Preserving the Privacy and Sharing the Data using Classification on Perturbed Data

P.Kamakshi
Assistant Professor
Dept. of Information Technology
Kakatiya Institute of Technology & Science
Warangal

Dr. A. Vinaya Babu
Director of Admissions
Professor, Dept. of CSE
J.N.T.U., Kukatpally
Hyderabad

Abstract* — Data mining is a powerful tool which supports automatic extraction of unknown patterns from large amounts of data. The knowledge extracted by data mining process support a variety of domains like marketing, weather forecasting, and medical diagnosis .The process of data mining requires a large data to be collected from diverse sites. With the rapid growth of the Internet, networking, hardware and software technology there is tremendous growth in the amount of data collection and data sharing. Huge volumes of detailed data are regularly collected from organizations and such datasets also contain personal as well as sensitive data about individuals. Though the data mining operation extracts useful knowledge to support variety of domains but access to personal data poses a threat to individual privacy. There is increased concern on how sensitive and private information can be protected while performing data mining operation. Privacy preserving data mining algorithms gives solution for the privacy problem. PPDM gives valid data mining results and also guarantees privacy protection for sensitive data stored in the data warehouse. In this paper we analyzed the threats to privacy that can occur due to data mining process. We have proposed a framework that allows systemic transformation of original data using randomized data perturbation technique and the modified data is submitted as a result of query to the parties using decision tree approach. This approach gives the valid results for analysis purpose but the actual or true data is not revealed and the privacy is preserved.

***Keywords* —Data perturbation, Data mining, Decision tree, Privacy preservation, Sensitive data.**

I . INTRODUCTION

Data mining is an emerging field which connects different major areas like databases, artificial intelligence and statistics. The process of data mining requires a large amount of data to be collected into a central site. In modern days organizations are extremely [9][10]dependent on data mining results to provide better service, achieving greater profit, and better decision-making. To reach their goals organizations collect huge amount of data about the consumers for marketing purposes and improving business strategies, medical organizations collect medical records for better treatment and medical research. With the rapid advance of the Internet, networking, hardware and software technology there is remarkable growth in the amount of data that can be collected from different sites or organizations. Huge volumes of data collected in this manner also include sensitive data about individuals. It is obvious that if a data mining algorithm is run against the union of such databases, the extracted knowledge not only consists of discovered patterns and correlations that are hidden in the data but it also reveals something about the contents of the other databases, which are considered to be private. Although Data mining operation efficiently discover valuable non-obvious information from large databases, it is very sensitive to privacy concerns. Privacy is an important issue in many data mining applications that deal with health care, security, financial and other types of sensitive data. On one hand the data mining process gives the knowledge which can be used to support a variety of domains like marketing, weather forecasting, and medical diagnosis. But, on the other hand, easy access to personal data poses a danger to individual privacy. The actual anxiety of people is that their private information should not be misused behind the scenes without their knowledge. The real threat is that once information is unrestricted, it will be impractical to stop misuse. Privacy can for instance be threatened when data mining techniques uses the identifiers which themselves are not very sensitive ,but are used to connect personal identifiers such as addresses, names etc., with other more sensitive personal information .The simplest solution to this problem is to completely hide the sensitive data or not to include such sensitive data in the database. But this solution is not ideal and accurate because in many applications, like medicine research, DNA research etc. different organizations or institutions wish to conduct a joint research on their databases because combining their data will definitely provide better results and mutual benefit to the organizations. In this scenario organizations want to share the data but neither of the institute or organizations want to disclose its database or private information about their clients to other party. In such a situation it is not only necessary to protect private and sensitive information but it is also essential to facilitate the use of database for investigation or for other purposes. Privacy preserving data mining [20] is a special data mining technique which has emerged to deal with the privacy issue in data mining. PPDM uses special techniques to protect the privacy of sensitive data and also give valid data mining results. In this paper we propose a novel method to preserve the privacy by perturbing the original data using randomized data perturbation privacy preserving data mining technique and then constructing a decision tree classifier on the perturbed data.

II . PREVIOUS WORK

Recently the application of data mining is increased in various domains like business, academia, communication, bioinformatics, medicine field .The data mining not only gives the valuable results hidden in these databases, but sometimes reveals private information about individuals. The difficulty is that by means of linking different attributes data mining process extracts the individual data which is considered as private. The true problem is not data mining, but the way data mining is done. PPDM is an emerging technique in data mining where privacy and data mining can coexist. It gives the summarized results without any loss of privacy through data mining process.

In general there are two main approaches in PPDM:

i) Data transformation based
ii) Cryptographic-based methods.

The data transformation based approach modifies sensitive data in such a way that it loses its sensitive meaning .In this process statistical properties of interest can be retained but exact values cannot be determined during the mining process. Various data modification techniques are noise addition [1] [2] [3], data swapping [4], aggregation [5], suppression and signal transformation.

In Cryptographic techniques the data is encrypted with encryption methods and still allow the data mining operation. These methods use certain set of protocols such as secured multiparty computation (SMC). Secure multi-party computation is a computation process performed by group of parties with distributed data set where each party has in its control a part of the input data needed to perform the computation. In SMC the participating parties should only learn the final result of the computation and no additional information is supposed to be revealed at the end of computation. Perfect privacy in the SMC [6] [7] is achieved because no information is released to any third party. The basic SMC PPDM techniques are secure sum, secure set union, secure size of set union etc.

A. Overview of randomization perturbation technique

In randomization perturbation approach the privacy of the data can be protected by perturbing [13] sensitive data with randomization algorithms before releasing to the data miner. The perturbed data version is then used to mine patterns and models. The algorithm is so chosen that combined properties of the data can be recovered with adequate accuracy while individual entries are considerably distorted. In this method privacy of confidential data [16]can be obtained by adding small noise component which is obtained from the probability distribution .The method of randomization can be described as follows. Consider a set of data records denoted by $X = \{x_1 \dots x_N\}$. For record $xi \in X$, we add a noise component which is drawn from the probability distribution $f_y(y)$. Commonly used distributions are the uniform distribution over an interval $[-\alpha, \alpha]$ and Gaussian distribution with mean $\mu = 0$ and standard deviation σ. These noise components are drawn independently, and are denoted $y_1 \dots y_N$. Thus, the new sets of distorted records are denoted by $x_1 + y_1$ ….. $x_N + y_N$. We denote this new set of records by $z_1 \dots z_N$. In general, it is assumed that the variance of the added noise is large enough, so that the original record values cannot be easily guessed from the distorted data. Thus, the original records cannot be recovered, but the distribution of the original records can be recovered. One key advantage of the randomization method is that it is relatively simple, and does not require knowledge of the distribution of other records in the data. Our experiment was performed on numerical database by applying Gaussian technique to all the attributes in a given database. The same technique can be applied to only selected attributes, which the database administrator considers as more sensitive.

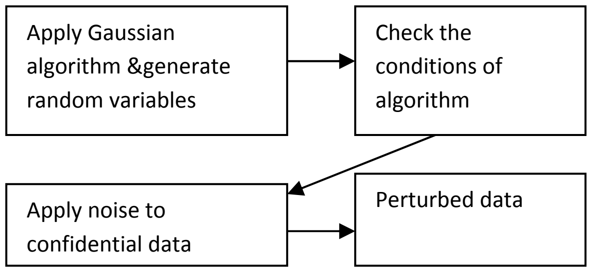

Figure1. **B**lock diagram for implementing perturbation technique

B. Overview of decision tree

Classification is one of the forms of data analysis that can be used to extract models describing important data classes or to predict future data. Decision trees are powerful and popular tools for classification and prediction. The attractiveness of decision trees is due to the fact that it is represented using rules. Rules can readily be expressed so that humans can understand them or even directly used in a database access language like SQL so that records falling into a particular category may be retrieved. Decision tree represents a tree structure, where each node is either a leaf node indicating the value of the target class of given datasets or a decision node on which some test can be performed resulting one branch or sub-tree for each possible outcome of the test.

A decision tree is a class discriminator that recursively partitions the training set until each partition entirely or dominantly consists of examples from one class. The main task for building a decision tree is to identify an attribute for the splitting point based on the information gain that measures how well a given attribute separates the training examples according to their target classification. Information gain can be computed using entropy. The attribute with highest information gain will form the root of the tree and algorithm iteratively continues splitting the data to form a decision tree. A decision tree [15] can be used to classify an example by starting at the root of the tree and moving through it until a leaf node, which provides the classification of the instance.

ID3 Decision Tree Algorithm

function ID3

Input: (R: a set of non-target attributes, C: the target attribute, S: a training set)
Output : a decision tree;

Begin

If S is empty, return a single node with value Failure;
If S consists of records all with the same value for the target attribute, return a single leaf node with that value;
If R is empty, then return a single node with the value of the most frequent of the values of the target attribute that are found in records of S;
Select test-attribute, the attribute among attribute-list with highest information gain;
Let A be the attribute with largest Gain (A, S) among attributes in R;
Let $\{a_j | j=1, 2... m\}$ be the values of attribute A;
Let $\{S_j \mid j=1, 2... m\}$ be the subsets of S consisting respectively of records with value a_j for A;
Return a tree with root labeled A and arcs labeled a_1, a_2... a_m going respectively to the trees (ID3(R-{A}, C, S_1), ID3 (R-A}, C,S_2)… ID3(R-{A}, C, Sm);
Recursively apply **ID3** to subsets $\{S_j \mid j=1, 2... m\}$ until they are empty

End.

Information gain calculatiuion

Each non leaf node of the decision tree contains a splitting point and the main task for building a decision tree is to identify an attribute for the splitting point based on the information gain. Information gain can be computed using entropy.

$$\text{Entropy (S)} = -\sum_{j=1}^{m} K_j \log K_j$$

where m represents total no. *of* classes in the whole training data set, K_j is the relative frequency of class *j* in *S*. Based on the entropy, information gain can be computed for any candidate attribute A if it is used to partition S.

$$\text{Gain(S, A)} = \text{entropy (S)} - \sum_{p \in A} \left(\frac{|Sp|}{|S|} \text{Entropy } (S_p) \right)$$

Where p represents any possible values of attribute A. S_p is the subset of S for which attribute A has value p,$|S_p|$is the number of elements in S_p ,|S| is the number of elements in S. To find the best split for a tree node, we compute information gain for each attribute. We then use the attribute with the largest information gain to split the node.

III . Our Framework

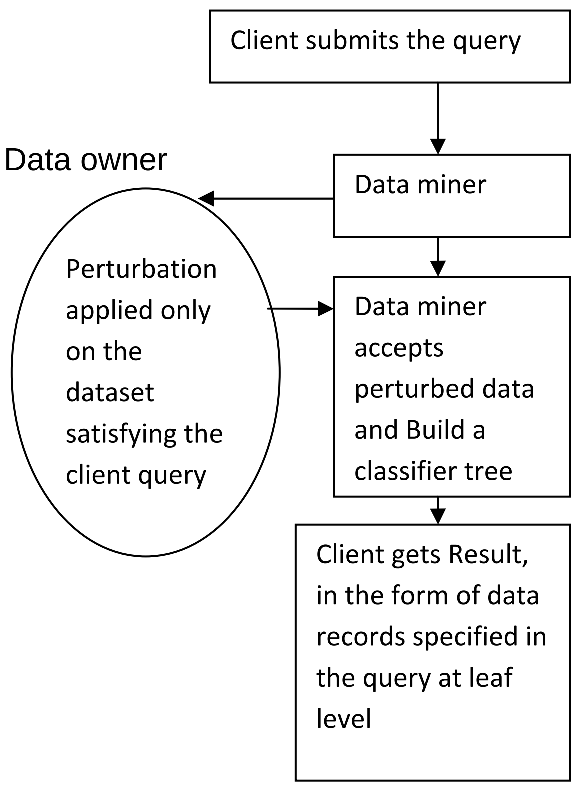

Figure 2. The framework to integrate perturbation and classification technique

In this novel framework we use two key components, data perturbation component at data provider site and classifier component in the data miner site. Our scheme is a Four -step process. In the first step, the data miner negotiates with different data provider depending on the query submitted by the user. In The second step the randomized perturbation technique is applied on the data set which satisfies the user query. In the third step data miner obtains the perturbed data from the data provider. In the fourth step a classifier is built on the perturbed data set.

This framework guarantees the privacy because the records on which the classifier is constructed is in the perturbed form. Confidentiality is also achieved because the data owner or provider does not learn anything about the classifier which has been constructed. The parameter like attribute selected at the root node, attribute used as class attributes and the records selection criteria remain hidden from the data owner. Figure below gives the example of classifier tree on perturbed data.

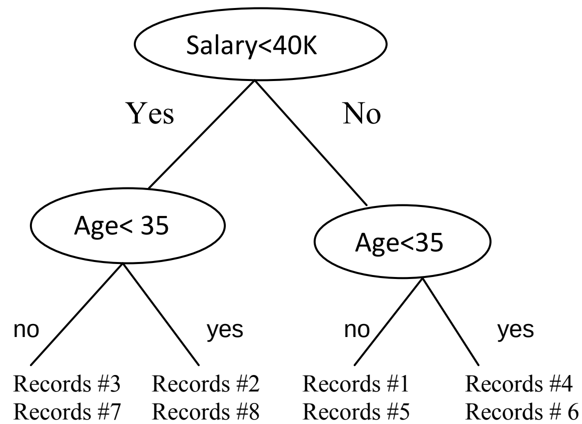

Figure 3. Example of classifier tree on perturbed data

IV. CONCLUSION

Data mining extracts useful patterns from large quantities of data stored in the data warehouse. The data mining process results valuable patterns to support decision making in different domains. But easy access to sensitive data poses threat to individual privacy. In this paper we presented a novel approach in which both data perturbation technique and classification are integrated to provide better data quality and individual privacy both at data owner site as well as at data mining site. The owner’s data consists of both categorical and numeric data types. To preserve the privacy of data at owner’s site perturbation technique is used in which small amount of noise is added to sensitive data such that the properties and the meaning of the original data is not changed. The problem with the randomization technique is that some privacy intrusion techniques can be used to reconstruct private information from the randomized data tuples. Hence to enhance the performance a decision tree is built on the perturbed data at data mining site, which reveals and gives only the required results and hides other information.

REFERENCES

[1] Agrawal R.,Srikant R., ``Privacy Preserving Data Mining.,'' In the Proceedings of the ACM SIGMOD Conference. 2000.

[2] K.Muralidhar.,R.Sarathy,“A General additive data perturbation method for data base security", journal of Management Science. ,45(10):1399-1415,2002.

[3] Agrawal D. Aggarwal C.C. “ On the Design and Quantification of Privacy Preserving Data mining algorithms.” ACM PODS Conference, 2002.

[4] Muralidhar K. and Sarathy R.,“ Data Shuffling- a new masking approach for numericadata”managements science, forthcoming, 2006.
[5] V.S. Iyengar.”Transforming data to satisfy privacy constraints” In Proc. of SIGKDD’02, Edmonton, Alberta,Canada,2
[6] Lindell Y., Pinkas B.”Privacy preserving Data Mining“ CRYPTO 2000.
[7] Yu.H.,Vaidya J.,Jiang X.”Privacy preserving SVM Classification on vertically partitioned data” PAKDD conference, 2006.
[8] IEEE Transactions on Knowledge and Data Engineering., Vol.18,No.1,2006
[9] D. Agarwal and C.C.Aggarwal, “ On the design and quantification of privacy preserving data mining algorithms”, In Proceedings of the 20th Symposium on Principles of Database systems,Santabarbara,California,USA, May2001.
[10] R.Agarwal and R.Srikant, “Privacy preserving data mining”, In Proceedings of the 19th ACM SIGMOD conference on Management of Data ,Dallas,Texas,USA, May2000.
[11] J. Canny, “Collaborative filtering with privacy”. In IEEE Symposium on security and privacy , pages 45-57 Oakland, May 2002.
[12] M. Kantarcioglu and C. Clifton, “Privacy-preserving distributed mining of association rules on horizontally partitioned data”, In the ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD’02), pp. 24–31,Madison, June 2002,
[13] K. Muralidhar, R. Sarathy, and R. A. Parsa, “A general additive perturbation method for database security,” Management Science, vol. 45, no. 10, pp. 1399–1415, 1999.
[14] R.Agrawal, A.Evfimievski, R.Srikant, “Information sharing across private databases”, In Proc.of ACM SIGMOD, 2003.
[15] J. Han and M. Kamber, “Data Mining: Concepts and Techniques”. Morgan Kaufmann Publishers, 2000.
[16] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar, “On the privacy preserving properties of random data perturbation techniques” In Proc. of 3rd IEEE Int. Conf. on Data Mining, Washington, DC, USA,, pages99–106, 2003.
[17] K. Muralidhar, R. Parsa, and R. Sarathy, “ A general additive data perturbation method for database security”,Management Science, 19:1399–1415, 1999.
[18] B. Pinkas, “ Cryptographic techniques for privacy preserving data mining “ SIGKDD Explorations, 12–19, 2002
[19] A. Evfimievski, “ Randomization in privacy preserving data mining”, In ACM SIGKDD Explorations Newsletter, volume 4, pages 43–48, 2002.
[20] Vaidya, J, Clifton, C., “ Privacy-Preserving Data Mining: Why, How, and When”, IEEE Security and Privacy, 2, 19-27, 2004.
[21] Clifton, C, Kantarcioglu, M, Vaidya,J Lin, X, Zhu, M Y. , “Tools for privacy preserving distributed data mining” , SIGKDD Explor. Newsl., 28-34, 2002.
[22] Weiss G M, “Data Mining in Telecommunications”, In Data Mining and Knowledge Discovery Handbook, A Complete Guide for Practitioners and Researchers. Kluwer Academic Publishers, 1189-1201, 2005.
[23] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar, “On the privacy preserving properties of random data perturbation techniques”, In Proceedings of the 3rd IEEE International Conference on Data Mining, pages 99–106, Melbourne, Florida, November 19-22, 2003.
[24] Muralidhar, K., Parsa, R. and Sarathy, R “ A General Additive Data Perturbation Method for database Security, Management” , 1399-1415,1999.
[25] Li Liu , Murat Kantarcioglu, Bhavani Thuraisingham “The applicability of the perturbation based privacypreserving data mining for real-world data”, Data & Knowledge Engineering 65 (2008) 5–21.
[26] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke. “Privacy preserving mining of association rules” InProceedings of 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, July 2002.
[27] Y. Lindell and B. Pinkas “ Privacy preserving data mining”. In Advances in Cryptology - crypto2000, Lecture Notes in Computer Science, volume 1880, 2000.
[28] J. Vaidya and C. Clifton, “ Privacy preserving association rule mining in vertically partitioned data”. In Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, July 23-26 2002.
[29] Kargupta. H., Datta,.S.,Wang.Q, And Sivakumar. K. “ On the privacy preserving properties of random data perturbation techniques”, Proc. of Intl. Conf. on Data Mining (ICDM) (2003).
[30] Agrawal.D, and aggarwal, C. “On the design and quantification of privacy preserving data mining algorithms”. Proc. of ACM PODS Conference2002.