International Journal of Pure and Applied Mathematics
Volume 119 No. 14 2018, 1863-1870
ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
Special Issue

ijpam.eu

A COMBINED APPROACH FOR PRIVACY PRESERVING CLASSIFICATION MINING

NagaPrasanthi Kundeti[1] , Chandra Sekhara Rao MVP[2]
[1]PhD Scholar, Department of CSE, Acharya Nagarjuna University, Guntur.
[2]Professor, Department of CSE, RVR & JC college of Engineering, Guntur.

ABSTRACT: In recent years, the growing capacity of information storage devices has led to increased storing personal information about customers and individuals for various purposes. Data mining needs extensive amount of data to do analysis for finding out patterns and other information which could be helpful for business growth, tracking health data, improving services, etc. This information can be misused for many reasons like identity theft, fake credit/debit card transactions, etc. To avoid these situations, data mining techniques that secure privacy of the data are proposed. Data Perturbation, Knowledge Hiding, Secure Multiparty computation and privacy aware knowledge sharing are some of the techniques of privacy preserving data mining. A combination of these approaches is applied to achieve better privacy. In this paper, geometric data perturbation, and k-anonymization techniques are discussed and also observed that data mining results after perturbation and anonymization are not changed much.

Keywords- Data Mining, Privacy preserving data mining, data perturbation, k-anonymization

1.1 INTRODUCTION

Data play an important role now a days in extracting knowledge. From decades, companies running software systems are flooded with lot of data which is of no use to them. Through Data Mining those large volumes of data can be processed and useful patterns can be identified. These patterns help managers to take decisions to improve their businesses. However, the collected information may contain some sensitive information which raises privacy concern and privacy does not has a benchmark definition[9].

Westin[1] defined privacy as "the assertion of individuals, groups or institutions to specify when, how and to what extent their information can be shared to others". Bertino[7] et.al. gave a similar definition as " the security of data about an individual contained in an electronic repository from unauthorized disclosure." Privacy preservation methods protect from information leakage by modifying the original data and protect owner's exposure[8],[12] but utility of data is reduced by data transformation. This data transformation results in inaccurate or infeasible knowledge extraction through data mining. Privacy preserving data mining (PPDM) methodologies equip with certain level of privacy, while not compromising data utility and still provide efficient data mining. PPDM consists of various techniques that preserve privacy while extracting knowledge from data. While carrying out data mining, there is a chance for private data that may be disclosed to the public. PPDM protects from this disclosure. PPDM is latest area of research and it has many algorithms. Different techniques preserve privacy in different stages of data mining process. There are three layers in PPDM framework namely Data Collection Layer(DCL), Data Pre-process Layer(DPL) and Data Mining Layer(DML).

1.2 PPDM FRAMEWORK

According to the PPDM framework defined by XueyunLi et.al[11] the PPDM techniques are categorized based on data mining process stages. DCL has huge collection of data providers and sensitive information may be part of this data. Data can be collected without losing privacy. In DPL, the data that is collected in DCL layer is stored in data warehouses and later processed by data warehouse servers. There are two aspects of privacy preservation in this layer. (i)data pre processing in such a way that privacy is preserved for doing data mining later and (ii)security of data access. Actual data mining is performed by data mining servers and data miners and results are provided in third layer. There are two aspects for privacy preservation in this layer. They are (a) incorporating privacy features into data mining methods,(b)combining lot of data sets from different parties and carrying out collaborative data mining without any private information revelation.

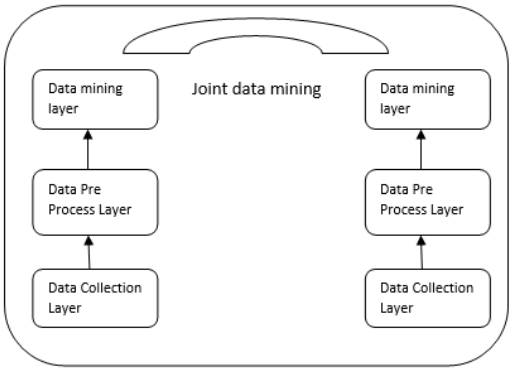


Fig1: A PPDM Framework

Privacy at Data Collection Layer:
In order to provide privacy at data collection time, raw data need to be randomized and stored. If original values are stored there is a chance of privacy leakage. So, randomization is performed for each value separately. According to statistical distribution, noise is calculated and added to data to modify data in randomization methods. Simple randomization approach is described as: if X is original data distribution, Y is noise distribution already known and Z is result of randomization then definition of Z can be given as

$$Z=X+Y \qquad (1)$$

Later X is constructed as X=Z-Y. We can not reconstruct entire X as it is but only X distribution can be reconstructed. This is known as additive noise. There is another way of randomization that is perturbing the data i.e. modify original data into perturbed data. Data mining algorithms which are based on distribution of data rather than individual values are used. But there is loss of data readability. When compared to the privacy preservation, the data readability loss is negligible. So, this method is followed. Randomization is a subset of data perturbation.

Data Privacy: Data privacy is generally identified as a level of difficulty, an attacker has to face in approximately identifying the original data from the available perturbed data. PPDM Techniques are said to provide higher level of privacy if estimation of original data from perturbed data is more difficult. Geometric data perturbation provides moderate level of data privacy but is more efficient compared to other algorithms [10].

Data Utility: Based on quantity of important data that is preserved after perturbation the level of data utility is defined. In this paper we present the steps of geometric data perturbation based on[10]. Many data mining models can be applied with geometric data perturbation for privacy preservation and they also provide better utility.

Some of the data perturbation techniques are mentioned as following.

Noise Additive Perturbation: "It is an additive randomization which is a column based one. This kind of technique is based upon the two factors i.e. 1) Data owner does not require to secure all components in a record equally, this gives the freedom to apply column based distortion on some sensitive fields. 2) Individual records are not needed for Classification model. Chosen Classification models only require value distributions and assume that they are independent columns "[10]. The fundamental method adds the certain amount of noise to the columns, keeping the structure intact and can be easily recreated from bewildered data.

A classic random noise addition model is outlined as following. Let a variable K having some distributions, be described as (k1,k2,k3... kn). The random noise addition process changes its original value by adding some kind of noise R and generates perturbed value Y. Now Y will be K + R, resulting into(k1+r1,k2+r2,k3+r3 …kn+rn). Using this noise R, the original value K can be recovered by applying reconstruction algorithm on the perturbed values [10].
While this approach is simple, it has some cons. Several researchers have found that it is easy to perform reconstruction based attacks, which is major weakness of randomized noise addition approach. Also,resembling properties of the perturbed data can become handy to identify and remove noise from the perturbed data. Moreover, algorithms like association rule mining and decision tree are based on the autonomic columns assumption and work only on column distributions. These algorithms can be modified to reconstruct the column distributions from modified datasets [10].

Condensation-based algorithm: "This is a multi-dimensional data perturbation technique. This technique preserves the dispersion matrix for multiple columns. Decision boundary which is a geometric property is preserved well. This algorithm unlike the randomize approach, disturbs multiple columns at a

time and generates the entire new dataset. Because of above mentioned properties, modified data sets can be directly used in many existing data mining algorithms without any change or need to develop new algorithms" [10].

"The approach is outlined as follows. Algorithm begins by partitioning the original dataset D into number of groups of records, say k-record groups. There are two parts in each group. One is a center of the group, selected randomly from the original dataset and the other part is of (k-1) members from original dataset, found using k-1 nearest neighbours. These chosen k records are first deleted from the original dataset. Then the remaining groups are materialized. Advantage of having small locality of the group, it is achievable to revive k records set to maintain the covariance and distribution.

The size of the locality is reciprocal of the preservation of covariance with regenerated k records. If in each group, size of locality is smaller, then it offers better quality of covariance preservation for regenerated k records" [10].

Rotation perturbation: "For privacy preserving data clustering this technique is nominated. Geometric data perturbation has rotation perturbation as one of its major component. The definition of Rotation perturbation is given as G(X) = R*X where $X_{d\times n}$ is the original dataset and $R_{d\times d}$ is rotation matrix which is randomly generated. Distance preservation is the unique benefit as well as major weakness of this method. This method is vulnerable to distance-inference attacks"[10].

Random Projection Perturbation: "In this technique data points from original multi-dimensional space are projected into another arbitrarily chosen multidimensional space. Let $Q_{k\times d}$be a random projection matrix. Here, Q contains orthonormal rows.

$$G(x)=\sqrt{d}/k\mathrm{QX} \quad (2)$$

The above formula is administered to ruffle the dataset C. According to Johnson – Lindenstrauss Lemma, projection perturbation approximately preserves the distance. A given data set in Euclidean space can be mapped into another space. This mapping should preserve the pairwise distance of any two points with least error. This results in model quality preservation" [10].

Privacy at Data Pre-process Layer

Data Anonymization is the most prevalent method used for preserving privacy at DPL(data pre-process layer). This data anonymization(k-anonymization) prevents the identity disclosure of data owners in public[6]. The
k-anonymization technique works by specifying k-value so that there are k identical records in data.
In this, each record is identical to at least k-1 other records. Table1 shows an example data for a number of patients which is 4-anonymous. There are some distinctive attributes which identify a patient individually like age, country, disease, pincode. These attributes are categorized into two sets. They are attributes which are sensitive and non-sensitive attributes. Opponents should not be able to find these sensitive attributes ex.Ailment. Non-sensitive attributes like pincode, country and age are also called quasi-identifier attributes for the given data set.
According to L seweney's survey [3], we can not protect individual's privacy by simply eliminating explicit unique identifier .In the given table there are at least 4 records which have identical values for every set of quasi-identifier attributes. K-anonymization is often performed by data generalization and suppression.[2]

Table1 4-Anonymous Data Example

	Non-Sensitive Attributes			**Sensitive Attributes**
	PinCode	Age	Country	Ailment
1	210**	<=30	*	Flu
2	210**	<=30	*	Flu
3	210**	<=30	*	Cancer
4	210**	<=30	*	orthoritis
5	250**	>40	*	Flu
6	250**	>40	*	Cardiomyopathy
7	250**	>40	*	Cancer
8	250**	>40	*	Diabetes
9	313**	>=55	*	Cancer
10	313**	>=55	*	Diabetes
11	313**	>=55	*	orthoritis
12	313**	>=55	*	cancer

1.3GEOMETRIC DATA PERTURBATION

Translation transformation (Ψ),Multiplicative transformation(R) and distance perturbation (Δ) are applied in a specific sequence to obtain geometric data perturbation.

$$G(X)= RX+\Psi+\Delta \quad (3)$$

Multiplicative transformation (R): Generally rotation matrix or random projection matrix are part of this. Distances are preserved exactly by rotation matrix. Exact distances are preserved by rotation matrix. Only Approximate distances are preserved by random projection. Rotation matrix protects the Euclidean distance. One of the crucial component of geometric perturbation is rotation perturbation.

Rotation perturbation protects ruffled data from naive estimation attacks. Rotation perturbation can be protected from more complicated attacks by using other components of geometric perturbation. The definition for random projection matrix $R_{k\times d}$ is given as $R=sqrt((d/k)R_0)$ .The Johnson- Lindenstrauss Lemma state that approximate Euclidean distances can be preserved by random projection when certain conditions are satisfied. [10].

Translation transformation Ψ: In original space consider two points x and y, with translation the new distance will be || (x-t) – (y-t) || = | x – y ||. Therefore, distance is always preserved by translation. Translation perturbation alone can not furnish data protection.

Attacker can identify original data by cancelling translation perturbation if only it is applied alone. In order to resist attacks translation is combined with rotation perturbation.

Distance Perturbation: The distance relationship is preserved by above two components. However, distance-inference attacks can still be performed on distance preserving perturbation. The main aim of distance perturbation is to resist distance-inference attacks while preserving distances. Here, distance perturbation can be noise. As noise intensity is low, applying only other two components will not carry out privacy preservation.

The major issue of distance perturbation is a trade off between reduction of model accuracy and gain of privacy guarantee. The data owner may opt not to use distance perturbation with the assumption that data is secure and attacker does not know about the original data. Hence, distance-inference attacks are avoided.

The below graph will help summarizing about Random rotation, random projection and geometric data perturbation.

Table 2. Comparision Of Perturbation Techniques

Random Rotation	**Geometric Perturbation**	**Random Projection**
Y = R*X X is the original dataset for all three formulas Y is the perturbed dataset for all three formulas R is the random rotation matrix	Y = RX + T + D R is the secret rotation matrix (preserves Euclidean distances) T is the secret random translation matrix. D is the secret random noise matrix.	Y = A*X A is the random projection matrix.
Distances are preserved. Less secured [4].	Distances are approximately preserved[5].	Distance is not well preserved. Loss of Data [5].
Accuracy depends on the rotation Matrix	Good accuracy than any other perturbation techniques.	Worse accuracy than Geometric data perturbation

1.3.1Algorithm: Geometric Data Perturbation

The idea behind using Geometric Data Perturbation algorithm is its simplicity. Geometric perturbation is nothing but the improvement to the rotation perturbation by coupling it with additional components like random translation perturbation and noise addition to the basic form of multiplicative perturbation Y = R £ X. Two additional components are added to normal rotation perturbation.

When compared to normal rotation based perturbation, geometric perturbation is more robust and efficient.

For each attribute of G(X), let T be the translation, random rotation R, D be a Gaussian Noise and X be the original dataset. The value of the attribute G(X) can be found using following formula.

$$G(X) = R*X + T + D \quad (4)$$

Procedure: Geometric transformation based Multiplicative data perturbation

Input: Dataset D, Sensitive attribute S.

Output: Perturbed dataset D’

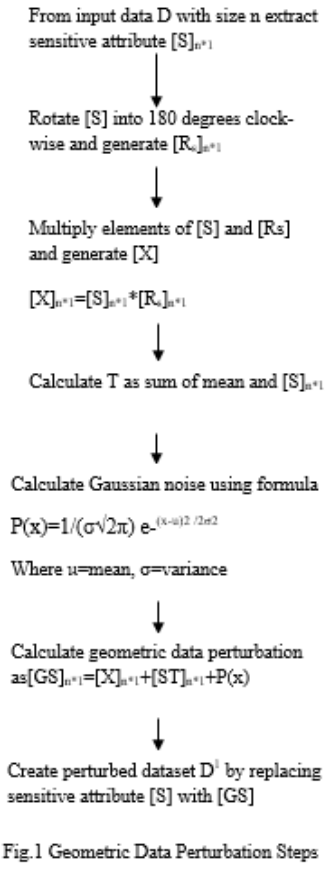


Fig.1 Geometric Data Perturbation Steps

Now apply classification algorithm on data set D with sensitive attribute S and obtain results. Apply classification algorithm on perturbed data set D^1 and obtain results. Compare both the results and analyze the accuracy.

1.4.K-ANONYMIZATION

In the perturbed data set, D^1, there are categorical attributes which can not be applied with geometric data perturbation. For those attributes we applied k-anonymization technique by generalizing the quasi identifiers wherever possible. The generalization hierarchy is followed for categorical attributes wherever necessary.

The generalization of categorical attributes can be obtained from the following hierarchical trees. In this the Adult data set from UCI machine learning repository is used for implementation.

The hierarchy tree for native-country is

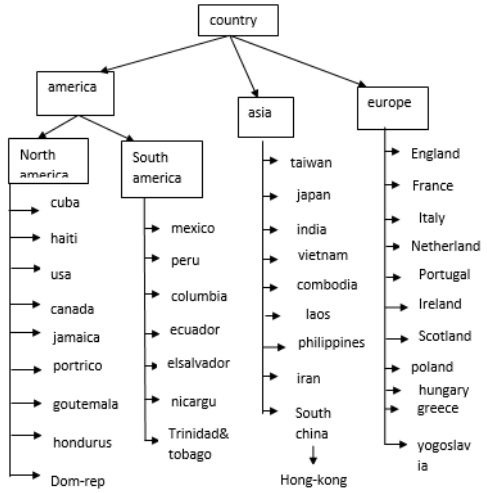


Fig2. Generalization Tree For Attribute Country

Similarly the hierarchy tree for education and work class is shown below.

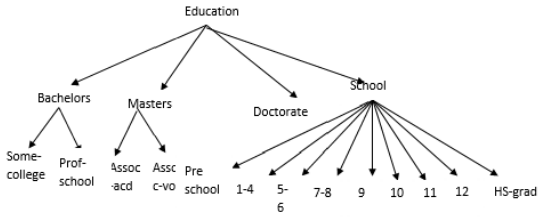


Fig3. Generalization Tree For Attribute Education

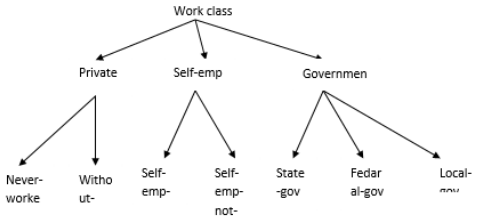


Fig4. Generalization Tree For Attribute Workclass

1.5 IMPLEMENTATION

In this paper, *Adult* data set from UCI Machine learning repository[13] is considered for implementation. The data set contained 48,842 instances with 15 attributes. The Implementation part is carried out in two steps namely Step1: Applying geometric data perturbation and perform classification. Step2:Applying K-anonymization after perturbation and perform classification.

The geometric random perturbation technique can be applied on only numeric data. So this method is applied on attributes age and education-num. The data set is described as below. The adult data set contains the following attributes Age(Numeric), Fnlwgt(Numeric), Work class(Text), Education(Text), Education num(Numeric), Marital Status(Text), Occupation(Text), Relationship(Text), Race(Text), Sex(Text), Capital gain(Numeric), Hours per week(Numeric), Native country(Text), Capital loss(Numeric), Class label(Text).

Results obtained:

WEKA tool is used for carrying out classification task on Adult data set. Two classification algorithms,J48 and Naive Bayes are applied on the original *Adult* data set from UCI machine learning repository[13]. Result characteristics like classification accuracy, Mean absolute error etc. are tabulated. The geometric data perturbation algorithm is implemented using MATLAB. The numerical attributes Age and Education num are modified by applying geometric data perturbation technique and the modified data set is obtained. On the modified data set, the two classification algorithms,J48 and Naive Bayes are applied and the results are tabulated.

Table 4: Classification Results After Geometric Data Perturbation And K-Anonymization Using Naive Bayes

	Naive Bayes				
	Original		Preprocessed		
	Original	Perturbed	Preprocessed	Perturbed	K-anony-mized
Correctly classified instances	0.8359	0.8363	0.8302	0.8347	0.8269
Incorrectly classified instances	0.1640	0.1636	0.1697	0.1652	0.1730
Kappa Statistics	0.4979	0.4897	0.5011	0.4997	0.4935
Mean Absolute error	0.1716	0.1711	0.1793	0.1731	0.1802
Root Mean Squared error	0.3692	0.3711	0.3798	0.3708	0.3815
Relative absolute error	0.4707	0.4718	0.4769	0.4679	0.4792

Table 5: Classification Results After Geometric Data Perturbation And K-Anonymization Using J48

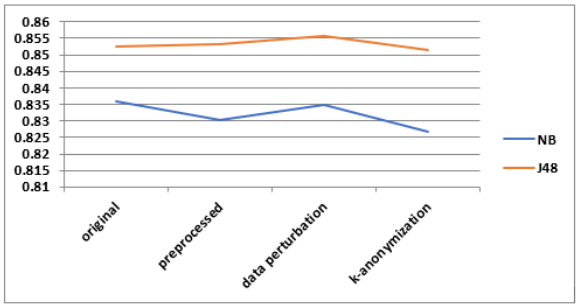
	J48				
	Original		Preprocessed		
	Original	Perturbed	Preprocessed	Perturbed	K-anony-mized
Correctly classified instances	0.8524	0.8550	0.8533	0.8557	0.8513
Incorrectly classified instances	0.1475	0.1449	0.1466	0.1442	0.1486
Kappa Statistics	0.5665	0.5722	0.59	0.5754	0.585
Mean Absolute error	0.1974	0.1974	0.1989	0.2013	0.1983
Root Mean Squared error	0.3247	0.3208	0.332	0.325	0.3297
Relative absolute error	0.5415	0.5444	0.5291	0.5441	0.5274

Since geometric data perturbation can be applied only to numerical attributes, for categorical attributes k-anonymization technique is applied. All the categorical attributes are generalized in such a way that at least k records in perturbed data will have same values for the categorical quasi identifier attributes education, native country. The modified data set in step1 is applied with K-anonymization and final modified data set is obtained. The classification algorithms Naive Bayes and J48 are applied on final modified dataset. The results obtained after k-anonymization with k=3 are shown below. The classifier accuracy can be represented using a graph as shown below for differently perturbed datasets of original Adult Data Set.

Table 6: Final Classification Result After Applying K-Anonymization

	NB		J48	
	Correctly classified instances	Incorrectly classified instances	Correctly classified instances	Incorrectly classified instances
Original Data Set	0.8359	0.1640	0.8524	0.1475
Preprocessed Data Set	0.8302	0.1697	0.8533	0.1466
After applying Geometric data perturbation	0.8347	0.1652	0.8557	0.1442
After applying K-anonymization	0.8269	0.1730	0.8513	0.1486

The following graph depicts the accuracy of classification after applied with different kinds of data perturbation techniques like geometric data perturbation and K-anonymization.



Conclusion

In this paper, we proposed an effective combined perspective for preserving privacy during data mining. We applied geometric data perturbation technique for numerical data and for categorical data one of k-anonymization technique, specifically generalization method is applied. It is shown that even after applying privacy preserving methods, the data mining results do not vary much. In future, a different k-anonymization techniques can be applied for better accuracy.

References

[1] F. Westin, "Privacy and freedom," Washington and Lee Law Review,vol. 25, no. 1, p. 166, 1968.

[2] P. Samarati, "Protecting respondents' indentities in microdata release", IEEE Transactions on Knowledge and Data Engineering, vol.13, pp.1010-1027, Nov 2001

[3] L. Sweeney, "Achieving k-anonymity privacy protection using generalization and suppression," International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 05, pp. 571–588, 2002.

[4] Stanley R. M. Oliveria, Osmar R. Zaiane, "Data Perturbation by rotation for privacypreserving Clustering", Technical Report TR 04-17, August 2004.

[5] Kun Liu, HillolKargupta, Senior Member, IEEE and Jessica Ryan, "Random ProjectionbasedMultiplicative Data Perturbation for Privacy Preserving Distributed Data Mining",IEEE transaction on knowledge, Vol. 18, No. 1, January 2006

[6] C.C. Agarwal, "On randomization, public information and the curse of Dimensionality", IEEE 23rd International conference on Data engineering, pp.136-145, April 2007.

[7] E. Bertino, D. Lin, and W. Jiang, "A survey of quantification of privacypreserving data mining algorithms," in Privacy-preserving data mining.Springer, 2008, pp. 183–205.

[8] C. Aggarwal and P. S. Yu, "A General Survey of Privacy-PreservingData Mining Models and Algorithms," Privacy preserving data mining,pp. 11–52, 2008.

[9] M. Langheinrich, "Privacy in ubiquitous computing," in Ubiquitous Computing Fundamentals. CRC Press, 2009, ch. 3, pp. 95–159

[10] Keke Chen, Ling Liu, "Geometric data perturbation for privacy preserving outsourced data mining", Knowledge Information and Systems,2010

[11] XueyunLi,ZhengYan,PengZhang,"A Review on Privacy-Preserving DataMining" in IEEEInternational Conference on Computer and Information Technology,2014.

[12] C. Aggarwal, Data mining: the textbook. Springer, 2015.

[13] Ronny Kohavi and Barry Becker UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Adult, CA: University of California, School of Information and Computer Science.