Hybrid Approach for Privacy Preserving Data Mining Using Piecewise Vector Quantization

[1]JOTHI.R, ANITHA.P, [2]INDUMATHI.P,

[1] Assistant Professor, Department of Computer Applications.

[2] Master of Computer Application, Department of computer applications.

Dhanalakshmi Srinivasan College of Arts and Sciences for women(Autonomous), Perambalur, Tamil Nadu, India

Abstract- In Recent time, the privacy-preserving data-mining (PPDM) has been extremely analyzed, due to the vast eruption of the private information over the internet. Various algorithmic technologies have been observed for the privacy-preserving data-mining. Now a day's detailed personal data from large data bases is regularly collected and analyzed by many applications with data mining, sometimes sharing of these data is beneficial to the application users. In this paper, we propose an Adaptive Privacy Policy Prediction (A3P) system which aims to provide users a hassle free privacy settings experience by automatically generating personalized policies. The A3P system handles user uploaded medical images, and factors in the following criteria that influence one's privacy settings of images. We design the interaction flows between the two building blocks to balance the benefits from meeting personal characteristics and obtaining community advice. This paper also investigates a descriptive analysis of the various approaches for the privacy preserving data mining like Randomization, K-anonymization technique, Perturbation Technique, Cryptographic technique, etc, for the sharing of information and for its privacy.

KEYWORDS

Vector quantization, code book generation, privacy preserving data mining, k-means clustering

I INTRODUCTION

Privacy preserving data mining (is one of the important areas of data mining that aims to provide security for secret information from unsolicited or unsanctioned disclosure. Data mining techniques analyzes and predicts useful information. Analyzing such data may opens treat to privacy .The concept of privacy preserving data mining is primarily concerned with protecting secret data against unsolicited access. It is important because Now a day's Treat to privacy is becoming real since data mining techniques are able to predict high sensitive knowledge from huge volumes of data.

- ✓ What data mining causes is social and ethical problem by revealing the data which should require privacy?
- ✓ Providing security to sensitive data against unauthorized access has been a long term goal for the database security research community and for the government statistical agencies.

- ✓ Hence, the security issue has become, recently, a much more important area of research in data mining.

The concept of privacy preserving data mining means preserving personal information from data mining algorithms. For privacy preservation of data sometimes data is altered before delivery and sometimes it is altered after delivery (before showing it to the third party). Now days we have different data mining applications for which security is a must for example- stock market, finance etc. In data mining system there are so many privacy preserving methods. The objective of privacy preserving data mining is to search some technique in which the original information is transformed in some way so that the private data and private knowledge remain confidential after the mining process. Data cleansing only takes effect on certain type of errors and cannot result in perfect data, eliminating noisy data may lead to information loss. Noise corrupted data can be modified by the use of noise knowledge. Data privacy can also be preserved by random non-linear data transformation.

There are many approaches for privacy preserving data mining. Privacy preserving data mining techniques can be classified based on the following dimensions

1. Data distribution
2. Data modification
3. Data mining algorithm
4. Data or rule hiding
5. Privacy preservation

First dimension refers to data distribution; data can be distributed vertically or horizontally over the systems,

Second dimension refers to modifying the original data to other form, so that we can prevent deidentification of sensitive data, there are several methods are there for data modification like randomization, swapping, sampling, anonymity, blocking,…etc

Third dimension is Data mining algorithm, when mining is performed on data we could be able to preserve privacy of individuals.

Fourth dimension refers to Hiding, part of data or result of data mining, Can be hided.

Fifth dimension is the most important issue i.e, providing privacy during data mining .This paper mainly concentrates on fifth dimension.

II. PROBLEM DEFINITION

Due to the increase in sensitive information in databases privacy preservation is an important concern for each and every data miner. The need for privacy is sometimes due to law (for medical databases) or can be influenced by medical interest. For scientific, economic and market oriented databases confidentiality is an important issue. So it is important to develop such a technique for privacy preservation that the data utility and integrity remains constant without affecting the data confidentiality.

III LITERATURE REVIEW

The term "privacy preserving data mining" was first introduced in papers by Agrawal & Srikant, they worked on Randomization. Lindell and Pinkas introduced a cryptographic protocol for decision tree construction over a dataset horizontally partitioned between two parties.A Protocol for secure association rules (Kantarcioğlu and Clifton), k-means clustering (Lin et al.), k-nn classifiers (Kantarcioğlu and Clifton).Again, secure protocols for the vertically partitioned case have been developed for mining association rules (Vaidya and Clifton), decision trees (Du and Zhan) and k-means clusters (Jagannathan and Wright).

Other areas that influence the development of PPDM include cryptography and secure multiparty computation (Goldreich,Stinson), database query auditing for disclosure detection and prevention (Kleinberg et al.) (Dinur & Nissim) (Kenthapadi et al.), database privacy and policy enforcement (Agrawal et al.) (Aggarwal et al.), database security (Castano et al.), and of course, specific application domains[3][4]. Now a day's privacy preserving data mining is becoming one of the focusing area because data mining predicts more valuable information that may be beneficial to the business, education systems, medical field, political ,…etc.

IV. METHODOLOGY

A. Randomization Technique

In randomization technique the original data is hidden by randomly modifying the data values. It provides some deeper statistical approach to security and privacy. On dataset randomization is applied either by adding or by multiplying random values to original records. Randomization technique is Easy to implement and it has very High search accuracy. It is computationally efficient and Suitable for different user requirements. But it results High information loss.

B.Perturbation Technique

This technique is also known as noise addition technique. Here noise is introduced either to the data or to the result of the queries. Perturbation is done by use of Gaussian and uniform perturbing functions. Here each and every attribute is treated independently but the Data Confidentiality has been compromised sometimes.

C.Cryptographic Technique

In cryptographic technique data value is altered by some encryption technique like secure sum, secure set union, and secure size set intersection etc. on all data or only on confidential data values. The output of a computation is not protected by cryptographic techniques instead it prevents privacy leaks during computation. It provide High search Accuracy but with high Computational complexity.

Cryptographic Technique In cryptographic technique data value is altered by some encryption technique like secure sum, secure set union, and secure size set intersection etc. on all data or only on confidential data values. The output of a computation is not protected by cryptographic techniques instead it prevents privacy leaks during computation. It provide High search Accuracy but with high Computational complexity.

D.The A3P-core Classification

The A3P-core classifies the image and determines whether there is a need to invoke the A3Psocial. In most cases, the A3P-core predicts policies for the users directly based on their historical behavior. If one of the following two cases is verified true, A3P-core will invoke A3Psocial: (i) The user does not have enough data for the type of the uploaded image to conduct policy prediction; (ii) The A3P-core detects the recent major changes among the user's community about their privacy practices along with user's increase of social networking activities (addition of new friends, new posts on one's profile etc.). The A3P system handles user uploaded images, and factors in the following criteria that influence one's privacy settings of images. We design the interaction flows between the two building blocks to balance the benefits from meeting personal characteristics and obtaining community advice. Data is mined to anticipate behavior patterns and trends.

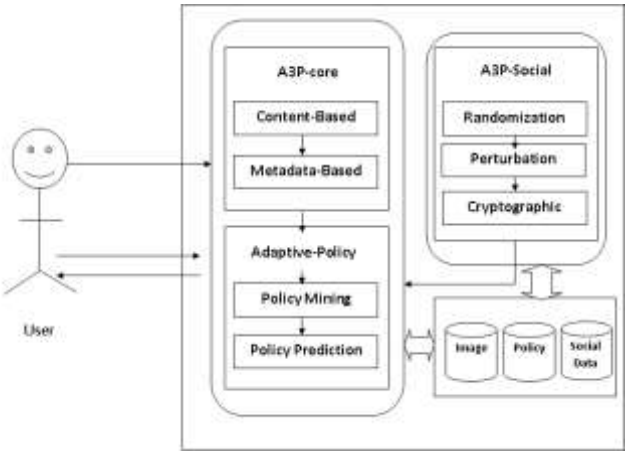

Figure 1: The Hybrid A3P Architecture

E.Content-Based Classification

Our approach to content-based classification is based on an efficient and yet accurate image similarity approach. Specifically, our classification algorithm compares image signatures defined based on quantified and sanitized version of Haar wavelet transformation. For each image, the wavelet transform encodes frequency and spatial information related to image color, size, invariant transform, shape, texture, symmetry, etc. Then, a small number of coefficients are selected to form the signature of the image. The content similarity among images is then determined by the distance among their image signatures.

F.Metadata-Based Classification

The metadata considered in our work are tags, captions, and comments. The second step is to derive a representative hypernym (denoted as h) from each metadata vector. The third step is to find a subcategory that an image belongs to. This is an incremental procedure. At the beginning, the first image forms a subcategory as itself and the representative hypernyms of the image becomes the subcategory's representative hypernyms.

G.Adaptive Policy Prediction

The policy prediction algorithm provides a predicted policy of a newly uploaded image to the user for his/her reference. More importantly, the predicted policy will reflect the possible changes of a user's privacy concerns.

V CONCLUSION

In this paper, privacy issues of dataset have been described and the new approach for

the privacy-preservation has also been suggested. The randomization and mod approaches are utilized to hide the sensitive data. Data randomization approach is implemented such that the private details cannot be introduced by data-mining approaches. Vector-quantization is the recent technique for the privacy-preserving data-mining, upon implementing this encoding-procedure one may not leak the actual data therefore the privacy is got preserved

VI FUTURE WORK

As future work new and effective quantization method can be used rather than K means approach that we have used. K nearest neighbor approach is one of the approach which can give better result in more work in the field of fuzzy Dataset, Mobility of different Dataset, The development of uniform framework for various privacy preserving across all data mining algorithms.

REFERENCES

[1] Agrawal, R., Srikant, R.: Privacy Preserving Data Mining. In: Proc. of ACM SIGMOD Conference on Management of Data

[2] Evfimievski, A., Grandison, T.: Privacy Preserving Data Mining. IBM Almaden Research Center 650 Harry Road, San Jose, California 95120, USA

[3] Agarwal Charu, C., Yu Philip, S.: Privacy Preserving Data Mining: Models and Algorithms. Springer, New York

[4] Oliveira, S.R.M., Zaiane Osmar, R.: A Privacy-Preserving Clustering Approach Toward Secure and Effective Data Analysis for Business Collaboration. In: Proceedings of the International Workshop on Privacy and Security Aspects of Data Mining in Conjunction with ICDM.

[5] Somasundaram, K., Vimala, S.: Codebook Generation for Vector Quantization with Edge Features. CiiT International Journal of Digital Image Processing

[6] Manish Sharma, AtulChaudhary, Manish Mathuria, ShaliniChaudhary and Santosh Kumar, An Efficient Approach for Privacy Preserving in Data Mining, International Conference on Signal Propagation and Computer Technology

[7] Madhavi Latha, M., Satya Sai Ram, M., Siddaiah, P.: Multi Switched Split Vector Quantization. Proceedings of World Academy of Science, Engineering and Technology

[8] Arunadevi M, Anuradha R (2014) Privacy preserving outsourcing for frequent itemset mining. Int J Innov Res Comp Commun Eng

[9] LidanShou, He Bai, Ke Chen, and Gang Chen "Supporting Privacy Protection in Personalized Web Search", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING,

[10] Madhavi Latha, M., Satya Sai Ram, M., Siddaiah, P.: Multi Switched Split Vector Quantization. Proceedings of World Academy of Science, Engineering and Technology..