World Applied Sciences Journal 29 (Computer Sciences, Engineering and Its Applications): 19-24, 2014
ISSN 1818-4952
© IDOSI Publications, 2014
DOI: 10.5829/idosi.wasj.2014.29.csea.2209

Random Projection Based Data Perturbation Using Geometric Transformation

R. Udayakumar, K.P. Thooyamani and Khanaa

School of Computing Science, Bharath University, Chennai - 73

Abstract: This paper explores the use of random projection matrices as a tool for privacy preserving data mining. The problem discussed here is as follows: Suppose there are N organizations O1, O2 ….On. Each organization is having a private transaction database. A third party data miner wants to learn the statistical properties of the union of these databases. These organizations don't want to disclose their raw data.The existing work is based on orthogonal transformation matrix. The problem with this method is that the data is not masked after the transformation. So the security of the data is questionable. In this paper, privacy preservation in data mining is done by using random projection and multiplicative data perturbation. To determine how to perturb the data, the row wise and column wise projection of the random matrix is done. It proves that after perturbation the statistical properties are maintained without changing dimensionalities and exact values. Finally this is combined with geometric transformations to obtain better performance results. Here geometric transformation means we combine the random projection perturbation with transformation, scaling and translation. Privacy in data mining requires understanding how privacy can be violated and possible means of preventing privacy violation.

Key words: Data mining · Privacy preserving data mining · Geometic transformation

INTRODUCTION

Data Mining and knowledge discovery in databases (KDD) are two new research areas that investigate the automatic extraction of previously unknown patterns from large amounts of data. Privacy preserving data mining is a novel research direction in data mining and statistical databases. Protecting privacy is an important concern of the society. Recent interest in the collection and monitoring of data using data mining technique for the purpose of security and business related application has raised serious concerns about privacy issues. In general, one major factor contributes to privacy violation in data mining is data misuse.

Preserving private data is becoming an increasingly important issue in many data mining applications. Privacy of the data can depend on many different aspects often dictated by the characteristics of the domain. Sometimes individuals or organizational entities may not be willing to divulge the individual values of the records. Some times the patterns can be detected by a data mining system may be used in a counter productive manner that violates the privacy of an individual or a group of individuals. Given a set of privacy constraints the goal of privacy preserving data mining system is to extract well defined product computations.

The main consideration in privacy preserving data mining is two fold:

Sensitive raw data like identifiers, names, addresses and the like should be modified or trimmed out from the original database, in order for the recipient of the data not to be able to compromise another person's privacy.

Sensitive knowledge which can be mined from a database by using data mining algorithms should be excluded, because such knowledge can equally well compromise data privacy.

Classification of Privacy Preservation Techniques: There are many approaches which have been adopted for privacy preserving data mining. We can classify them based on the following dimensions:

Corresponding Author: R. Udayakumar, School of Computing Science, Bharath University, Chennai – 73,

- Data Distribution
- Data modification
- Data mining algorithm
- Data or rule hiding
- Privacy preservation

The first dimension refers to the distribution of data. The distributed scenarios can also be classified as:Horizontal distribution refers to the cases where different database records reside in different places. Vertical data distribution refers to the case where all values for different attributes reside in different places.

The second dimension refers to the data modification scheme. The data modification is used in order to modify the original values of a data base that needs to be released to the public and in this way to ensure high privacy protection. The methods of modification include: perturbation, which is accomplished by the alteration of an attribute value by a new value (changing a l-value to a 0-value, or adding noise),blocking, which is the replacement of an existing attribute value with a"?". Aggregation or merging which is the combination of several values into a coarser category. Swapping that refers to interchanging values of individual records and Sampling, which refers to releasing data for only a sample of a population? The third dimension refers to the data mining algorithm, for which data modification is taking place. The most important mining algorithms are Decision tree inducers, Association rule mining, Clustering algorithms and Rough sets and Bayesian Networks [2].

The fourth dimension refers to whether raw data or aggregated data should be hidden. The last dimension refers to privacy preservation technique used for the selective modification of the data.

Privacy Preserving Data Mining: The goal of data mining is to extract knowledge from a large data base. Most data mining applications operate under the assumption that all data is available at a single repository called a datawarehouse.This posses huge privacy problem because violating a single repository's security exposes all the data. Whether the data warehouse is real or virtual is irrelevant; if the data mining algorithm can access data, the opportunity exists for an attacker to get it [3].

One real life example that describes privacy preserving data mining is as follows:

A hospital shares some data for research purposes (e.g. concerning a group of patients who have similar diseases). The hospital's security administrator may suppress some identifiers from patient records to meet privacy requirements. However the released data may not be fully protected. A patient record may contain other information that can be linked with other datasets to reidentify the individuals or entities.

Generally when people talk of privacy, they say "keep information about me from being available to others". However, their real concern is that their information not be misused. The fear is that once information is released, it will be impossible to prevent misuse. Utilizing this distinction -ensuring that a data mining project won't enable misuse of personal information - opens opportunities that "complete privacy" would prevent. To do this, we need technical and social solutions that ensure data will not be released. Another view is corporate privacy - the release of information about a collection of data rather than an individual data item. I may not be concerned about someone knowing my birth date, mother's maiden name, or social security number; but knowing all of them enables identity theft. This collected information problem scales to large, multiindividual collections as well. A technique that guarantees no individual data is revealed may still release information describing the collection as a whole. Such corporate information is generally the goal of data mining.

Data Perturbation: Data perturbation approaches fall into two main categories- the probability distribution approach and the value distortion approach (Fig 1). The probability distribution approach replaces the data with another sample from the same (estimated) distribution or by the distribution itself. On the other hand, the value distortion approach perturbs the values of data elements or attributes directly by some additive or multiplicative noise before it is released to the data miner. In this paper, we mainly focus on the value distortion techniques [8]. A value distortion technique to protect the privacy by adding random noise from a Gaussian distribution to the actual data. They showed that this technique appears to mask the data while allowing extraction of certain patterns like the original data distribution and decision tree models with good accuracy. The use of rando m additive noise pointed out that additive noise can be easily filtered out in many cases that may lead to compromising the privacy.

Given the large body of existing signal-processing literature on multiplicative noise is added for protecting the privacy of the data while maintaining some of the original analytic properties.

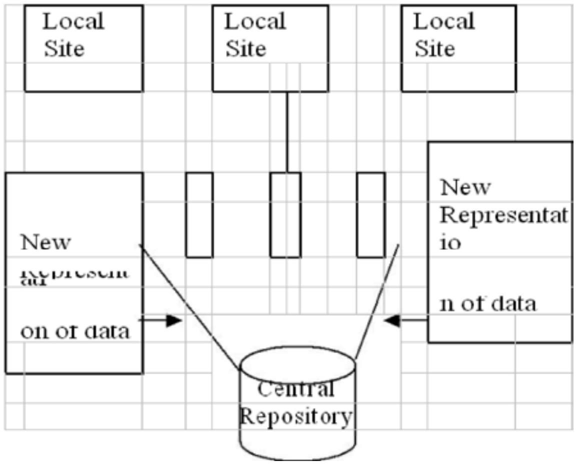

Fig. 1: An idea of data perturbation approach

Two basic forms of multiplicative noise are introduced in the statistics community.

- Based on generating random numbers that have a truncated Gaussian distribution with meanone and small variance, and multiplying each element of the original data by the noise.
- To take a logarithmic transformat ion of the data first(for posit ive data only), compute the covariance, andgenerate random noise following a multivariate Gaussian method which assures higher security than the first one and still maintains the data utility very well in the log-scale. Finally take the antilog of the noise-added data.

Multiplicative perturbation overcomes the scale problem and it has been proved that the mean and variance/covariance of the original data elements can be estimated from the perturbed version. In practice, the first method is good if the data disseminator only wants to make minor changes to the original data; however the second method assures higher security than the first one and still maintains the data utility very well in the log-scale.

One of the main problems of the traditional additive perturbation and multiplicative perturbation is that they perturb each data element independently and therefore the similarity between attributes or observations which are considered as vectors in the original data space is not well preserved. In this paper, we propose an alternative approach to perturb data using multiplicative noise. We make use of random projection matrices for constructing a perturbed representation of the data.

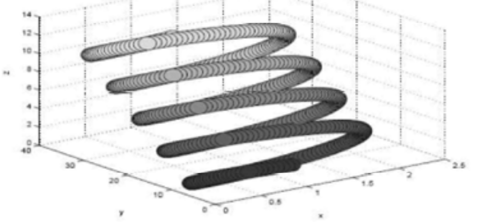

Fig. 2: The perturbed data after a random orthogonal transformation

Random Orthogonal Transformation: This section presents a multiplicative perturbation method using random orthogonal matrices in the context of computing inner product matrix. Later we shall analyze the deficiency of this method and then propose a more general case that makes use of random projection matrices for better protection of the data privacy [10].

An orthogonal transformation is a linear transformation which preserves the length of vectors as well as the angles between them.

Usually orthogonal transformations correspond to and may be represented using orthogonal matrices. In the meantime, we can imagine both the privacy sensitive data and the transformation procedure are inside a black box, the perturbed data is the only output to the third party, the observer. Since only transformed data is released there are an infinite number of input and transformation procedures that can stimulate the output.Thus random orthogonal transformations are a good way to protect privacy while preserving its utility (Fig 2).

From the geometric point of view, an orthogonal transformation is either a pure rotation when the determinant of the orthogonal matrix is 1; or a roto inversion (a rotation followed by a flip) when the determinant is -1 and therefore it is possible to identify the real values of the data through a proper rotation vectors are statistically independent and they do not follow Gaussian distribution, it is possible to estimate their original forms quite accurately using Independent Component Analysis (ICA). In next section we propose a random projection-based multiplicative perturbation technique to improve the privacy level while preserving the data utilities.

Random Projection Based Data Perturbation: Random projection refers to the technique of projecting a set of data points from a high dimensional space to a randomly

1

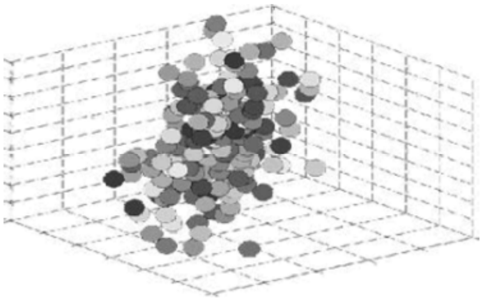

Fig. 3: The perturbed data after a row wise projection which reduces 50 percent of the data points

chosen lower dimensional subspace. The key idea of random projection arises from the Johnson-Linden Strauss lemma [10]. It is as follows:For any 0<n<1 and any integer s, let k be a positive integer such that k>=k0 =O (e2log n). For every set S of n points in R d there exists f:Rd->Rk such that for all u, v element of P,

(1-_) ||u-v||2|| f(u)-f(v)||2(1+_)||u-v||2

It is possible to change data's original form by reducing the dimensionality but still maintains its statistical characteristics. To perturb the data the row-wise projection (Fig 3) preserves the column-wise inner product and the column-wise projection preserves the row-wise inner product. The beauty of this property is that if the data is properly normalized, the inner product is directly related to the cosine angle, the correlation and even the Euclidean distance of the vectors. Random projection based technique guarantees that both dimensionality and exact value of each element of original data are kept confidential. An analysis of how much preservation can be obtained is explained by considering some properties of random matrix [11].

Geometric Transformations: There are mainly three transformations:

- Translation
- Rotation
- Scaling

Translation Data Perturbation: Translation is to move a point with coordinates(X,Y) to a new location by using displacements(X0,Y0).The translation is accomplished by using a matrix representation V^0 = Tv,where T is a 2*3 transformation matrix v is a vector containing the original coordinates. V^0 is a column vector whose coordinates are transformed coordinates.

Algorithm:

Input: V, N
Output: V 0

Step 1: For each confidential attribute Aj in V, where 1 ● j ● d do

- Select the noise term ej in N for the confidential attribute Aj
- The j-th operation opj ● {Add}

Step 2: For each vi V do For each aj in vi = (a1,,,,,,, ad), where aj is the observation of the j-th attribute do a^0 ● Transform (a^j, op^j, e^j) *j* End

Scaling Data Perturbation: In Scaling Data Perturbation, the attributes are perturbed using multiplicative noise perturbation. The noise term is constant and can be either positive or negative.

Algorithm:

l

Step 1: For each confidential attribute Aj in V, where 1 ● j ● d do

- Select the noise term ej in N for the confidential attribute Aj
- The j-th operation opj ● {Mult}

Step 2: For each vi V do

For each aj in vi = (a1,... ad), where aj is the observation of the j-th attribute do a_0 ● Transform (aj, opj, ej) End

Rotation Data Perturbation: The Rotation Data Perturbation Method, denoted by RDP. In this case, the noise term is an angle ●. The rotation angle ●, measured clockwise, is the transformation applied to the observations of the confidential attributes. The set of operations Di(OP) takes only the value {Rotate} that identifies a common rotation angle between the attributes Ai and Aj. Unlike the previous methods, RDP may be applied more than once to some confidential attributes. For instance, when a rotation transformation is applied this affects the values of two coordinates. In a 2D discrete space, the X and Y coordinates are affected. In a 3D discrete space or higher, two variables are affected and the others remain without any alteration. This requires that one or more rotation transformations are applied to guarantee that all the confidential attributes are distorted in order to preserve privacy [3].

Algorithm:

Input: V, N
Output: V_0

Step 1: For every two attributes Aj,Ak in V, where 1 ● j ● d and 1 ● k ● d do

- Select an angle ● for the confidential attributes Aj,Ak
- The j-th operation opj ● {Rotate}
- The k-th operation opk ●{Rotate}

Step 2: For each viV do

For each al in vi = (a_1,... a_d), where al is the observation of the l-th attribute do a_0 ● Transform(a_l, op_l, e_l) End

The privacy provided by a perturbation technique has been measured as the variance between the actual and the perturbed values.This measure is given by Var(X ● Y) where X represents a single original attribute and Y the distorted attribute. This measure can be made scale invariant with respect to the variance of X by expressing security as

Sec = Var(X ● Y)/Var(X).

The procedure to improve the privacy level of the geometric transformation is applied to transformed database only. This procedure is composed of three steps as follows:

Step 1: We select a probability distribution for each confidential attribute

Step 2: We randomly select some of the vectors by to reinforce privacy by adding some noise term to each observation corresponding to probability distribution selected in previous step.

Step 3: Based on the previous steps we distort the data using the projection technique.

In this work, the introduced family of geometric transformation which ensures that the mining process will not violate privacy up to a certain degree of security.

CONCLUSION

Privacy-preserving data mining from multi-party distributed data is playing an increasingly important role in many different application domains. Many of the recent efforts in this field are motivated by the idea of perturbing the data using randomized techniques. This paper explores the use of random projection matrices as a tool for privacy preserving data mining. The analysis shows that it is difficult to recover the exact values of any elements of the data even when the secret random matrix is disclosed. The random projection-based technique is more powerful when used with some geometric transformation techniques like scaling, translation and rotation. Combining this with SMC-based techniques offer another interesting direction.

REFERENCES

1. Sweeney, L., 2002. “k-anonymity: a model for protecting privacy,” International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10(5): 557- 570, [Online].
2. Udayakumar, R., V. Khanna, T. Saravanan and G. Saritha, 2013. Retinal Image Analysis Using Curvelet Transform and Multistructure Elements Morphology by Reconstruction, Middle-East Journal of Scientific Research, ISSN:1990-9233, 16(12): 1798-1800.
3. Udayakumar, R., V. Khanna, T. Saravanan and G. Saritha, 2013. Cross Layer Optimization For Wireless Network (Wimax), Middle-East Journal of Scientific Research, ISSN: 1990-9233, 16(12): 1786-1789.
4. Kargupta, H., S. Datta, Q. Wang and K. Siva Kumar, 2003. “On the privacy preserving properties of random data perturbation techniques,” in Proceedings of the IEEE International Conference on Data Mining, Melbourne, FL,
5. Kim, J.J. and W.E. Winkler, 2003. “Multiplicative noise for masking continuous data,” Statistical Research Division, U.S. Bureau of the Census, Washington D.C. Tech. Rep. Statistics #2003-01, April 2003. [10] K. Muralidhar, D. Batrah and P.J. Kirs, Accessibility, security and accuracy in statistical databases: The case for the multiplicative fixed data perturbation approach,” Management Science, 41(9): 1549- 1584, 1995. “Protocols for secure remote database access with approximate matching,” in 7th ACM Conference on Computer and Communications

Security(ACMCCS 2000). The f i rst workshop on Security of Privacy in E-Commerce, Athens, Greece, November 2000.

6. Atallah, M.J. and W. Du, 2001. "Secure multi-party computational geometry," in WADS2001: Seventh Internat ional Workshop on Algorithms and Data Structures, Providence, Rhode Island, pp: 165- 179.
7. Du, W. and M.J. Atallah, 2011. "Privacy-preserving cooperative statistical analysis," in Proceedings of the 17th Annual Computer Security Applications Conference, New Orleans, LA,
8. Saravanan, T. and R. Udayakumar, 2013. Optimization of Machining Hybrid Metal matrix Composites using desirability analysis, Middle-East Journal of Scientific Research, ISSN: 1990-9233, 15(12): 1691-1697
9. Clifton, C., M. Kantarcioglu, J. Vaidya, X. Lin and M. Zhu, 2003. "Tools for privacy preserving distributed data mining," ACM SIGKDD Explorations, 4: 2.
10. Thooyamani, K.P. V. Khanaa and R. Udayakumar, 2013. Blue tooth broad casting server, Middle-East Journal of Scientific Research, ISSN: 1990-9233, 15(12): 1707-1712.
11. Thooyamani, K.P., V. Khanaa and R. Udayakumar, 2013. Improving Web Information gathering for personalised ontology in user profiles, Middle-East Journal of Scientific Research, ISSN:1990-9233 15(12): 1675-1679.
12. Kantarcioglu, M. and C. Clifton, 2002. "Privacy-preserving distributed mining of association rules on horizontally partitioned data," in ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD'02), June 2002.
13. Lin, X., C. Clifton and Y. Zhu, 2004. "Privacy preserving clustering with distributed mixture modeling," 2004, International Journal of Knowledge and Information Systems. To appear.
14. Lind Y., ell and B. Pinkas, 2000. "Privacy preserving data mining," in Advances in Cryptology (CRYPTO'00), ser. Lecture Notes in Computer Science, vol.1880. Springer-Verlag, pp: 36- 53.