Improving the Performance of Various Privacy Preserving Databases using Hybrid Geometric Data Perturbation Classification Model

Sk. Mohammed Gouse[1], Dr.G.Krishna Mohan[2]
Department of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation
Vaddeswaram, 522502
Andhra Pradesh. India

***Abstract*—As the size of the privacy preserving databases is increasing, it is difficult to improve the privacy and accuracy of these databases due to dimensionality and runtime. However, most of the traditional privacy preserving models are independent of privacy and runtime. Also, it is essential to preserve the privacy of the large sensitive attributes before publishing it to the third-party servers. As a result, a novel framework is required to improve the privacy as well as accuracy on the high dimensional privacy preserving data with less runtime. In order to improve the privacy, accuracy and runtime of the traditional privacy preserving models, a hybrid perturbation based privacy preserving classification model is proposed on the multiple databases. In this work, a new data transformation approach, hybrid geometrical perturbation approach and hybrid boosting classifier are proposed in order to enhance the overall efficiency of the model on the privacy preserving databases. In this work, a hybrid geometric perturbation approach is used to enhance the privacy preserving on the sensitive attributes. Initially, a pre-processing method is applied on the input dataset in order to remove the noise in the feature values. A hybrid machine learning classifier is proposed to predict the privacy preserving class label based on the training data. Experimental results represents the proposed hybrid geometric perturbation based boosting classifier has better statistical accuracy, recall, precision and runtime than the conventional models.**

Keywords*—Privacy preserving databases; machine learning; perturbation; high dimensionality; data filtering; data classification

I. INTRODUCTION

Data mining focuses on the problem of discovering patterns that are unknown or hidden. It includes building data models, providing a human-comprehensible statistical summary of data, deciding strategies based on mined information [1]. Recently, researchers have drawn much attention to integrate utility constraints into data mining tasks. Utility mining is commonly used in many practical applications. A sensitive pattern is the repeated object with a sensitive information. The datasets used for data mining are represented in centralized or distributed way. In the centralized way, data are stored in the physical location, but that data accessibility / possession is involved. In the distributed manner, data are shared by two or more parties who do not really have trust in their personal information but are interested in the extraction of their common data. The dataset can be heterogeneous, i.e. horizontally partitioned, if each group has the same set of records with various sub-sets of attributes. Centralized data is usually more complete than a portion of the distributed data, as it contains complete records and attributes for collecting and mining purposes. Many real-time applications, telecommunications networks, internet traffic flows, online banking and financial transactions, retail markets, manufacturing process data, sensor-based application data flows, satellite data, research laboratory data, electrical grids, engineering data, and other dynamic environments often use data mining tools and techniques. Data streams are enormous in volumes and possibly infinite. To recognize trends and patterns, these data streams need to be analysed, which benefit us in isolating anomalies and predicting future behaviour. However, due to some reasons, most notably privacy considerations, data proprietors or originators may not be willing to accurately discover the true values of their data. A certain amount of privacy preservation must therefore be done on the data before it can be made widely accessible. Data understanding is important and is combined with the need to use appropriate algorithms to preserve privacy. Various approaches such as data perturbation, k-anonymity, association rule mining, masking and encryption have been suggested for this purpose. It is not possible to apply existing techniques directly to data streams. In addition, robust assurances on the maximum permitted interval between incoming data and its anonymous output with minimum data losses and maximum privacy gain are required in data mining applications. Another approach to privacy preservation is to perform anonymization that ensures that the record of any individual in a dataset cannot be distinguished from a group of similar individuals. The availability of raw data is the most significant consideration in data mining privacy. For detailed statistical details about the data, the data miner should not be able to access all sensitive information into its original form. This calls for more rigorous data mining techniques, which will intentionally modify data in order to mask sensitive information and preserve the data statistics inherent in mining. The latest trend in corporate cooperation is that they want to exchange data and mining findings to help each other. Nevertheless, the disclosure of sensitive information also increased the potential threat. Sanitization of information is the process that covers the sensitive items in the source

database by appropriate modification and exposes the updated database [2]. In this work, they presented an efficient algorithm to maintain the privacy of high-value items from mining that extends our proposal to weighted utilities. The majority of data mining techniques that safeguard privacy turn original data into technologies or algorithms for data mining to decrease performance. There is also a common compromise between privacy and accuracy, but this compromise is endured by certain particular algorithms used for protecting privacy. Deep learning is a multi-layered data processing network that consists of multiple levels of abstraction to train the data for pattern analysis [3]. This network uses a non-linear transformation approach to transform and learn the data in each level. Recently, a large number of composite functions have been used in the deep learning framework for pattern analysis.

Data partitioning, there are two scenarios that require using of cluster analysis in a distributed way. In the first, the volume of data that is to be analysed is fairly great. Therefore, this requires a huge amount of computational effort—so much so, sometimes, it is not feasible to complete this computation. In such a case, a better alternative is to split the data and cluster it in a distributed manner and, finally, unify the distributed results. In centralized database, data will be located and maintained at single place where as in distributed database, data may be distributed vertically or horizontally to various sources. When the database is centralized, all the data is stored in one place. This type of database is completely different from the distributed database. One of the issues the centralized database faces is that as the entire data resides at one central location [4], there can be problems with bottle-necks occurring at key points where the data is released or assimilated.

Anonymity is "nameless." Anonymity is the identification of the information with their identity. Data anonymization is the process of removing personal information from the dataset to protect the privacy of individuals and allows data users and holders to safely reveal data for data analysis, decision making, testing and other purposes so that people whose information is in the dataset remain anonymous. Even if the specific identifiers are removed, the availability of individual's background information (e.g. in the public voter list) makes it easier for the adversary to re-identify individuals by linking the released data making it very hard to publish data without disclosing privacy [5]. Once the data is released to the third party, it is hard for the owners to control the way the data is manipulated.

K-anonymity protects privacy against the identification of records; however, it is not generally successful for protecting privacy against inference attacks of the sensitive attributes. k-anonymity is characterized as the degree of inference data protection. For example, a politician who intends to be elected to a post in the governance of a state utilizes the medical history of his opponent in demonstrating to the populace that his opponent cannot or is not ready to deal with the obligations as an agent of the state due to his medical problems. In the former scenario, l-diversity [6] fails to prevent attribute disclosure because the distribution for the real population is different from the dataset. K-anonymity is designed for single data set where each row represents a different person. In case of relational database, k-anonymity might distort data too much or leak privacy. They proposed L-diversity to avoid attribute linkage attack. L-diversity demands that at least one responsive attribute value in each quasi identifier (QID) class [7]. This provision also satisfies the k-anonymity criterion where k= l. L-diversity varies from k-anonymity, while k-anonymity demands that a group contain at least k individuals with the same QID, l-diversity means that a group contain at least l of sensitive attributes.

L-diversity does not offer sufficient protection against probabilistic attack because some attributes appear more often than others [8]. In probabilistic, the sensitive attribute is inferred because it appears more frequently than other sensitive attributes and therefore attacker can infer that his victim must also have that value for the sensitive attribute. Isolating the sensitive attributes are considered as anonymous. The underlying principle here is isolation: if it cannot be isolated from its neighbours, a record is personal. In particular, when removed from the database, an opponent takes advantage of discovering the identity of the data. This is embedded in the breach of privacy that anonymizes a server. The attacker targets a server when entire data is accessed as a single large entity. If the selected data are removed from the server, the opponent cannot detect missing data and must change the attack strategy. Re-identification of individual records through quasi-identifiers is one of the major types of privacy outbreaks. Anonymization solves this type of attack. The idea behind k-anonymity is to suppress or generalize the publicly available selected data in order to make each record very similar from at least k-1 other records. Sensitive data can therefore be linked to collections of at least k size records. Quasi-identifier attribute values are a set of minimum values for the information attribute that can identify individuals in combination with other dataset. K-anonymity is intended to prevent the privacy of individuals without altering the attribute values. The traditional k-anonymity cannot be applied directly to the census data primarily for static dataset. The K-anonymity approach is the most widely used in PPDM while maintaining confidentiality.[9] proposed a K-anonymity approach by splitting the original dataset into data estimates, so that each one follows the K-anonymity. A classifier was trained on each projection and then an unknown instance was classified by combining all classifiers.

Perturbation is known for its long history, simplicity and effectiveness. It works by replacing original data with synthetic data which has similar statistical properties. Attacker cannot gain sensitive information from perturbed data because it does not correspond to original data. The downside of perturbation is that the data is meaningless for humans and it is only useful for computing statistical properties such as minimum, maximum, average, mean and so on. Additive noise is perturbation method that works by adding some random value to original value so that statistical properties of the original table would not differ too much from original ones. The downside of additive noise is that it does always offer sufficient protection to sensitive attribute. For example, when there is high correlation between QID and sensitive attribute and noise is low, the sensitive attribute's original value can be

covered from perturbed data [10]. The perturbation function requires a minor or major alteration of the problem-solving scenario to mathematically obtain the expected return. The perturbation functions were concerned with mathematical issues dealing with duality and primacy. The name of the function is appropriate for those which alter or trigger function changes at the start of the problem, and the function is twofold which is generally used to modify the limitations in order to obtain the desired solution. This contrasts with the previously proposed data mining strategies focused on additive random perturbation in order to show a significant breach of privacy. It also discusses the possibilities of proposed feature filtering techniques on various data types and interference approaches such as discrete and exclusive data or noise. Such data are widely available as statistical or categorical data. Numeric data are values that can be enumerated by categorical data. As the data of a database typically consists of ordered objects like tables and instances, the whole table or instance is not affected by the identity as a whole. The analyst or the miner is aware of the table or example but the information within the organizations are held privately. The sections or structural elements of the object are therefore chosen to cause randomization. In a database, each user typically comes up with a table consisting of multiple attributes where the user may pick the set of attributes for the query or where the attributes are appropriate for the query operations.

Increasing amounts of personal data collected and processed by companies also increases the complexity of information systems that protect information. Mainly, Privacy Preserving Data Mining (PPDM) problem focuses on two important aspects. Research's first facet: maintaining server confidentiality based on analysts ' confidence rates and key attributes for their data mining queries. The second facet of analysis is to determine the level of sensitivity of the information disseminated from the database based on the queries of the analysts. In centralized database, data will be located and maintained at single place whereas in distributed database, data may be distributed vertically or horizontally to various sources. When the database is centralized, all the data is stored in one place. This type of database is completely different from the distributed database. One of the issues the centralized database faces is that as the entire data resides at one central location, there can be problems with bottle-necks occurring at key points where the data is released or assimilated. As a result, when looking for the availability of data, the efficiency with which it is retrieved is not as strong as in the distributed database system.

The rest of the paper is organized as follows. Section 2, describes the related works of the privacy preserving models and its limitations. Section 3, describes the proposed solution to the privacy preserving based machine learning framework on high dimensional data. Section 4, describes the experimental results and analysis. Finally, we conclude the paper in Section 5.

II. Related Works

Privacy Preservation Data Mining (PPDM) is a data-protection research field focused on personally identifiable information that is considered for the creation of data-mining information systems. Therefore, numerous efforts have been made to integrate data protection techniques with data mining algorithms. The current data storage technologies for data extraction are viewed in four dimensions: (i) data delivery (central or distributed); (ii) modification used (encryption, perturbation, generalization, etc.) to sanitize data; (iii) data mining algorithms optimized for the protection of privacy techniques; (iv) data mining techniques; This study incorporates techniques for noise generation that represent the sensitivity of the attributes and disturbance techniques. Data analysis, usually a realistic, multi-story business procedure, involves people using standardized methods to detect and analyse suitable problems, find approaches and techniques for implementation, and achieve measurable results. In general, information on privacy for data mining is taken as in tuples that contain several attributes. Each privacy data is scanned and transformed into normalized continuous data. The main issues of the privacy datasets are high dimensionality and imbalance nature. Traditional machine learning classifiers consider subset of features for classification and privacy prediction with high true negative rate and error rates. Attribute selection is used to compute the measure for each feature and rank them accordingly. These ranking methods select the top ‘k’ features based on highest rank and eliminate those having lower feature ranks.

The Privacy Preserving Data Mining (PPDM) problem in this traditional work concentrates on two important aspects. The first facet of the research: Assuring privacy of database based on the trust levels of the analysts and with respect to the key attributes for their data mining queries. The second facet of the research is to assess the sensitivity level of the information that is disseminated from the database based on the analysts‟ queries. The issue of utility-based privacy controlling data mining was reviewed in [11]. In [12], a technique for the suppression of anonymization of data. Disclosure top-down does not require a tree of taxonomy. The process begins with a set of deleted records and identifies the best specific candidate value that satisfies the privacy constraint for disclosure. The multidimensional k-anonymity is a multidimensional QID global recoding technique. In order to determine the optimum generalization, they consider Discernibility metric and Equivalence Class Size metric parameters. Multidimensional partitioning compared to single-dimensional partitioning to achieve the generalization error rate. The principle of t-closeness is that the distribution of sensitive values is as close to the distribution of sensitive values in the original data set in each equivalence class.

Support vector machine is an optimization technique for solving a variety of approaches such as classification, learning and outlier problems. The basic support vector machine (SVM) solves the two class problems, in which the data are partitioned by a hyper-plane using support vectors. If the support vector machine fails to separate two classes, then it solves this problem using a kernel function. Various kernel functions can be used in the SVM model such as linear, polynomial, Gaussian, regression, etc. to preserve the privacy on training dataset [13]. The author in [14] studied the utility-based problem of PPDM on large dataset. The idea was to extend the cursed dimensionality by distributing disjointed

matrices covering efficient attributes (Utility), but it is also challenging for privacy to be preserved. In Xu et al. proposed the use of local utility-based data mining method. The method is based on the fact that different attributes have varied utility from a software point of view. In local data partitioning, the data space is separated into many areas and the instance plotting to generalize value is local to that area. Another alternative way of using utility-based PPDM to anonymize data is that its residues beneficial to specific types of knowledge discovery process. This form of approach is frequently modelled with the k-anonymity framework and its derivatives: l-diversity, t-closeness, etc. Another popular model of privacy is that of π-differential privacy, which ensures that the addition or removal of data from a dataset results in a maximum change in any published information relative to π [15]. This ensures that a particular individual's presence or absence in the dataset has a limited impact on the information released, thus protecting the privacy of each individual. Data will be located and maintained at a single location in a centralized database, where data can be distributed vertically or horizontally to different sources, as in the distributed database. All data is stored in one place when the database is centralized. This database type is entirely different from the distributed database. One of the issues facing the centralized database is that since the entire data is located at one central location, bottle-neck problems can occur at key points where the data is released or published. As a result, the efficiency with which it is retrieved when searching for data availability is not as strong as in the distributed database system. Some of the traditional approaches, including k-anonymity,'-diversity, t-closeness and incognito, provide solutions to the problem of disclosure. They introduced a solution, namely, k-anonymity, which is considered a standard approach to dealing with the problem of linking attack. The anonymization-based study to protect individual privacy has become popular for the past decade. They conducted [16] a survey of U.S. census summary data to state the privacy risk of individuals.

III. Proposed Geometric Perturbation based Privacy Preserving Classifier

In this proposed an advanced privacy preserving classification model is designed and implemented on the various datasets. Initially, each input data is pre-processing using the novel data filtering method. This transformation method is used to transform the numerical and nominal values and to fill the sparsity values on large datasets. After the data pre-processing step, a hybrid geometric perturbation method is developed to improve the classification rate on the filtered data. Finally, a novel boosting classification model is applied on the perturbation data for privacy preserving as shown in Fig. 1.

In this work, a hybrid data filtering method is designed and implemented on each PPDM input dataset. In the proposed data filtering method, each numerical attribute is normalized using the hybrid data transformation equation.

Algorithm 1: Privacy preserving based data filter (PPDF)

Input: PPDatasets PD={D1,D2…Dn}, Attributes List: AL ,Max attribute value M_x, Minimum attribute value M_n.

Maximum attribute value:Mx(A),Minimum Attriubte value :Mn(A), Mean of the attribute:μ_A, Standard deviation of the attribute:σ_A.

1. Read input PPDM datasets D
2. To each dataset PD_i
3. Do
4. To each attribute $PD_{A[j]}$
5. Do
6. For each attribute value of $PDV_{A[j][k]}$
7. Do
8. If ($PD_{A[j]}$ is numerical attribute and NOTNULL)
9. Then
10. Transform $PDA_{[j]}$ using the following eq .(1)
11.

$$PDV_{A[j][k]} = \frac{|PDV_{A[j][k]} - \max\{\mu_{PD_{A[j]}}, (M_x(PD_{A[j]}) - M_n(PD_{A[j]}))/\sigma_{PD_{A[j]}})|}{2*(M_x(PD_{A[j]}) - M_n(PD_{A[j]}))} \quad \text{--(1)}$$

12. End if
13. If ($PD_{A[j]}$ is nominal && $PD_{A[j]}$ is not null)
14. Then
15. Replace $PDV_{A[j][k]}$ using the eq.(2)
16.

$$PDV_{A[j][k]} = \frac{\sum_{i=j,m=c} Pr(PD_{A[j]}/c_m).\max\{Pr(c_m)\}}{|c|.\sigma_X.\min\{Pr(PD_{A[j]}/c_m)\}} \quad \text{-----(2)}$$

$i = 1..n; m = 1..c(\#classes)$

17 Done

18 Done

19 Done

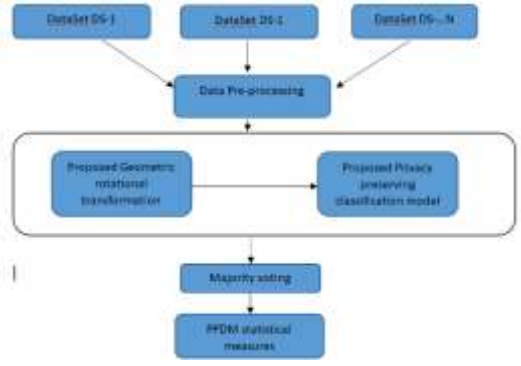

Fig. 1. Proposed Ensemble Deep Learning Framework for Privacy Preserving.

In the algorithm 1, each attribute of the input privacy preserving dataset is taken as input and transform the value using the equation 1 and 2. Initially, each attribute is tested for numerical data type or nominal type. If the attribute is numerical and it is not empty then each value in the attribute is transformed to new value by using eq.1. Similarly, if the attribute is nominal type then each value is estimated by using the maximization and minimization of its class probabilities.

Algorithm 2: Geometric Homo Perturbation (Attribute A, Value V):

1: Input: Transformed Sensitive attributes list S (AL).

2: Parameter initialization for homomorphic based attribute perturbation.

In the homomorphic based geometrical transformation, each attribute is perturbated using the additive and multiplicative transformation as given below.

$: h \to h_0 = \begin{pmatrix} h_1 & h_2 \\ h_2 & h_1 \end{pmatrix}$, where $h = h_1 + h_2$

(1) The addition homomorphism holds since

$$h_0 + \bar{h}_0 = \begin{pmatrix} h_1 & h_2 \\ h_2 & h_1 \end{pmatrix} + \begin{pmatrix} \bar{h}_1 & \bar{h}_2 \\ \bar{h}_2 & \bar{h}_1 \end{pmatrix} = \begin{pmatrix} h_1 + \bar{h}_1 & h_2 + \bar{h}_2 \\ h_2 + \bar{h}_2 & h_1 + \bar{h}_1 \end{pmatrix}$$

(2) The subtraction homomorphism holds since

$$h_0 - \bar{h}_0 = \begin{pmatrix} h_1 & h_2 \\ h_2 & h_1 \end{pmatrix} - \begin{pmatrix} \bar{h}_1 & \bar{h}_2 \\ \bar{h}_2 & \bar{h}_1 \end{pmatrix} = \begin{pmatrix} h_1 - \bar{h}_1 & h_2 - \bar{h}_2 \\ h_2 - \bar{h}_2 & h_1 - \bar{h}_2 \end{pmatrix}$$

Then, $h + \bar{h} = h_1$

(3) The multiplication $+ h_2 - (\bar{h}_1 + \bar{h}_2) = (h_1 - \bar{h}_1) + (h_2 - \bar{h}_2) = (h_0 - \bar{h}_0)_{11} + (h_0 - \bar{h}_0)_{12}$, Hence, Dec $(C^*_\Theta) = h - \bar{h}$.

$$h_0 \cdot \bar{h}_0 = \begin{pmatrix} h_1 & h_2 \\ h_2 & h_1 \end{pmatrix} \cdot \begin{pmatrix} \bar{h}_1 & \bar{h}_2 \\ \bar{h}_2 & \bar{h}_1 \end{pmatrix}$$

Here, $h_1, \bar{h}_1 \in F$ are pairwise commutative. Furthermore, $h = h_1 + h_2, \bar{h} = \bar{h}_1 + \bar{h}_2$ and $m \cdot \bar{h} = (h_1\bar{h}_1 + h_2\bar{h}_2)$ $(h_1\bar{h}_2 + h_2\bar{h}_1) = (h_0 \cdot \bar{h}_0)_{11} + (h_0 \cdot \bar{h}_0)_{12}$. Hence, Dec $(C^*_\varnothing) = h\bar{h}$

The division homomorphism holds since

$$\bar{h}_0^{-1} = \begin{pmatrix} \bar{h}_1 & \bar{h}_2 \\ \bar{h}_2 & \bar{h}_1 \end{pmatrix}^{-1} = \frac{1}{(\bar{h}_1 - \bar{h}_2) \cdot (\bar{h}_1 + \bar{h}_2)} \cdot \begin{pmatrix} \bar{h}_1 & -\bar{h}_2 \\ -\bar{h}_2 & \bar{h}_1 \end{pmatrix}$$

and $\bar{h}_1 + \bar{h}_2 = \bar{h}$ for $\bar{h}_1 \neq \bar{h}_2$. Therefore,

$$(\bar{h}_0^{-1})_{11} + (\bar{h}_0^{-1})_{12} = \frac{1}{(\bar{h}_1 - \bar{h}_2) \cdot (\bar{h}_1 + \bar{h}_2)} \cdot \bar{h}_1 + \frac{1}{(\bar{h}_1 - \bar{h}_2) \cdot (\bar{h}_1 + \bar{h}_2)} \cdot (-\bar{h}_2)$$

$$= \frac{\bar{h}_1 - \bar{h}_2}{(\bar{h}_1 - \bar{h}_2) \cdot (\bar{h}_1 + \bar{h}_2)} = \frac{1}{\bar{h}_1 + \bar{h}_2} = \frac{1}{\bar{h}}$$

In the geometrical homomorphic perturbation, two keys are generated to each communication parties for data sharing and data re-construction process. The two keys public key and private keys are generated using the non-linear cyclic group elements.

Choose two cyclic group elements with prime orders k1,k2.

μ_A = mean

η_A = VAR(AL);//Variance

$$h_p = \text{gdf}(AL) = \frac{\eta_A{}^{\alpha} x^{\alpha-1} e^{-\eta_A x}}{\Gamma(\alpha)}, \quad \text{for } \eta_A, \alpha > 0$$

$$h_q = \log\left(\frac{\eta_A e^{-\eta_A (ALV[i]-\tau)}}{(1+e^{-\eta_A (AL[i]-\tau)})^2} * \text{gdf}(AL).\text{mean}\right)$$

n= h_p * h_q;

s=n*n;

Choose a random noise $r_n \in (0,1)$

$$\psi = \frac{h_p . h_q}{(n^{(h_p)} \bmod(r_n))^{(h_q)} \bmod(r_n)}$$

$$\theta = \psi_A{}^{\gcd(h_p . h_q, r_n)}$$

Step 3: Geometric attribute perturbation is given as

GP[]=E(PB[])=

$$r_n{}^{ALV[i]} \bmod(s) . \theta^n \bmod(s) . \bmod(s)$$

Step 4: Geometric data re-construction process is given as D(CB[])

$$h_1 = r_n^{\theta} \bmod(s) - (\frac{\theta}{n})^{-1} \bmod(n)$$

$$\text{PB[]} = \text{GP[i]}^{\theta} \bmod(s) - \frac{\theta}{n} . h_1 \bmod(n)$$

Algorithm 3: Boosting Privacy Preserving Classification model

In this algorithm, a hybrid privacy preserving based classification model is designed and implemented on the input datasets. This algorithm is used to check the performance of the privacy preserving model on the geometric perturbation data and the original data. Here, multiple boosting classifiers are integrated to improve the voting rate of the overall classification model. In this proposed classification model, a novel random tree and non-linear kernel function based multi-class SVM approach. In the boosting classification model KNN, random tree and non-linear kernel based SVM are used to improve the overall accuracy on the perturbation data.

Algorithm: Boosting Privacy Preserving Classification model

Random Tree

1. To each input dataset PD
2. Do
3. Partition data into k number of classes and compute the best feature ranking measure using the following measure.

4. To each partition PFD
5. Do
6. For each attribute $FD(A_i)$ in PFD
7. Do
8. $$\text{RandomTree Ranking Measure=RTRM}[FD(A_i),k]= \frac{-Prob(C_k).\sum \log(FD(A_i))*Prob(FD(A_i)/C_k)}{FD(A_i).\sqrt{Entropy(FD(A_i))}}$$
9. End for
10. Done
11. Done

Non-linear SVM

Apply SVM multi-class optimization models as

$$\min_{W_k, a_k} \frac{1}{2} \| W_k \|_1^2 + \ker < v, m > . \sum_{i=1}^{n} \exp(W_k) + \eta$$

s.t

$$W_k^T D_i + b_k \geq 1 - \eta, \text{if } m_i = k$$

$$W_k^T D_i + b_k \leq -1 + \eta, \text{if } m_i \neq k$$

$$\eta > 0; m = 1 \ldots \text{classes}$$

In the above multi-objective function, a new kernel function is defined to improve the performance of the privacy preserving classification model. Here kernel function ker(x,y) defines the v input values that are mapped to m dimensional space as:

$$F_0(v) = 1$$

$$F_1(v) = v$$

$$F_{k+1}(v) = 2vT_k(v) - T_{k-1}(v)$$

$$Ker(v, m) = \min\{\frac{\sum_{i=0}^{n} T_i(v)T_i^T(m)}{\sqrt{a - vm}}, \frac{\sum_{i=0}^{n} T_i(x_j)T_i(m_j)}{\sqrt{1 - v_j m_j}}\} \times \prod_{i=1}^{m} (\cos(2.\frac{v_j - m_j}{a}) \exp(-\frac{\| v - m \|^2}{2a^2})$$

To each pattern in the decision tree construction, rule type is considered as either left side or right side of the pattern for privacy preserving.

IV. Experimental Results

Experimental results are carried out in java environment with multiple privacy preserving datasets. In this experimental results, proposed privacy preserving model is simulated on original datasets and transformed datasets. Different statistical measures such as accuracy, recall, precision and runtime are computed on the different datasets. These statistical metrics are used to check the performance of the privacy preserving based model on the perturbation dataset. Here, all the sensitive features are perturbated in order to preserve the privacy on machine learning decision patterns. Experimental results are compared on different privacy preserving models such as geometric perturbation, rotational perturbation and PABIDOT. Our models is tested on different datasets such as FRDS, WQDS, ELDS, LRDS taken from [16].

The proposed algorithm is applied on Bank Marketing dataset from UCI repository. The dataset contains 17 attributes and 45211 rows along with other datasets. The attributes in bank dataset are age(numeric), job(categorical), marital status(categorical), education(categorical), credit default(categorical), housing loan(categorical), personal loan(categorical), contact communication type (categorical), last contact month(categorical), contact day of month(categorical), duration (numeric), campaign (numeric), pdays (numeric), previous(numeric), pout come (categorical), client subscribed to term deposit(yes or no)(categorical).

Among the attributes of bank marketing dataset, 'client subscribed to term deposit' attribute is sensitive attribute. There are no identifier attributes to be removed from given dataset. Attributes age, job, marital status and education are considered as quasi identifiers. Age is numerical quasi identifier and job, marital status, education are categorical quasi identifiers. Various utility measurements are used to measure the usefulness of generalized data. Some are loss metrics, ambiguity metrics, differentiation in discernibility, KL, entropy-based loss of information, and so on. In this work, traditional model PABIDOT and other perturbation models are used to compare the proposed model on the input training data. These traditional models have issues on high dimensionality and sparsity problems.

Metric Loss (LM): LM is calculated for each tuple attribute. The value t[A] is widespread tox where t is tuple and A is categorical. Suppose the domain size of A is and M is the number of values generalized tox and the value of t[A] (M-1)/(A-1) is lost. Loss of attributes is calculated for all tuple t as average loss t[A]. LM is the sum of losses for each attribute for the dataset.

Discernability metric (DM): Each tuple in the database has a penalty based on the number of other tuples that cannot be distinguished from it. For a size n database, DM assigns n to each deleted tuple as a penalty. Penalty shall be the total number of tuples with the same quasi-identifier values for unrestrained tuples. Thus, if tuples are grouped by a quasi-identifier, the DM shall be defined as the total number of squared groups plus n times the number of deletes.

Metric ambiguity: This metric is highly suitable for the k-anonymity framework. AM calculates the number of tuples for every tuple t, generalized to tuple t*, in the sanitized data domain. This is the ambiguity of t*. The AM for sanitized data is an average ambiguity of all tuples in the sanitized dataset.

KL-Divergence: The original table is treated as a distribution probability p1 to use KL-divergence. P1(t) is the tuple fraction equal to t. The sanitized data will also be converted to p2 (possible ways to do this will be discussed). The KL-divergence among the two is the same as for p1(t) log (p1(t)/p2(t)).

Table I illustrates the performance of the present proposed hybrid perturbation-based privacy preserving model to the traditional models on different training datasets. Here, the average of F-measure is computed on the training datasets. As

shown in the table, it is noted that the proposed geometric perturbation based boosting classifier has better F-measure than the traditional models.

Fig. 2 illustrates the performance of the present proposed hybrid perturbation-based privacy preserving model to the traditional models on different training datasets. Here, the average of recall measure is computed on the training datasets. As shown in the figure, it is noted that the proposed geometric perturbation based boosting classifier has better recall measure than the traditional models.

Fig. 3 illustrates the performance of the present proposed hybrid perturbation-based privacy preserving model to the traditional models on different training datasets. Here, the average of precision measure is computed on the training datasets. As shown in the figure, it is noted that the proposed geometric perturbation based boosting classifier has better precision measure than the traditional models.

Fig. 4 illustrates the performance of the present proposed hybrid perturbation-based privacy preserving model to the traditional models on different training datasets. Here, the average of accuracy measure is computed on the training datasets. As shown in the figure, it is noted that the proposed geometric perturbation based boosting classifier has better accuracy measure than the traditional models.

TABLE I. PERFORMANCE ANALYSIS OF HYBRID PERTURBATION BASED PRIVACY PRESERVING MODEL TO THE CONVENTIONAL MODELS USING STATISTICAL F-MEASURE

TestData	GP	RP	PABIDOT	ProposedGPBC
Test(#1)	68.31	69.96	79	85.11
Test(#2)	69.77	69.2	79.41	87.11
Test(#3)	70.05	69.03	81.87	90.9
Test(#4)	69.25	69.17	79.79	91.03
Test(#5)	69.54	70.99	80.07	91.81
Test(#6)	68.56	75.18	79.03	88.47
Test(#7)	69.47	70.67	78.16	91.96
Test(#8)	67.57	75.9	80.52	89.43
Test(#9)	70.61	75.96	81.24	88.22
Test(#10)	67.63	73.58	81.6	87.55

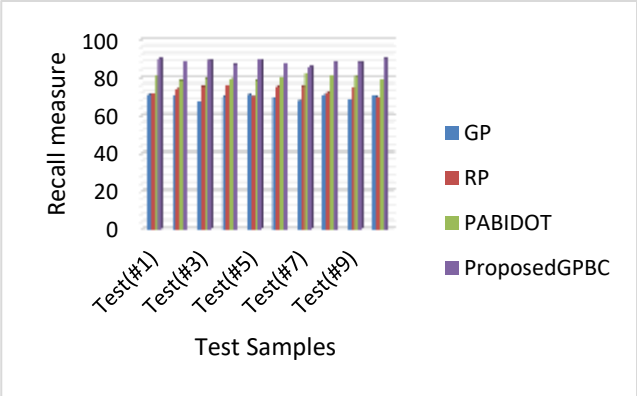

Fig. 2. Performance Analysis of Hybrid Perturbation based Privacy Preserving Model to the Conventional Models using Statistical Recall measure.

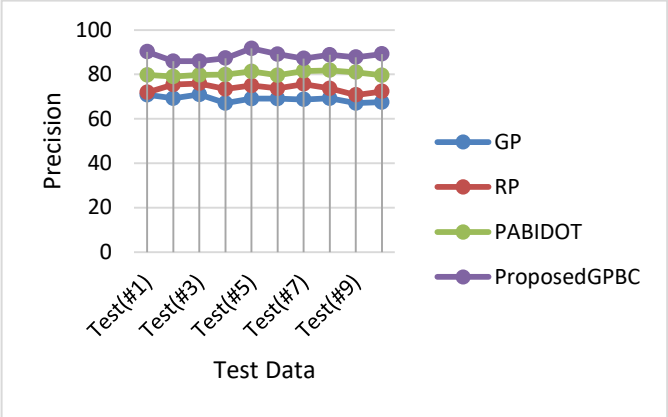

Fig. 3. Performance Analysis of Hybrid Perturbation based Privacy Preserving Model to the Conventional Models using Statistical Precision.

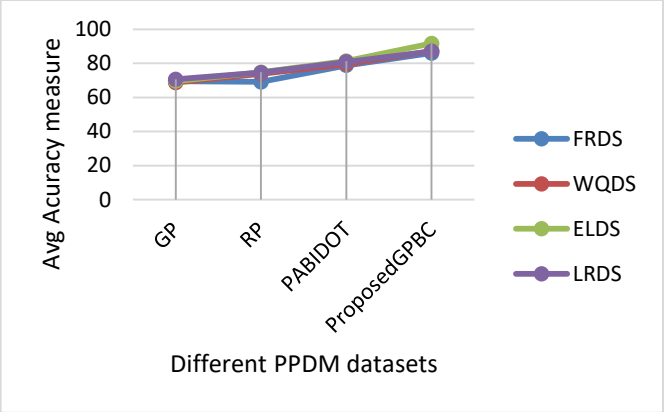

Fig. 4. Performance Analysis of Hybrid Perturbation based Privacy Preserving Model to the Conventional Models using Statistical Accuracy on different Datasets.

Fig. 5 illustrates the performance of the present proposed hybrid perturbation-based privacy preserving model to the traditional models on different training datasets. Here, the average of error rate measure is computed on the training datasets. As shown in the figure, it is noted that the proposed geometric perturbation based boosting classifier has better error rate than the traditional models.

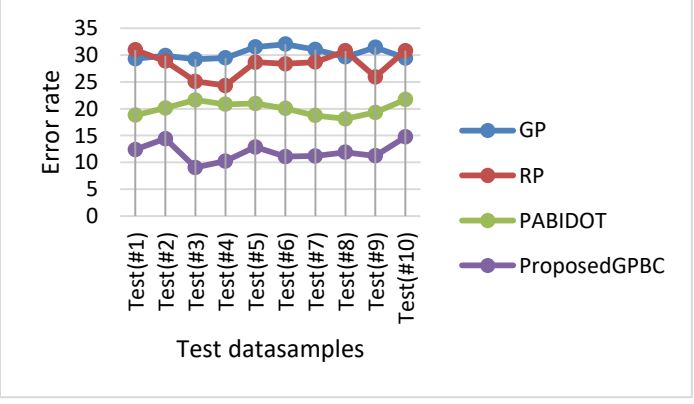

Fig. 5. Performance Analysis of Hybrid Perturbation based Privacy Preserving Model to the Conventional Models using Statistical Error Rate on different Datasets.

TABLE II. PERFORMANCE ANALYSIS OF HYBRID PERTURBATION BASED PRIVACY PRESERVING MODEL TO THE CONVENTIONAL MODELS USING STATISTICAL RUNTIME (MS) ON DIFFERENT DATASETS

TestData	GP	RP	PABIDOT	ProposedGPBC
Test(#1)	4203	4853	3947	3138
Test(#2)	4217	4376	3707	3052
Test(#3)	4440	4183	4147	3106
Test(#4)	4265	4792	4608	3247
Test(#5)	3903	5164	4008	2894
Test(#6)	3622	3826	5204	2910
Test(#7)	4335	4257	4042	2921
Test(#8)	4118	4633	5007	2880
Test(#9)	4437	4906	3706	3092
Test(#10)	4000	4814	3691	2962

Table II illustrates the performance of the present proposed hybrid perturbation-based privacy preserving model to the traditional models on different training datasets. Here, the average of runtime (ms) measure is computed on the training datasets. As shown in the table, it is noted that the proposed geometric perturbation based boosting classifier has better runtime (ms) measure than the traditional models.

A. Results Analysis

A new privacy preserving data mining method is proposed. The proposed method is applied on various data sets and results were observed. The proposed method retains the classification accuracy while balancing data utility. Traditional approaches are limited to fixed sensitive attributes for privacy preserving. Also, these models are not appropriate on large data size. Also, the experimental results simulated on the perturbation anonymization bank data were improved by nearly 2% than the original data and nearly over 1% on the perturbation bank data. Experimental results suggested that the proposed geometric perturbation model achieves better efficiency in terms of high dimensionality and large data size than the conventional models.

V. CONCLUSION

In this work, a novel filtered based privacy preserving model is designed and implemented on the different datasets. Since, most of the conventional privacy preserving models are depend on the data size and number of features, it is difficult to provide the privacy to a large number of attributes due to computational time and accuracy. Also, it is essential to preserve the privacy of the large sensitive attributes before publishing it to the third-party servers. As a result, a novel framework is required to improve the privacy as well as accuracy on the high dimensional privacy preserving data with less runtime. In this work, a filter-based hybrid privacy preserving model is designed and implemented on the different complex datasets in order to optimize the privacy preserving accuracy and the runtime. Experimental results proved that the proposed privacy preserving model has better efficiency on the different domain datasets compared to the conventional models. In the future work, this work can be extended to a cryptographic based perturbation method for big datasets in order to minimize the error rate and to improve the privacy preserving policies.

REFERENCES

[1] R. M. Alguliyev, R. M. Aliguliyev, and F. J. Abdullayeva, “Privacy-preserving deep learning algorithm for big personal data analysis,” Journal of Industrial Information Integration, vol. 15, pp. 1–14, Sep. 2019, doi: 10.1016/j.jii.2019.07.002.

[2] M. Amiri-Zarandi, R. A. Dara, and E. Fraser, “A survey of machine learning-based solutions to protect privacy in the Internet of Things,” Computers & Security, vol. 96, p. 101921, Sep. 2020, doi: 10.1016/j.cose.2020.101921.

[3] A.Boulemtafes, A. Derhab, and Y. Challal, “A review of privacy-preserving techniques for deep learning,” Neurocomputing, vol. 384, pp. 21–45, Apr. 2020, doi: 10.1016/j.neucom.2019.11.041.

[4] M. A. P. Chamikara, P. Bertok, I. Khalil, D. Liu, and S. Camtepe, “Privacy Preserving Face Recognition Utilizing Differential Privacy,” Computers & Security, vol. 97, p. 101951, Oct. 2020, doi: 10.1016/j.cose.2020.101951.

[5] Y. Chen, F. Luo, T. Li, T. Xiang, Z. Liu, and J. Li, “A training-integrity privacy-preserving federated learning scheme with trusted execution environment,” Information Sciences, vol. 522, pp. 69–79, Jun. 2020, doi: 10.1016/j.ins.2020.02.037.

[6] Y. Dong, X. Chen, L. Shen, and D. Wang, “EaSTFLy: Efficient and secure ternary federated learning,” Computers & Security, vol. 94, p. 101824, Jul. 2020, doi: 10.1016/j.cose.2020.101824.

[7] J. Duan, J. Zhou, and Y. Li, “Privacy-Preserving distributed deep learning based on secret sharing,” Information Sciences, vol. 527, pp. 108–127, Jul. 2020, doi: 10.1016/j.ins.2020.03.074.

[8] Y. Fan et al., “Privacy preserving based logistic regression on big data,” Journal of Network and Computer Applications, p. 102769, Aug. 2020, doi: 10.1016/j.jnca.2020.102769.

[9] C. Fang, Y. Guo, N. Wang, and A. Ju, “Highly efficient federated learning with strong privacy preservation in cloud computing,” Computers & Security, vol. 96, p. 101889, Sep. 2020, doi: 10.1016/j.cose.2020.101889.

[10] M. Gong, J. Feng, and Y. Xie, “Privacy-enhanced multi-party deep learning,” Neural Networks, vol. 121, pp. 484–496, Jan. 2020, doi: 10.1016/j.neunet.2019.10.001.

[11] M. Gong, K. Pan, Y. Xie, A. K. Qin, and Z. Tang, “Preserving differential privacy in deep neural networks with relevance-based adaptive noise imposition,” Neural Networks, vol. 125, pp. 131–141, May 2020, doi: 10.1016/j.neunet.2020.02.001.

[12] Z. Guan, Z. Lv, X. Du, L. Wu, and M. Guizani, “Achieving data utility-privacy tradeoff in Internet of Medical Things: A machine learning approach,” Future Generation Computer Systems, vol. 98, pp. 60–68, Sep. 2019, doi: 10.1016/j.future.2019.01.058.

[13] P. Li, T. Li, H. Ye, J. Li, X. Chen, and Y. Xiang, “Privacy-preserving machine learning with multiple data providers,” Future Generation Computer Systems, vol. 87, pp. 341–350, Oct. 2018, doi: 10.1016/j.future.2018.04.076.

[14] X. Li, Y. Gu, N. Dvornek, L. H. Staib, P. Ventola, and J. S. Duncan, “Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results,” Medical Image Analysis, vol. 65, p. 101765, Oct. 2020, doi: 10.1016/j.media.2020.101765.

[15] Y. Liu, Z. Ma, Z. Yan, Z. Wang, X. Liu, and J. Ma, “Privacy-preserving federated k-means for proactive caching in next generation cellular networks,” Information Sciences, vol. 521, pp. 14–31, Jun. 2020, doi: 10.1016/j.ins.2020.02.042.

[16] M. A. P. Chamikara, P. Bertok, D. Liu, S. Camtepe, and I. Khalil, “Efficient privacy preservation of big data for accurate data mining,” Information Sciences, vol. 527, pp. 420–443, Jul. 2020, doi: 10.1016/j.ins.2019.05.053.

AUTHORS' PROFILE

Shaik Mohammed Gouse research scholar. He obtained his Bachelors degree in Electronics From Acharya Nagarjuna University, M.C.A degree from Madurai Kamaraj U niversity, M.Tech (CSE) from Jawaharlal Nehru Technological University, Kakinada, He is currently pursuing Ph.D (CSE) degree with Department of Computer Science and Engineering Koneru Lakshmaiah Education Foundation, Vaddeswaram, 522502 Andhra Pradesh, India. His research interests lie in Bigdata Analytics, AI and Data science.

Dr. G. Krishna Mohan, working as Professor in the Department of Computer Science & Engineering, KL University. He obtained his M.C.A degree from Acharya Nagarjuna University, M.Tech(CSE) from Jawaharlal Nehru Technological University, Kakinada, Ph.D(CSE) from Acharya Nagarjuna Universi ty. Qualified, AP State Level Eligibility Test. His research interests lie in Data Mining and Software Engineering. He published 26 research papers in SCOPUS indexed, 45 research papers in various National and International journals. Editorial board member of SAS Publishers (Scholars Academic & Scientific Publishers) & SCIREA journal of computer. Reviewer of Australasian Journal of Information Systems & International Journal of Engineering &Technology. Authored three book chapters in Springer.