THE INTERNATIONAL JOURNAL OF SCIENCE & TECHNOLEDGE

A Randomized Approach for Privacy Preservation

Sharath Kumar Jagannathan
Assistant Professor, Department of School of Computing Science and Engineering,
VIT University Chennai, Tamil Nadu, India
Dr. N. Maheswari
Professor, Department of School of Computing Science and Engineering, VIT University Chennai, Tamil Nadu, India
Anita Susan John
M.Tech. (Big Data) Student, Department of School of Computing Science and Engineering,
VIT University Chennai, Tamil Nadu, India

Abstract:
In this time of big data boom, data mining has become a service. The main idea of data mining is to form generalized knowledge rather than finding out personalized information about a particular individual. When privacy regulations are there it prevents data owners from sharing information. So data owners should devise a strategy that guarantees privacy as well as data mining results. This paper discusses the possibility of using a randomized approach with maximization and minimization for privacy preservation.

Keywords: *Randomization, data mining, optimization, Gaussian, privacy preservation models—randomized*

1. Introduction

Since the big data boom, there is an increased ability to efficiently store data. So the concerns of the personal data for using malicious purposes have increased. Data mining tasks can be performed in a privacy preserving way. There are many techniques such as data modification, cryptographic techniques, protocols for data sharing, statistical techniques for disclosure and inference control, query auditing methods, randomization and perturbation based techniques [i]. The partial information hiding techniques is divided into, the perturbation based approaches and partition based approaches [vi]. In perturbation based approaches privacy is preserved by adding noises to data [iv]. But the noise must follow certain criteria like some kind of distribution is followed in order to make the modified data having several characteristics as that of original data [vi] [ii] [iii]. Authors use perturbation. In partition based approach data is divided into disjoint group's. Generalized information about each group can be released. Some of the popular approaches are k-anonymity [iv] and condensation [iii]. There are two types of privacy preserving techniques. Pertubative and non pertubative techniques. In perturbation the two types are Input perturbation and output perturbation. In input perturbation first data is perturbed and then it is published where in output perturbation, a statistical data base is created and a little amount of noise is added to the result of the query. [xv] The application of privacy preserving data mining is vivid and vast. It can be used in health care [vii], counter terrorism [ix], home land security [viii], medicine [x], business collaboration [xi], surveillance of bio-terrorism [xii], and so on. In the case of medical research, privacy preserving data mining problem is there. Suppose there is large number of medical institutes who want to mine patient data [xiv]. Assume that there are some privacy laws which is a hindrance to these hospitals in pooling the patient records, because of the confidentiality of the patient records. Hence there is a problem where in the normal data mining techniques cannot be used. So there arises a need to find the solution which lets the hospital to mine the patient records for research without revealing the personal identity or breaching privacy of the patient. Institutes cannot allow other institutes to directly access their records or share their records, but they must carry out data mining for research purposes. In the case of homeland security also this case is very common. Different agencies cannot allow others to view their records. But in some areas they must jointly carry out some operation. A method is proposed where it chooses aggregation of the data by accessing the risk of disclosure if it is released [xiv]. Describes about different types of terrorist threats. It also provides an overview of how data mining can be used to give solutions to defeating terrorism [ix].

2. Related Work

Algorithms for privacy are classified into three main categories. Value distortion methods, Value transformation methods and cryptographic methods. A brief outline of each is presented in this section [xvi].

2.1. Pertubative Methods

It transforms original data, for that some mathematical transformation methods are used. There are two main classes one is probability distortion and other is value distortion. In probability distortion data is replaced with a new sample from the same distribution. In Randomization Approach, Data is perturbed by noise which can be additive, multiplicative or some other randomization techniques. Some additive noise can be added to distort probability distribution of data. A noise n can be taken from a known probability distribution and can be added to an attribute a so as to modify it [xvii]. The result is x=a+n which is the perturbed attribute, n is drawn from a uniform distribution over a segment or it can be drawn from Gaussian distribution. If it is used for multiple attribute then, each attribute is perturbed independently [xvii] Bayes function can be used for reconstruction [xxi]. Discusses about discovering association rules over randomized data [xxi] For e.g. Consider a transaction with product preferences and data miner wants to find out the item sets whose support is equal or above a certain threshold. Privacy is achieved in this case by adding or discarding new item sets before being given to data miner. Other techniques are multiplicative noise, Random projection [xviii] based transformation and Rotation based transformation [xix]. They are used for distributed data mining.

2.2. Non-Pertubative Methods

It is based on partial suppressions or reductions and it does not depend upon the distortion of original data. Few of these techniques are discussed here in k-Anonymity model [iv] data is generalized or suppressed to protect the privacy. If the record is not distinguished from n-1 people, then the data is not released in suppression, some values are replaced with a symbols such as asterisk. There are two tables one is the original table and another anonymzed table. In anonymized table certain attributes are replaced by some symbols like asterisk. In generalization the wider category values are used to replace individual values. We will take the example of age, if it is 39 It is replaced by ¡=40 and if it is 45, it is replaced by ¡=50. So broader group replaces the individual values.

2.3. Cryptographic Techniques

In this technique patterns are extracted by giving a very little information among the other nodes. There are several cryptographic techniques and one is secure multiparty computation. Here a secret input is derived from two parties and no one knows anything and it will only know the output which is a designated one. Circuit valuation protocol, homomorphic encryption, etc act as the foundation stone of SMC. The protocol used in SMC is MPC. In SMC there are many types of adversaries. First one is the corruption strategy. In that model corruption of parties are taken into account, like when are they corrupted and how. In SMC there are two other models.

In Static corruption model there are set of parties, in this set those who are honest will remain honest throughout and those who are corrupted will remain corrupted throughout.

In Adaptive corruption model there is no fixed number of parties; it has the ability to corrupt in the course of computation. The choice entirely lies with adversary so this model is called adaptive. In this method whenever a party gets corrupted, it will remain like that. This technique has longer computational time and bigger overhead.

3. Proposed Work

3.1. Randomized Approach with Maximization and Minimization Function

The Gaussian randomized method with optimization functions for privacy preservation is discussed in this section. This technique is used in privacy preserving data mining [ii] Consider a set of data S={s1..sN}.Consider record s ∈ S, add a noise which is taken from a probability distribution fR(r). We get noise that are s1+r1.....sN+rN.consider this set as a1...aN. If the variance is large, then original values cannot be guessed. If the original data follows any distribution that can be recovered. So suppose S is the random number showing original distribution and R showing noise distribution. Then M = S + R S = M−R suppose if we know an occurrences of probability distribution of the original data. For a set of N numbers M can be approximated by different methods. One of such methods are Kernel density estimation. If we want to get the original distribution, we can subtract by using iterative methods. These methods have more accuracy that sequentially approximating and then subtracting. At the end we have distribution which has the properties of S. Therefore, only the algorithms which can work with uni-variate distributions can handle this situation. Multi-variate and density estimation techniques becomes hard. Curse of dimensionality is a problem in this approach. One of the advantages of this approach is it is simple. We do not want to know about the distribution of other records. During data collection period this method can be done. It has certain weakness also. It considers all records equally. So outliers are more prone to attack Randomization is used in many problems. It is used in classification [ii]. Mining rules which are used in association are proposed in [iii] Behaviour of the attribute depends on presence or absence of the records. Sometimes some of the records are dropped or some of the records are added. LAP also uses randomized function.SVD based collaborative filtering also uses this approach.

3.2. The Model

Assume there are m clients connected to a Proposal Engineering server [xxiv]. Clients have some private information. It can be shown to only some users with some privileges. For others it need not be shown. Hence the clients send its data in a privacy preserved form. If users without having these privileges see the data, it may cause loss to the company.

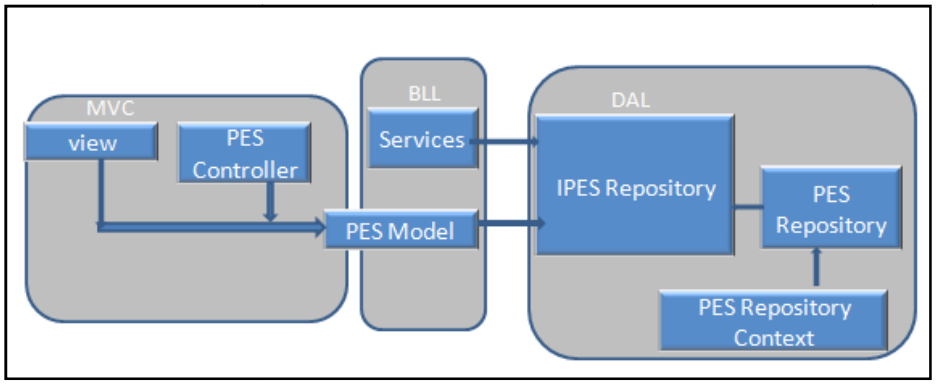

Figure 1: Architecture Diagram of the model

3.3. Proposed Algorithm

The client has the privilege to see the data. Suppose the data is composed of design details like Thickness of grade slab, offset of excavation, Pcc thickness, steel consumption etc. Since it is a personal data to a particular client it does not want to share the information A normalization function is applied on the given data The calculation for finding normalized value in the case of minimization is:

$Nx = Dmax - Dx/Dmax - Dmin$, $x = 1,2,...,X$ (3.1)

For maximization, normalized value is calculated as discussed below.

$Nx = Dx - Dmax/Dmax - Dmin$, $x = 1,2,...,X$, Dmin and Dmax are the values of minimum and maximum values in the data set (3.2)

$Z0 = sqrt(-2ln(x1))cos(2pix2)$ (3.3)

$Z1 = sqrt(-2ln(x1))sin(2pix2)$ (3.4)

$\cos\theta = u/R$ (3.5)

$\sin\theta = v/R$ (3.6)

$R = sqrt(u*u + v*v) = s$ (3.7)

For each point accepted, the polar transformation $Z0= sqrt(-2*log(s)/s)*u$ (3.6)

$Z1= sqrt(-2*log(s)/s)*v$ (3.7)

The original data is first modified by applying minimization and maximization function and then the random numbers are added to the modified data.

4. Results and Discussion

In this architecture data is passed through a randomized model, where the data is optimized with respect to maximization and minimization and random Gaussian noise are added. Hence, the original data is masked and decrypted values are obtained. According to [1], if original value can be found out with a confidence m% confidence and if it lies in the interval [α1,α2],Interval width is equal to(α2−α1). It estimates the amount of privacy at c% confidence level [i] A Gaussian distribution cannot fall under a finite range. But a truncated Gaussian can fall under a particular range. Hence privacy quantification, the measure by which how closely one can estimate the actual value is relatively high in this case. When adding large amount of noise, it is difficult to get the original data. However, distribution of the data can be drawn and it is used in data mining. In this method the maximization and minimization function is applied on the data and Gaussian noise is added which makes the data can be retrieved by the data owner and since it does not fall under a range, it is difficult to predict.

There are many advantages to this approach. One of the main advantages of this method is it is simple. We do not want to know about the distribution of other records. During data collection period this method can be done. This is a method of probability distribution. It has certain weakness also. It considers all records equally. So outliers are more prone to attack. Gaussian randomization multiplication is also there. In that approach, not a number but a matrix is added to the data set so that it can be perturbed. Randomization is used in many problems. It is used in classification [ii] Mining rules which are used in association are proposed in [iii] Behaviour of the attribute depends on presence or absence of the records. Sometimes some of the records are dropped or some of the records are added. LAP also uses randomized function.SVD based collaborative filtering also uses this approach. For generating random numbers polar form is used.

Original Data(RCC Thickness)	Distorted data(data after maximization)
1.2	1.13
3.2	2.43
4.3	4.06
0.9	0.26
0.6	0.56
6.7	5.74
56.4	53.34
34.6	32.14
24.6	23.26
18.5	16.91

Table 1: Original Data and Distorted Data after maximization

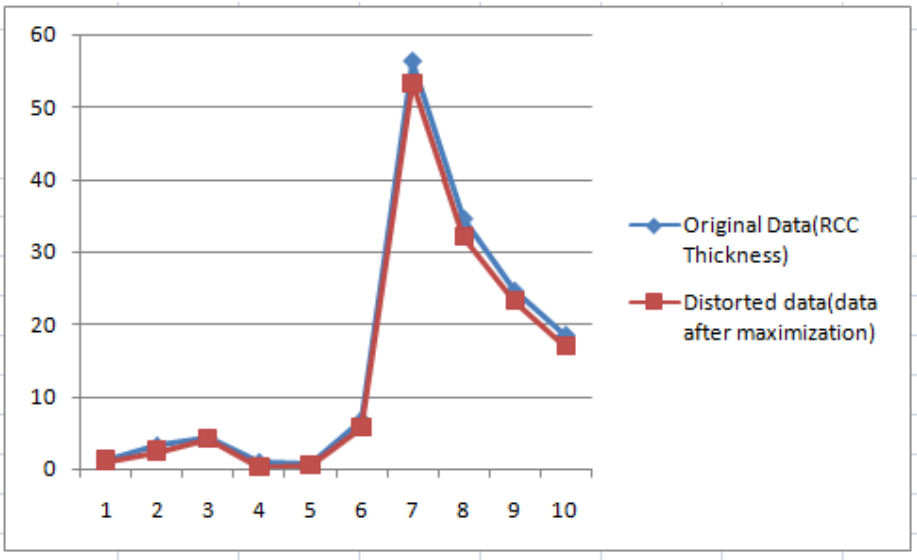

Figure 2: Data after maximization

Original Data(RCC Thickness)	Distorted data(data after minimization)
1.2	-17.105
3.2	-13.688
4.3	-10.9
0.9	-18.28
0.6	-18.3
6.7	-6.688
56.4	93.2951
34.6	49.116
24.6	29.695
18.5	16.911

Table 2: Original Data and Distorted Data after minimization

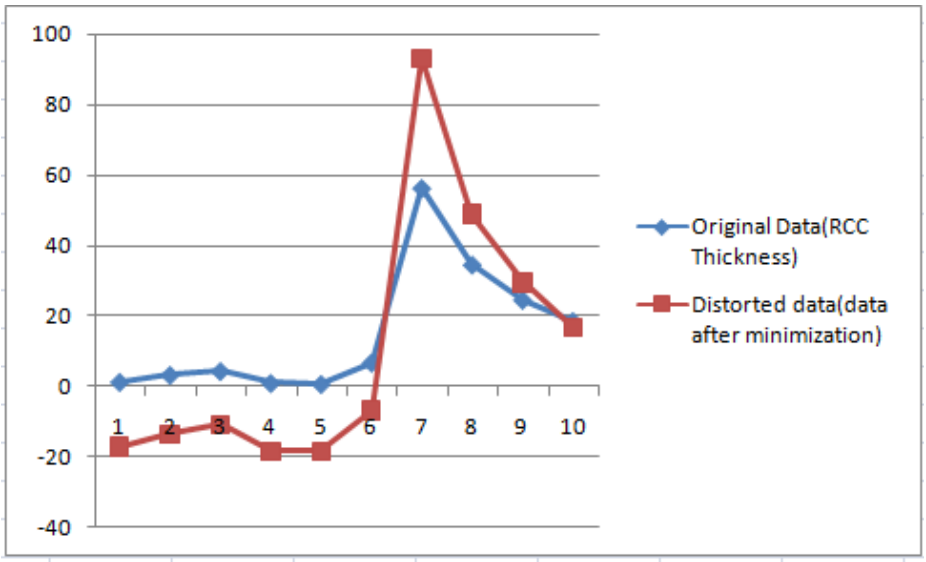

Figure 2: Data after minimization

5. Conclusion

The proposed algorithm can be used for preserving privacy. Since the model uses randomized approach, the operation can be done at the time of data collection itself. Since it follows a Gaussian distribution with maximization and minimization optimization, it is difficult for the miner to find out the data and thus the privacy is preserved. Some of the optimization algorithms can be applied on the data which makes the algorithm still more difficult to predict, which is the future scope of this work

6. References

i. Aggarwal Charu C.,Yu, PhilipS, Privacy preserving Data Mining Models and Algorithms Springler pp 47
ii. S. Papadimitriou, F. Li, G. Kollios, and P.S. Yu, Time Series Compressibility and Privacy, Proc. 33rd Intl Conf. Very Large Data Bases (VLDB), pp. 459-470, 2007.
iii. L. Singh and M. Sayal, Privacy Preserving Burst Detection of Distributed Time Series Data Using Linear Transforms, Proc. IEEE Symp. Computational Intelligence and Data Mining (CIDM), pp. 646-653, 2007.] .
iv. L. Sweeney, k-Anonymity: Privacy Protection Using Generalization and Suppression, Intl J. Uncertainty Fuzziness and Knowledge-Based Systems, vol. 10, no. 5, pp. 571-588, 2002.
v. C.C. Aggarwal and P.S. Yu, A Condensation Approach to Privacy Preserving Data Mining, Proc. Ninth Intl Conf. Extending Database Technology (EDBT), pp. 183-199, 2004.
vi. Lidan Shou, Xuan Shang, Ke Chen, Gang Chen, and Chao Zhang Supporting Pattern-Preserving Anonymization for Time-Series Data ieee transactions on knowledge and data engineering, vol. 25, no.4, 2013
vii. Behlen.F.M, Johnson. S.B, Multicenter Patient Records Research: Security Policies and Tools, J Am Med Inform Assoc. 6(6) 435(1999)
viii. Fienberg.S.E (2005), Homeland insecurity: Data mining, terrorism detection, and confidentiality, Bull. Internat. Stat. Inst., 55th Session. Sydney
ix. Thuraisingham.B (2003), Web Data Mining and its Applications in Business Intelligence and Counter-terrorism, CRC Press.
x. Berman. J.J. (2002), Confidentiality Issues for Medical Data Miners, Artif Intell Med. 26(1-2): 25-36.
xi. Oliveira S.R.M and Zaiane. O.R (2007), A privacy-preserving clustering approach toward secure and effective data analysis for business collaboration, Journal of Computer and Security, 26, pp.81-83.
xii. Sweeney.L (2005), Privacy-Preserving Bio-terrorism Surveillance, AAAI Spring Symposium, AI Technologies for Homeland Security
xiii. Yehuda Lindell and Benny Pinkas(2009),Secure Multiparty Computation for Privacy-Preserving Data Mining, The Journal of Privacy and Condentiality 1, pp. 5998
xiv. Boyens.C, Krishnan.R and Padman.R (2004), On privacy-preserving access to distributed heterogeneous healthcare information, System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference.
xv. Xun Yi n, Yanchun Zhang(2013)97-107, Equally contributory privacy-preserving k-means clustering over vertically partitioned data
xvi. Vaidya.J, Clifton.C and Zhu.M (2006),Privacy Preserving Data Mining, ISBN: 978-0-387-258867, Advances in Information Security, Springer, 19.
xvii. Evfimievski.A (2002), ”Randomization in Privacy Preserving Data Mining,” ACM SIGKDD Explorations Newsletter, 4, 2, pp: 43-48.
xviii. Kargupta.H, Liu.K and Ryan.J (2006), Random Projection-Based Multiplicative Data Perturbation for Privacy Preserving Distributed Data Mining, IEEE Transactions on Knowledge and Data Engineering, 18,1, pp: 92 106.

xix. Ketel. Mand Homaifar. A (2005), Privacy-Preserving Mining by Rotational Data Transformation, Proceedings of the 43rd annual southeast regional conference, 1, pp: 233 236.
xx. Sweeney.L (2002), k-Anonymity: A Model for Protecting Privacy, Intl J. Uncertainty, Fuzziness, and Knowledge-Based Systems, 10,5, pp. 557-570.
xxi. A. Evmievski, R. Srikant, R. Agrawal, and J. Gehrke, Privacy preserving mining of association rules, In Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery in Databases and Data Mining, pages 217228, Edmonton, Alberta, Canada, July 2326 2002;