National Conference on Recent Research in Engineering and Technology (NCRRET-2015)
International Journal of Advance Engineering and Research Development (IJAERD)
e-ISSN: 2348 - 4470 , print-ISSN:2348-6406

A Review on Privacy Preserving Data Mining Approaches

Anu Thomas
Asst.Prof. Computer Science & Engineering Department
DJMIT,Mogar,Anand
Gujarat Technological University
Anu.thomas@djmit.ac.in

Jimesh Rana
Asst.Prof. Computer Science &Engineering Department
DJMIT,Mogar,Anand
Gujarat Technological University
Jimesh.rana@djmit.ac.in

***Abstract* - The field of privacy has seen rapid advances in recent years because of the increase in the ability to store data. In particular, recent advances in the data mining field have lead to increased concerns about privacy. While the topic of privacy has been traditionally studied in the context of cryptography and information hiding, recent emphasis on data mining has lead to renewed interest in the field.**

A fruitful direction for future data mining research will be the development of techniques that incorporate privacy concerns. Specially, we address the following question. Since the primary task in data mining is the development of models about aggregated data, can we develop accurate models without access to precise information in individual data records? We consider the concrete case of building a decision-tree classier from training data in which the values of individual records have been perturbed. The resulting data records look very different from the original records and the distribution of data values is also very different from the original distribution. While it is not possible to accurately estimate original values in individual data records, we propose a novel reconstruction procedure to accurately estimate the distribution of original data values. By using these reconstructed distributions, we are able to build classifiers whose accuracy is comparable to the accuracy of classifiers built with the original data.

I. Introduction

This document is the problem of privacy-preserving data mining has become more important in recent years because of the increasing ability to store personal data about users, and the increasing sophistication of data mining algorithms to leverage this information. A number of techniques such as randomization and k-anonymity have been suggested in recent years in order to perform privacy-preserving data mining. Furthermore, the problem has been discussed in multiple communities such as the database community, the statistical disclosure control community and the cryptography community. In some cases, the different communities have explored parallel lines of work which are quite similar. This book will try to explore different topics from the perspective of different communities, and will try to give a fused idea of the work in different communities. The key directions in the field of privacy-preserving data mining are as follows:

A. Privacy-Preserving Data Publishing

These techniques tend to study different transformation methods associated with privacy. Another related issue is how the perturbed data can be used in conjunction with classical data mining methods such as association rule mining. Other related problems include that of determining privacy-preserving methods to keep the underlying data useful (utility-based methods), or the problem of studying the different definitions of privacy, and how they compare in terms of effectiveness in different scenarios. Changing the results of Data Mining Applications to preserve privacy: In many cases, the results of data mining applications such as association rule or classification rule mining can compromise the privacy of the data. This has spawned a field of privacy in which the results of data mining algorithms such as association rule mining are modified in order to preserve the privacy of the data. A classic example of such techniques are association rule hiding methods, in which some of the association rules are suppressed in order to preserve privacy.

B. *Query Auditing*

Such methods are akin to the previous case of modifying the results of data mining algorithms. Here, we are either modifying or restricting the results of queries. Cryptographic Methods for Distributed Privacy: In many cases, the data may be distributed across multiple sites, and the owners of the data across these different sites may wish to compute a common function. In such cases, a variety of cryptographic protocols may be used in order to communicate among the different sites, so that secure function computation is possible without revealing sensitive information.

C. Theoretical Challenges in High Dimensionality

Real data sets are usually extremely high dimensional, and this makes the process of privacy preservation extremely difficult both from a computational and effectiveness point of view. It has been shown that optimal k-anonymization is NP-hard. Furthermore, the technique is not even effective with increasing dimensionality, since the data can typically be combined with either public or background information to reveal the identity of the underlying record owners.

II. Privacy-Preserving Methods

The basic approach to preserving privacy is to let users provide a modified value for sensitive attributes. The modified value may be generated using custom code, a browser plug-in, or extensions to products. In this method, the values for an attribute are partitioned into a set of disjoint, mutually-exclusive classes. We consider the special case of discretization in which values for an attribute are discretized into intervals. All intervals need not be of equal width. For example, salary may be discretized into 10K intervals for lower values and 50K intervals for higher values. Instead of a

true attribute value, the user provides the interval in which the value lies. Discretization is the method used most often for hiding individual values. Value Distortion Return a value xi + r instead of xi where r is a random value drawn from some distribution. We consider two random distributions.

A. Uniform

The random variable has a uniform distribution, between [-α; + α]. The mean of the random variable is 0.

B. Gaussian

The random variable has a normal distribution, with mean μ= 0 and standard deviation.

III. Quantifying Privacy

For quantifying privacy provided by a method, we use a measure based on how closely the original values of a modified attribute can be estimated. If it can be estimated with c% confidence that a value x lies in the interval [x1; x2], then the interval width (x2 > x1) defines the amount of privacy at c% confidence level. Table 1 shows the privacy offered by the different methods using this metric. We have assumed that the intervals are of equal width W in discretization. Clearly, for 2α=W, Uniform and Discretization provide the same amount of privacy. As α increase, privacy also increases. To keep up with Uniform, Discretization will have to increase the interval width, and hence reduce the number of intervals. Hence Discretization will lead to poor model accuracy compared to Uniform since all the values in interval are modified to the same value. Gaussian provides significantly more privacy at higher confidence levels compared to the other two methods.

	Confidence		
	50%	95%	99.9%
Discretization	$0.5 \times W$	$0.95 \times W$	$0.999 \times W$
Uniform	$0.5 \times 2\alpha$	$0.95 \times 2\alpha$	$0.999 \times 2\alpha$
Gaussian	$1.34 \times \sigma$	$3.92 \times \sigma$	$6.8 \times \sigma$

Table 1: Privacy Metrics

IV. Application Scenarios

A. Surveys and Data Collection

Companies collect personal preferences of their customers for targeted product recommendations, or conduct surveys for business planning; political parties conduct opinion polls to adjust their strategy. The coverage of such data collection may significantly increase if all respondents are aware that their privacy is provably protected, also eliminating the bias associated with evasive answers.

B. Monitoring for Emergencies

Early detection of large-scale abnormalities with potential implications for public safety or national security is important in protecting our well-being. Disease outbreaks, environmental disasters, terrorist acts, and manufacturing accidents can often be detected and contained before they endanger a large population. The first indication of an Impending disaster can be difficult to notice by looking at any individual case, but is easy to see using data mining: an unusual increase in certain health symptoms or non prescription drug purchases, a surge in car accidents, a change in online traffic pattern, and so forth.

C. Product Traceability

Before a product (e.g. car or a drug) reaches its end user, it usually passes through a long chain of processing steps, such as manufacturing, packaging, transportation, storage, and sale. In the near future, many products and package units will carry radio frequency identification (RFID) tag and will be automatically registered at every processing step. This will create a vast distributed collection of RFID traces, which can be mined to detect business patterns, market trends, inefficiencies and bottlenecks, criminal activity such as theft and counterfeiting, and so on.

D. Medical Research

Personal health records are one of the most sensitive types of private data; their privacy standards have been codified into law in many countries, for example HIPAA (Health Insurance Portability and Accountability Act) in the United States (Office for Civil Rights [OCR], 2003). On the other hand, data mining over health records is vital for medical, pharmaceutical, and environmental research. For example, a researcher may want to study the effect of a certain gene A on an adverse reaction to drug B. However, due to privacy concerns, the DNA sequences and the medical histories are stored at different data repositories and cannot be brought together. Then, PPDM over vertically partitioned data can be used to compute the aggregate counts while preserving the privacy of records.

V. Future Trends

The main technical challenge for PPDM is to make its algorithms scale and achieve higher accuracy while keeping the privacy guarantees. The known proof techniques and privacy definitions are not yet flexible enough to take full advantage of existing PPDM approaches. Adding a minor assumption (from the practical viewpoint) may slash the computation cost or allow much better accuracy if the PPDM methodology is augmented to leverage this assumption. On the other hand, proving complexity lower bounds and accuracy upper bounds will expose the theoretical limits of PPDM.

VI. Conclusions

Privacy-preserving data mining emerged in response to two equally important (and seemingly disparate) needs: data analysis in order to deliver better services and ensuring the privacy rights of the data owners. Difficult as the task of addressing these needs may seem, several tangible efforts have been accomplished. In this article, an overview of the

popular approaches for doing PPDM was presented, namely,

suppression, randomization, cryptography, and summarization. The privacy guarantees, advantages, and disadvantages of each approach were stated in order to provide a balanced view of the state of the art.

VII. The Broad Areas of Privacy are as Follows

A. Privacy-preserving data publishing

This corresponds to sanitizing the data, so that its privacy remains preserved.

B. Privacy-Preserving Applications

This corresponds to designing data management and mining algorithms in such a way that the privacy re-mains preserved. Some examples include association rule mining, classification, and query processing.

C. Utility Issues

Since the perturbed data may often be used for mining and management purposes, its utility needs to be preserved. Therefore, the data mining and privacy transformation techniques need to be designed effectively, so to to preserve the utility of the results.

D. Distributed Privacy, cryptography and adversarial collaboration

This corresponds to secure communication protocols between trusted parties, so that information can be shared effectively without revealing sensitive information about particular parties.

REFERENCES

[1] Agrawal R., Srikant R. Privacy-Preserving Data Mining. Proceedings of the ACM SIGMOD, Conference, 2000.
[2] Aggarwal C. C. On k-anonymity and the curse of dimensionality. VLDB Conference, 2005.
[3] Agrawal, R., & Srikant, R. (2000). Privacy preserving data mining. Proceedings of ACM SIGMOD Conference on Management of Data (SIGMOD'00), Dallas, TX.
[4] Mrs. Kiran M. Jha, Mehul Barot, Privacy Preserving Data Mining, International Journal of Futuristic Trends in Engineering and Technology.