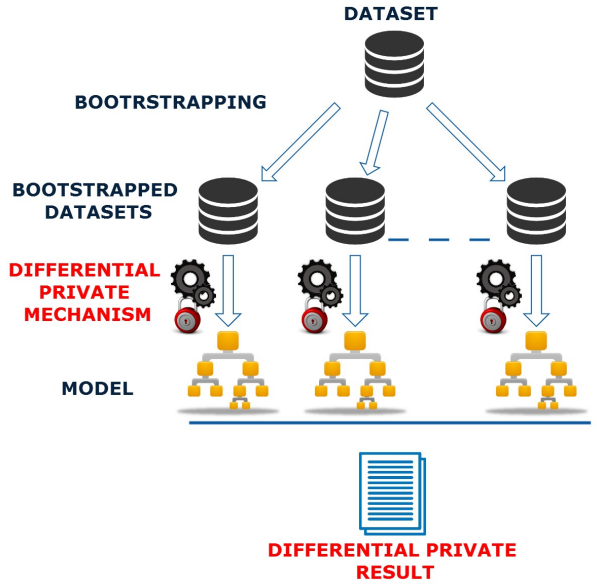

Fig. 1. Schematic diagram for providing differential privacy using random forest algorithm.

is no explanation about where to add the noise during the attribute selection process.

Several studies have compared the performance of different splitting criteria for decision tree induction, their results do not, in general, attest to the superiority of any one criteria in terms of tree accuracy, although the choice may affect the resulting tree size [10]. John Mingers[10], considers a number of different measures and experimentally examines their behavior in four domains (different type of datasets). The results show that the choice of measure affects the size of a tree but not its accuracy, which remains same even when attributes are selected randomly. McSherry and Mironov studied the application of differential privacy to collaborative recommendation systems [11] and demonstrated the feasibility of differential privacy guarantees without a significant loss in recommendation accuracy.

Arik Friedman and Assaf Schuster [5] have proposed, data mining with differential privacy using decision tree induction as an example. They have concluded that, introduction of formal privacy guarantee into a system requires the data miner to take a different approach to data mining algorithms. The major limitation of their work is, the sensitivity of quality functions such as information gain [4] and max operator [5] with respect to noise, which affects the final classification accuracy. They have fed different quality functions such as information gain, max operator and gini index into exponential mechanism and found some interesting experimental results. Their results suggests that the use of different quality functions, affect the final outcome or accuracy. The rationale is the sensitivity of those functions due to added noise; the information gain is more sensitive to noise than the max operator. Differential Private ID3 [5] gives a differential private version of ID3 algorithm. By applying exponential mechanism on the process of attribute selection it has overcome the drawback of SuLQ-based ID3 [12]. In SuLQ-based ID3, each attribute is evaluated separately and so it leads to waste of privacy budget. Differential private ID3 evaluate all the attributes in one single query, and result of which is the attribute to use for splitting.

This algorithm works on both continuous and categorical datasets, and has good accuracy compare to SuLQ-based ID3. Low accuracy with respect to information gain as quality function is the major limitation of Differential Privacy ID3 algorithm. In our work we have used differential private ID3 algorithm.

III. Theoretical Background

This section briefs the required theoretical background which are the basis for the proposed algorithm, *Differential Private Random Forest*.

Let D be the dataset, $D = \{(X_i, Y_i)\}_{i=1}^{n}$, with $X_i \in \mathbb{R}^d$ and corresponds to the data record of an individual i. The d elements of vector X correspond to different features or attributes.

A. Differential Privacy

The ϵ-differential privacy model introduced by Dwork et al.[13] assures that the removal or addition of a single item in a database does not have a substantial impact on the output produced by a private database access mechanism. Differential privacy measures privacy risk by a parameter ϵ, that bounds the log-likelihood ratio of the output of a (private) algorithm under two datasets D_1 and D_2 differing in a single individual's data. When ϵ is small, the inferences that an adversary can make observing the output of the algorithm will be similar regardless of whether that individual is in the data set or not [1].

Definition 1: A randomized mechanism $\mathcal{M}$ satisfies ϵ-differential privacy if for databases D_1 and D_2 differing on at most one element, and all $S \in$ Range($\mathcal{M}$), then [14]

$$Pr[\mathcal{M}(D_1 \in S)] \leq e^{\epsilon} \times Pr[\mathcal{M}(D_2 \in S)]$$

The probability is taken over the coin tosses in $\mathcal{M}$.

The parameter $\epsilon > 0$ is public and specified by the data owner. Lower values of ϵ provide a stronger privacy guarantee. Typically, the values of ϵ should be small, such as 0.01, 0.1, or in some cases ln2 or ln3 [14]. The value of ϵ allows us to control the level of privacy. Typical, differential privacy is achieved by adding noise to the outcome of a query. To obtain ϵ-differential privacy, calibrate the magnitude of noise according to the *sensitivity* of the function. The sensitivity of a real-valued function expresses the maximal possible change in its value due to addition or removal of a single record.

Definition 2: For any function $f : D \mapsto \mathbb{R}^d$, the sensitivity of f is defined as

$$\Delta f = max_{D_1 \sim D_2} \|f(D_1) - f(D_2)\|_1$$

for all D_1, D_2 differing in at most one record.

Given the sensitivity of a function f, the addition of noise drawn from a calibrated Laplace distribution maintains ϵ-differential privacy.

Theorem 1: Given a function $f : D \mapsto \mathbb{R}^d$ over an arbitrary domain D, the computation

$$\mathcal{M}(X) = f(X) + (Laplace(\Delta(f)/\epsilon))^d$$

provides ϵ-differential privacy.
For example, the count function over a set S, $f(S) = |S|$, has sensitivity 1. Therefore, a noisy count that returns $\mathcal{M}(S) = |S| + Laplace(1/\epsilon)$ maintains ϵ-differential privacy.

B. Exponential Mechanism

Exponential mechanism [2] has a quality function q, that scores outcomes of a calculation, where higher scores are better. For a given database and ϵ parameter, the quality function induces a probability distribution over the output domain, from which the exponential mechanism samples the outcome. This probability distribution favors high scoring outcomes (they are exponentially more likely to be chosen), while ensuring ϵ-differential privacy.

Definition 3: Let $q : (\mathcal{D}^n \times \mathcal{R}) \mapsto \mathbb{R}$ be a quality function that, given a database $D \in \mathcal{D}^n$, assigns a score to each outcome $r \in \mathcal{R}$. Let $\Delta(q) = max_{r,D_1 \sim D_2} \|q(D_1, r) - q(D_2, r)\|_1$. Let $\mathcal{M}$ be a mechanism for choosing an outcome $r \in \mathcal{R}$ given a dataset instance $D \in \mathcal{D}^n$. Then the mechanism $\mathcal{M}$, is defined by

$$\mathcal{M}(D, q) = \left\{ \text{return } r \text{ with probability} \propto exp\left(\frac{\epsilon q(D, r)}{2\Delta(q)}\right) \right\}$$

maintains ϵ-differential privacy.

C. Random Forest

Random forest is an ensemble learning method for classification (and regression) that operate by constructing a multitude of decision trees at training time and gives the class that is the mode of the classes output by individual trees [3].

In random forest, initially it creates a B number of bootstrap samples [4] from dataset D, and each bootstrap sample $Z^{*b}, b = 1, 2, \ldots, B$, is used to construct random forest tree, T_b. To determine the decision at a node of the tree, it selects m numbers of input variables at random from d variables, note that here, $m < d$. Each tree is fully grown and not pruned. For prediction, a new sample is pushed down the tree and assigned a label for the training sample in the terminal node it ends up in. This procedure is iterated over all trees, and the mode vote of all trees is reported as the random forest prediction.

IV. Differential Private Random Forest

Now we consider our proposed algorithm wherein we incorporate the concept of differential privacy in the classical random forest algorithm. To make random forest private, we have used Differential private ID3 [5], to construct the random forest tree on each bootstrap sample. Differential private ID3, passes all the d attributes or variables to exponential mechanism in order to choose the best split attribute. While using random forest, we need to pass the m number of attributes to exponential mechanism instead of passing all d attributes.

The input to the differential private random forest algorithm is dataset D with attributes $\mathcal{A} = \{A_1, \ldots, A_d\}$ and a class attribute C. Each record in the dataset belongs to an individual person.

In our previous work [15], we have proposed a differentially private random forest algorithm for only categorical data sets. In [15], we have considered the datasets with only categorical attributes, wherein the input to the differential private random forest is a raw categorical datasets. However, in real life datasets may not always be with categorical attributes there are numerous scenarios where datasets are in the continuous form. In this paper we have extended our earlier work for continuous data sets as well. In order to modify our differentially private random forest algorithm for categorical data sets for continuous data sets, we need a preprocessing of the continuous data sets.

A. Preprocessing

Preprocessing of a given dataset is done by considering the discretization of continuous attributes. The discretization process used in our work is Fayyad and Irani's Entropy-based discretization [16]. The process is described as below.

Entropy-based discretization is a supervised discretization method, where the boundaries for discretization are selected by using class information entropy of candidate partitions. It considers one large interval containing all known values of an attribute then recursively partitions this interval into smaller sub-intervals. This recursive process stops until some stopping criterion, such as Minimum Description Length(MDL) principle [17] or an optimal number of intervals is achieved. Entropy is used as splitting criterion and MDL principle is used as stopping criterion. The Entropy is defined in definition.4.

Definition 4: The entropy of the dataset D, with respect to the class attribute C is defined as:

$$H_C(D) = -\sum_{c \in C} \frac{\tau_c}{\tau} \log \frac{\tau_c}{\tau}.$$

Definition 5: Information gain (IG), is defined as, the measure of the difference in entropy with and without split on an attribute A, in dataset D, and is given by

$$IG(A, D) = H_C(D) - H_{C|A}(D)$$

where, $H_{C|A}(D) = -\sum_{j \in A} \frac{\tau_j^A}{\tau} \cdot H_C(D_j^A)$.
Information gain can be approximated with noisy counts for τ_j^A and $\tau_{j,c}^A$ to obtain:

$$\overline{IG_A} = \sum_{j=1}^{|A|} \sum_{c=1}^{|C|} N_{j,c}^A \cdot \log \frac{N_{j,c}^A}{N_j^A}$$

where, $\overline{IG_A}$ is noisy information gain of attribute A.

Given a set of samples S, continuous attribute A and number of classes k. Each value of A, is considered as a potential split-point T. The split-point with the lowest entropy is chosen to split the range into two intervals. The splitting is continued until a stopping criterion is satisfied. The stopping criterion used is MDL principle, which stops the splitting when,

$$IG(S,T) = H(S) - H(S,T) < \delta$$

where, IG is information gain given in definition 5. and $H(S)$ is an entropy of S, T is the potential interval boundary that splits S into S_1(left) and S_2(right) parts, and

$$\delta = \frac{\log_2(n-1) + \log_2(3^k - 2) - [mH(S) - m_1H(S_1) - m_2H(S_2)]}{n}$$

where m, is the number of classes in each S_i and n is the total number of samples in S.

B. Algorithm

We use following notations in algorithm 1 : D refers to a set of records, $\tau = |D|$, r_A and r_C refer to the values that record $r \in D$ takes on attributes A and C respectively, $D_j^A = \{r \in D | r_A = j\}, \tau_j^A = |D_j^A|, \tau_c = |\{r \in D | \tau_C = c\}|$, and $\tau_{j,c}^A = |\{r \in D | r_A = j \wedge r_C = c\}|$. We use t to denote size of an attribute with maximum number of attribute values. $\overline{A}$ is the split attribute selected by exponential mechanism. The noisy counts that is, releasing of the number of records in a dataset perturbed by symmetric exponential (Laplace) noise [11] is referred by N for τ. We use P_ϵ as a overall privacy budget for an input dataset, that is, before bootstrapping. P_ϵ is distributed among the B number of bootstrap samples, because for each sample we are constructing a differential private random tree. We use ϵ' to refer to privacy budget for a single tree and ϵ refers to privacy budget for each level or depth, h of the tree.

The single tree budget ϵ' is used by limiting the depth h of the tree and assigning an equal share of the budget, ϵ for each level of the tree including leaves. According to composition property of differential privacy [18], queries on different nodes on the same level do not accumulate, as they are carried out on disjoint sets of records.

Within each node, half of the allocated budget is used to evaluate the number of instances and other half is used to determine the class counts (in leaves) or evaluate the attributes (in nodes). Class counts are calculated on disjoint sets, so each query can use the allocated ϵ'. To prevent the misuse of privacy budget, exponential mechanism is used to choose the best split attribute which avoids splitting of allocated budget ϵ' among multiple queries, and the entire budget is used to find the best attribute in a single query. The quality function, q, is provided to the exponential mechanism scores each attribute according to the splitting criterion which we explain next.

C. Splitting Criteria

To know the order, in which attributes must be chosen to split the data, we need some measures that would allow us to compare the attributes on some scale and choose one above the other. In our work we have considered two splitting criteria, information gain and max operator, and used them as a quality function for exponential mechanism.

Algorithm 1 Differential Private Random Forest Algorithm

1) **procedure** DiffPRandomForest $(D, \mathcal{A}, C, h, B, P_\epsilon)$
2) **Input:** D - private dataset, $\mathcal{A} = \{A_1, \ldots, A_d\}$- a set of attributes, *C*- class attribute, *h*- maximal tree depth, *B*- number of bootstrap sample, P_ϵ-differential privacy budget on D.
3) $\epsilon' = \frac{P_\epsilon}{B}$
4) $\epsilon = \frac{\epsilon'}{2(h+1)}$
5) for b=1 to *B*
   a) Draw a bootstrap sample Z^* of size τ from the training data D
   b) Grow a random-forest tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until $\mathcal{A}$=0 or *h*=0
      i) Select *m* variables at random from the *d* variables
      ii) Use exponential mechanism to select variable/split-point among the *m* variables with probability,
      $$\frac{exp\left(\frac{\epsilon}{2\Delta q} q(Z^*, A)\right)}{\sum_{A \in m} exp\left(\frac{\epsilon}{2\Delta q} q(Z^*, A)\right)}$$
      where, *A* is an attribute , $q(Z^*, A)$ is information gain.
      iii) Split the node into two child nodes.
6) Output the ensemble of trees $\{T_b\}_1^B$.
   To make a prediction at new point *x*:
   Classification: Let $\widehat{C}_b(X)$ be the class of the b^{th} random-forest tree.
   Then $\widehat{C}_{brf}(X)$ = *majority vote* $\{\widehat{C}_b(X)\}_1^B$

The quality function q scores each outcome and induces probability distribution over output domain, from which exponential mechanism samples the outcome. Each quality function has sensitivity.

In step 5b(ii) of algorithm 1, the exponential mechanism selects the attribute for splitting with the probability [9],

$$\frac{exp\left(\frac{\epsilon}{2\Delta q} q(Z^*, A)\right)}{\sum_{A \in m} \exp\left(\frac{\epsilon}{2\Delta q} q(Z^*, A)\right)}$$

where, $q(Z^*, A)$ is the information gain used to score each attribute $A \in m$, Z^* is the bootstrap sample and Δq is the sensitivity of scoring function q.

D. Stopping Criteria

The recursive process of tree construction stops either following conditions,

1) When attribute set $\mathcal{A}$, becomes empty, that is, $\mathcal{A}$=0
2) When all the samples belongs to the same class
3) When tree reaches the maximum height h

Once, either of these criteria succeeds, instead of accurate majority class, noisy count of majority class is used to determine the class.

In DiffPID3, when there are only few samples in the leaf, the noise will overcome the accurate counts and wrong class will be chosen. To avoid this, a threshold has been introduced, which depends on the noise added. Once the number of samples are less than this threshold, further recursive process stops. In our study, we found that, threshold criteria, reduces the classification accuracy of DiffPID3. And there is no effect on accuracy of different quality functions such as, information gain, max operator and gini index.

Random forest gives good accuracy for any size of dataset, and dataset with more or less number of attributes. The main reason behind this is the bootstrapping of data and randomness in variable selection. We tried, both the cases, that is, with and without threshold criteria in differential private random forest and found that, there is no much difference in classification accuracy. As, we are stopping the recursive process by the height of tree h and an empty set of attributes, there is no need to include, threshold criteria.

E. Sensitivity of Quality Functions

The sensitivity of quality functions: information gain and max operator is discussed further. We denote the quality function for information gain as,

$$q_{IG}(D, A) = IG(D, A) = H_C(D) - H_{C|A}(D).$$

The sensitivity of this function is $\Delta(q_{IG}) = log_2|C|$, where $|C|$ is the domain size of the class attribute C. It is because, the value of the entropy $H_C(D)$ is between 0 and $log_2|C|$. And, the value of conditional entropy $H_{C|A}(D)$ lies between 0 and $H_C(D)$. Therefore the maximum change of q_{IG} due to addition or removal of a record is bounded by $log_2|C|$.

Max operator corresponds to the node misclassification rate by picking the class with the highest frequency. Quality function for max operator is defined as,

$$q_{Max}(D, A) = \sum_{j \in A}(max_c(\tau^A_{j,c})).$$

The sensitivity of this function is $\Delta(q_{Max}) = 1$. Since a record can change the count only by 1.

Sensitivity of Gini index is $\Delta(q_{gini}) = 2$. Our results shows that information gain and max operator gives almost equal accuracy for different size of data. The reason for getting equal accuracy is, randomness in attribute selection (step 5b(i) of algo.1). It is one of the main property of Random Forest.

V. Experimental Results

This section gives experimental results of our work. We experimented the algorithms on synthetic datasets and real datasets. The datasets considered in experiments have both categorical as well as continuous attributes.

A. Synthetic Datasets

We generated synthetic datasets using WEKA [19], an open source machine learning software, where each dataset consists of 10 categorical attributes. The size of dataset ranging from 1000 to 10000. In differential private random forest we have generated 20 bootstrap samples and constructed a differential private tree on each bootstrap sample. The depth of each tree h is taken as 5. Privacy budget ϵ' considered in our work is 0.1, 0.25, 0.5, 0.75 and 1. We have not excluded the threshold criteria. The quality functions used to determined the effect on accuracy are *information gain*, *max operator* and *gini index*. The accuracy and deviation of DiffPID3 and Differential private random forest and classical random forest are given in Table.I. The results are shown in Fig. 2-3, where y-axis is the average accuracy and x-axis is the privacy budget ϵ'. The plot legend marker, *DiffPRF-ig* represents differential private random forest with information gain, *DiffPRF-Max* represents differential random forest with max operator, and *DiffPRF-Gini* represents differential random forest with gini index.

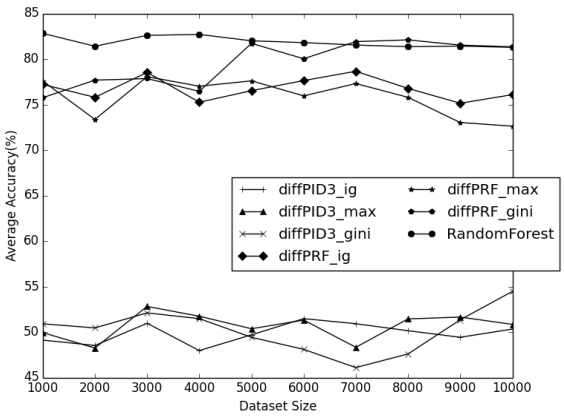

Fig. 2. Comparing accuracy with different size of dataset where, $\epsilon' = 0.1$ for DiffPID3 and differential private random forest.

B. Real Datasets

The real datasets which we have chosen for our experiment are given in Table.II with the required summary. The initial four datasets has only categorical attributes and rest has both categorical and continuous attributes. All datasets are real datasets and are taken from UCI Machine Learning Repository [20]. The classification accuracy of random forest and differential private random forest for real datasets is given in Table.III. We have considered the accuracy with information gain, max operator and gini index separately. Table.IV and Table.IV shows the classification accuracy on mushroom database and

TABLE I
ACCURACY OF DIFFPID3, DIFFERENTIAL PRIVATE RANDOM FOREST (DIFFPRF) AND RANDOM FOREST WITH DIFFERENT QUALITY FUNCTIONS AND PRIVACY BUDGET ϵ'= 0.1

No.of samples	DiffPID3-Ig	DiffPID3-Max	DiffPID3-Gini	DiffPRF-Ig	DiffPRF-Max	DiffPRF-Gini	RandomForest
1000	49.11±15	49.93±15	50.90±16	77.21±4	78.58±3	79.59±4	84.82±1
2000	48.52±15	48.24±14	50.45±15	75.79±4	73.36±3	76.36±4	84.40±1
3000	50.96±15	52.81±15	52.50±16	78.55±3	78.06±3	78.78±4	83.60±1
4000	47.96±14	51.73±15	51.47±15	75.25±4	77.01±4	77.59±4	84.70±1
5000	49.69±15	50.36±15	49.42±16	76.52±3	77.60±2	81.72±2	84.00±1
6000	51.47±15	51.31±15	48.09±14	77.63±3	75.95±3	80.03±2	83.80±2
7000	50.91±15	48.32±14	46.09±14	78.65±1	77.30±3	81.91±2	84.54±1
8000	50.14±15	51.44±15	47.59±14	76.77±1	75.79±3	82.10±4	84.35±2
9000	49.42±14	51.64±15	51.34±16	75.14±3	73.02±3	81.53±3	84.42±2
10000	50.32±14	50.84±15	54.46±16	76.08±3	72.62±2	80.34±2	84.28±2

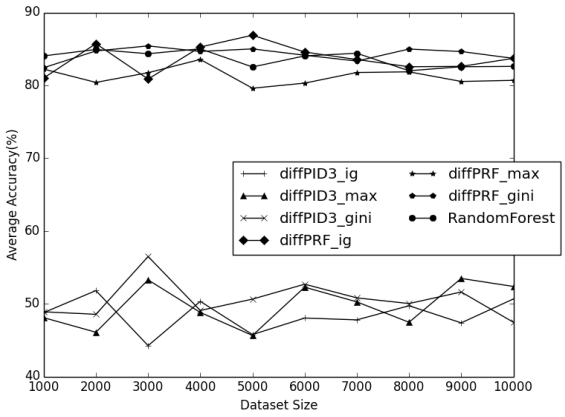

Fig. 3. Comparing accuracy with different size of dataset where, $\epsilon' = 1$ for DiffPID3 and differential private random forest.

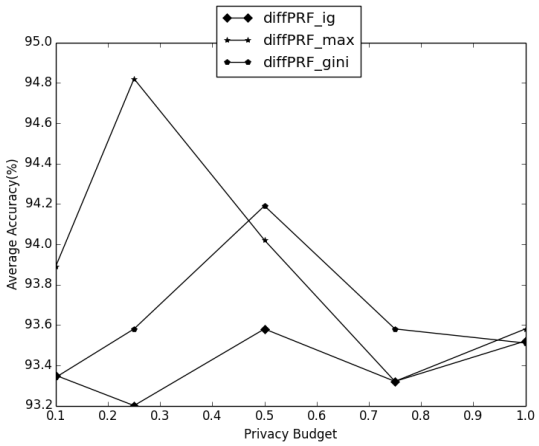

Fig. 4. Comparing accuracy with different quality functions for Mushroom dataset.

adult database respectively in different privacy budgets with different quality functions. The results are shown in Fig. 4-5 with legends similar to the synthetic databases. This shows that, differential private random forest has almost similar accuracy when compared to the classical random forest.

VI. DISCUSSION

Table.I gives the accuracy of DiffPID3, differential private random forest and random forest with different quality functions and privacy budget ϵ'= 0.1. The datasets are synthetic with categorical attributes. From Table. I we infer that the accuracy of differential private random forest is much better than DiffPID3. For this datasets, we have considered the threshold criteria in differential differential private algorithm which reduces the accuracy when compared to random forest. The Fig. 2 and 3 shows the same result with privacy budget 0.1 and 1 respectively.

Table.III gives the classification accuracy for random forest and differential private random forest with privacy budget 0.1 on the real datasets from Table.II where, both the algorithms, have almost same accuracy. We have not considered the threshold criteria for these datasets as it reduces the accuracy. In Table.IV and Table.V the accuracy of differential private

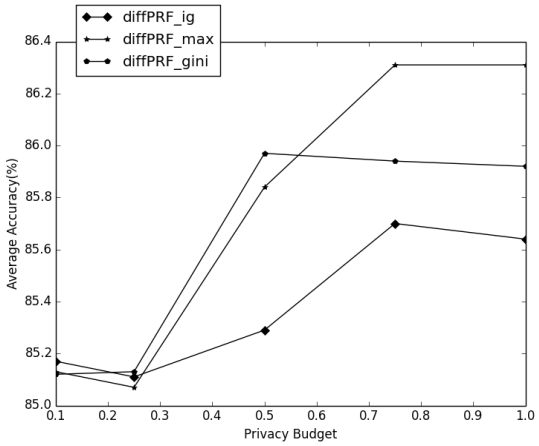

Fig. 5. Comparing accuracy with different quality functions for Adult dataset.

TABLE II
REAL DATASETS

Dataset	Number of samples	Number of attributes
Tic-Tac-Toe Endgame database [21]	958	36
Car Evaluation Database [22]	1728	9
Mushroom Database [23]	8124	21
Nursery Database [24]	12960	8
Adult datset [25]	48842	4
Credit Approval Dataset [26]	690	14
Iris Dataset [27]	150	4

TABLE III
CLASSIFICATION ACCURACY OF RANDOM FOREST AND DIFFERENTIAL PRIVATE RANDOM FOREST WITH PRIVACY BUDGET $\epsilon' = 0.1$

Dataset	Random Forest(%)	DiffPRF-InfoGain(%)	DiffPRF-Max(%)	DiffPRF-Gini(%)
Tic-Tac-Toe database	90.01±1	89.63±1	86.67±1	92.86±1
Car Database	92.03±2	92.04±1	89.95±2	92.04±2
Mushroom Database	93.70±1	93.58±1	93.14±1	93.51±1
Nursery Database	91.96.±1	87.17±1	87.35±1	88.74±1
Adult Database	86.24.±1	85.29±1	85.73±1	85.62±1
Credit Approval Database	89.07.±1	88.99±2	88.55±3	88.91±2
Iris Database	99.00±1	100.00±0	100.00±0	100.00±0

TABLE IV
CLASSIFICATION ACCURACY OF DIFFERENTIAL PRIVATE RANDOM FOREST FOR MUSHROOM DATASET WITH DIFFERENT QUALITY FUNCTIONS

Privacy budget(ϵ')	DiffPRF-InfoGain(%)	DiffPRF-Max(%)	DiffPRF-Gini(%)
0.1	93.35±1	93.89±1	93.34±1
0.25	92.04±1	93.82±2	93.58±2
0.5	93.58±1	94.07±1	94.19±1
0.75	93.32±1	93.82±1	93.58±1
1.00	93.48±1	93.14±1	93.51±1

TABLE V
CLASSIFICATION ACCURACY OF DIFFERENTIAL PRIVATE RANDOM FOREST FOR ADULT DATASET WITH DIFFERENT QUALITY FUNCTIONS

Privacy budget(ϵ')	DiffPRF-InfoGain(%)	DiffPRF-Max(%)	DiffPRF-Gini(%)
0.1	84.71±1	85.13±1	85.12±1
0.25	85.11±1	85.07±1	85.13±1
0.5	85.29±1	85.84±1	85.97±1
0.75	85.70±1	86.31±1	85.94±1
1.00	85.64±1	86.31±1	85.92±1

random forest with different privacy budgets and different quality functions has been given for mushroom dataset and adult dataset respectively. The attributes in Mushroom dataset are categorical where as attributes of adult dataset are both categorical and continuous. Fig. 4 and 5 gives the pictorial representation of the results given in Table.IV and Table.V.

In our work, we have considered all bootstrap samples are disjoint datasets, and so we have distributed the privacy budget P_ϵ in B number of bootstrap samples. But, actually bootstrap samples are not disjoint datasets as they are created by sampling with replacement from original dataset, D. Constructing trees on B bootstrap samples is a parallel process, therefore each sample should get an equal privacy budget.

We have achieved almost equal accuracy in both, random forest and differential private random forest, by considering different quality functions. In [5], their results says, as the information gain is most sensitive to the noise, it gives low accuracy, and max operator gives high accuracy since it is least sensitive to noise. However, with our proposed algorithm, *differential private random forest*, which makes use of such quality functions, it is not necessary to be true that the accuracy depends on their sensitivity. Since our algorithm gives same accuracy as the random forest, it can be used on any real time categorical and continuous datasets to achieve faster and better differential privacy.

VII. CONCLUSION

In this paper we have addressed the problem of providing differential privacy on datasets with categorical and continuous attribute with a faster and computationally efficient algorithm. To achieve this, we have incorporated the concept of differential privacy in classical random forest algorithm. Our results, demonstrates that, the random forest and differential private random forest has almost same accuracy for datasets

of different sizes. In addition, we conclude that, with proposed algorithm, the quality functions such as information gain and max operator gives almost equal accuracy regardless of their sensitivity towards the noise. Experimental results confirms that with any privacy budget the proposed differential private random forest algorithm gives acceptable accuracy without any additional computational cost. The discretization process loses some information of original dataset. To overcome this, in our future work we are going to make the discretization process private with the help of exponential mechanism.

REFERENCES

[1] A. Sarwate and K. Chaudhuri, "Signal processing and machine learning with differential privacy: Algorithms and challenges for continuous data," *Signal Processing Magazine, IEEE*, vol. 30, no. 5, pp. 86–94, 2013.

[2] F. McSherry and K. Talwar, "Mechanism design via differential privacy," in *Foundations of Computer Science, 2007. FOCS '07. 48th Annual IEEE Symposium on*, 2007, pp. 94–103.

[3] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[4] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, ser. Springer Series in Statistics. New York, NY, USA: Springer New York Inc., 2001.

[5] A. Friedman and A. Schuster, "Data mining with differential privacy," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '10. New York, NY, USA: ACM, 2010, pp. 493–502. [Online]. Available: http://doi.acm.org/10.1145/1835804.1835868

[6] A. Machanavajjhala, D. Kifer, J. Abowd, J. Gehrke, and L. Vilhuber, "Privacy: Theory meets practice on the map," in *Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on*, April 2008, pp. 277–286.

[7] K. Chaudhuri and C. Monteleoni, "Privacy-preserving logistic regression," in *NIPS*, 2008, pp. 289–296.

[8] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments," *ACM Comput. Surv.*, vol. 42, no. 4, pp. 14:1–14:53, Jun. 2010. [Online]. Available: http://doi.acm.org/10.1145/1749603.1749605

[9] N. Mohammed, R. Chen, B. C. Fung, and P. S. Yu, "Differentially private data release for data mining," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '11. New York, NY, USA: ACM, 2011, pp. 493–501. [Online]. Available: http://doi.acm.org/10.1145/2020408.2020487

[10] J. Mingers, "An empirical comparison of selection measures for decision-tree induction," *Mach. Learn.*, vol. 3, no. 4, pp. 319–342, Mar. 1989. [Online]. Available: http://dx.doi.org/10.1023/A:1022645801436

[11] F. D. McSherry, "Privacy integrated queries: An extensible platform for privacy-preserving data analysis," in *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '09. New York, NY, USA: ACM, 2009, pp. 19–30. [Online]. Available: http://doi.acm.org/10.1145/1559845.1559850

[12] A. Blum, C. Dwork, F. McSherry, and K. Nissim, "Practical privacy: The sulq framework," in *Proceedings of the Twenty-fourth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, ser. PODS '05. New York, NY, USA: ACM, 2005, pp. 128–138. [Online]. Available: http://doi.acm.org/10.1145/1065167.1065184

[13] C. Dwork, "Differential privacy: A survey of results," in *Proceedings of the 5th International Conference on Theory and Applications of Models of Computation*, ser. TAMC'08. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 1–19. [Online]. Available: http://dl.acm.org/citation.cfm?id=1791834.1791836

[14] ——, "A firm foundation for private data analysis," *Commun. ACM*, vol. 54, no. 1, pp. 86–95, Jan. 2011. [Online]. Available: http://doi.acm.org/10.1145/1866739.1866758

[15] A. Patil and S. Singh, "Differential private radom forest: An efficient technique for introducing differential privacy in categorical datasets," Communicated to ACM SIGKDD, 2014.

[16] U. M. Fayyad and K. B. Irani, "Multi-interval discretization of continuous-valued attributes for classification learning," in *IJCAI*, 1993, pp. 1022–1029.

[17] P. D. Grünwald, *The Minimum Description Length Principle (Adaptive Computation and Machine Learning)*. The MIT Press, 2007.

[18] Differential privacy. [Online]. Available: http://en.wikipedia.org/wiki/Differential_privacy

[19] Weka: Data mining software in java. [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka/

[20] UCI machine learning repository. [Online]. Available: http://archive.ics.uci.edu/ml/

[21] Tic tac toe game. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/TicTacToeEndgame/

[22] Car evaluation. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/CarEvaluation/

[23] Mushroom database. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Mushroom

[24] Nursery dataset. [Online]. Available: http://archive.ics.uci.edu/ml/machine-learning-databases/nursery/nursery.data

[25] Adult dataset. [Online]. Available: http://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data

[26] Credit approval datset. [Online]. Available: https://archive.ics.uci.edu/ml/machine-learning-databases/credit-screening/crx.names

[27] Iris datset. [Online]. Available: http://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data