ISSN 2229-5518

Privacy of Data, Preserving in Data Mining

Deepika Saxena

Abstract — Huge volume of detailed personal data is regularly collected and sharing of these data is proved to be beneficial for data mining application. Such data include shopping habits, criminal records,medical history, credit records etc .On one hand such data is an important asset to business organization and governments for decision making by analyzing it .On the other hand privacy regulations and other privacy concerns may prevent data owners from sharing information for data analysis. In order to share data while preserving privacy data owner must come up with a solution which achieves the dual goal of privacy preservation as well as accurate clustering result. Trying to give solution for this we implemented vector quantization approach piecewise on the datasets which segmentize each row of datasets and quantization approach is performed on each segment using K means which later are again united to form a transformed data set.Some experimental results are presented which tries to finds the optimum value of segment size and quantization parameter which gives optimum in the tradeoff between clustering utility and data privacy in the input dataset

Keywords—datamining, data privacy, preserving datamining, model, quantization approach, predictive information, ppdp, ppdm.

1 Introduction

Data mining is a technique that deals with the extraction of hidden predictive information from large database. It uses sophisticated algorithms for the process of sorting through large amounts of data sets and picking out relevant information. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. With the amount of data doubling each year, more data is gathered and data mining is becoming an increasingly important tool to transform this data into information. Long process of research and product development evolved data mining. This evolution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time. Data mining takes this Evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery. Data mining is ready for application in the business community because it is supported by three technologies that are now sufficiently mature:

- Massive data collection
- Powerful multiprocessor computers
- Data mining algorithms

Scope of data mining

Data mining gets its name from the similarities between finding for important business information in a huge database for example, getting linked products in gigabytes of store scanner data and mining a mountain for a vein of valuable ore. These processes need either shifting through an large amount of material, or intelligently searching it to find exactly where the value resides. Data mining technology can produce new business opportunities by providing these features in databases of sufficient size and quality, automated prediction of trends and behaviors. The process of finding predictive information in large databases is automated by data mining. Questions that required extensive analysis traditionally can now be answered directly from the data quickly with data mining technique. A typical example is targeted marketing. It uses data on past promotional mailings to recognize the targets most likely to maximize return on investment in future mailings. Other predictive problems include forecasting bankruptcy and other forms of default, and identifying segments of a population likely to respond similarly to given events.· Automated discovery of previously unknown patterns. Data mining tools analyze databases and recognize previously hidden patterns in one step. The analysis of retail sales data to recognize seemingly unrelated products that are often purchased together is an example of pattern discovery. Other pattern discovery problems include detecting fraudulent credit card transactions and identifying data that are anomalous that could represent data entry keying errors. Data mining techniques can provide the features of automation on existing software and hardware platforms, and can be implemented on new systems as existing platforms are upgraded and new products developed. On high performance parallel processing systems when data mining tools are used, they can analyze huge databases in minutes. Users can automatically experiment with more models to understand complex data by using faster processing facility. High speed makes it possible for users to analyze huge quantities of data. Larger databases, in turn, yield improved predictions.

Applications of data mining

There is a rapidly growing body of successful applications in a wide range of areas as diverse as:

- Analysis of organic compounds, automatic abstracting, credit card fraud detection, financial forecasting, medical diagnosis etc. Some examples of applications (potential or actual) are:

☞A supermarket chain mines its customer transactions data to optimize targeting of high value customers
☞A credit card company can use its data warehouse of customer transactions for fraud detection
☞A major hotel chain can use survey databases to identify attributes of a 'high-value' prospect.

Applications can be divided into four main types:
1 Classification
2 Numerical prediction
3 Association
4 Clustering.

Data mining using labeled data (specially designated attribute) is called supervised learning. Classification and numerical prediction applications falls in supervised learning. Data mining which uses unlabeled data is termed as unsupervised learning and association and clustering falls in this category.

Data mining and Privacy

Data mining deals with large database which can contain sensitive information. It requires data preparation which can uncover information or patterns which may compromise confidentiality and privacy obligations. Advancement of efficient data mining technique has increased the disclosure risks of sensitive data. A common way for this to occur is through data aggregation. Data aggregation is when the data are accrued, possibly from various sources, and put together so that they can be analyzed. This is not data mining per se, but a result of the preparation of data before and for the purposes of the analysis. The threat to an individual's privacy comes into play when the data, once compiled, cause the data miner, or anyone who has access to the newly compiled data set, to be able to identify specific individuals, especially when originally the data were anonymous.
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. Therefore, in recent years, privacy-preserving data mining has been studied extensively. We will further see the research done in privacy area .In chapter 3 general survey of privacy preserving methods used in data mining is presented.

PRIVACY-PRESERVING DATA MINING

The recent work on PPDM has studied novel data mining techniques that do not require accessing sensitive information. The general idea of PPDM is to allow data mining from a modified version of the data that contains no sensitive information.

CENTRALIZED MODEL

In the centralized model, all data are owned by a single data publisher. The key issues are how to modify the data and how to recover the data mining result from the modified data. Answers often depend on data mining operations and algorithms. One common technique is randomization, by introducing random noise and swapping values in the data. The randomized data preserves aggregate properties (such as means and correlations) in the collection of records, but has little use when each record is examined individually. Another common technique is encryption. The data publisher transforms the original data into an encrypted form for data mining at an external party. Since the data mining results are in the encrypted form and since the data publisher is the only one who can decrypt the results, this approach is applicable only if the data publisher himself is the user of data mining results.

DISTRIBUTED MODEL

In the distributed model, multiple data publishers want to conduct a computation based on their private inputs, but no data publisher is willing to disclose its own output to anybody else. The privacy issue here is how to conduct such a computation while preserving the privacy of the inputs. This problem is known as the Secure Multiparty Computation (SMC) problem. The aim of SMC is to enable multiple parties to carry out distributed computing tasks in a secure manner with the assumption that some attackers, who possibly are the participating parties themselves, want to obtain extra information other than the output. SMC has two major requirements, privacy and correctness. The privacy requirement states that parties should learn their output and nothing else during and after the SMC process. The correctness requirement states that each party should receive its correct output without alteration by the attackers. Extensive research has been conducted on secure protocols for data mining tasks including association rule mining, classification analysis, and clustering analysis. Refer to for surveys on this distributed model of PPDM.

COMPARING PPDP AND PPDM

In many real life applications, the data publisher wants to *publish* some data, but has little or no interest in data mining results and algorithms. For example, a hospital may publish the patient data to a drug research institute; although willing to contribute its data to drug research, the hospital is not interested in and has no expertise in data mining algorithms because drug research is not its normal business. This privacy-preserving data publishing (PPDP) scenario differs from PPDM in several major ways. PPDP focuses on the data, not data mining results; therefore, published records should be meaningful when examined individually. This implies that randomization and encryption are inapplicable. PPDP seeks to anonymize the data by hiding the identity of individuals, not hiding sensitive data. The anonymized data is expected to be analyzed by traditional data mining techniques; therefore, no new data mining techniques are needed. We did not intend to dismiss the contribution of the randomization and encryption approaches. They are effective anonymization methods if the data records will not be examined individually.

CONFIDENCE BOUNDING

solution is possible. In particular, we iteratively disclose domain values in a top-down
manner by suppressing all domain values. In each iteration, we disclose the suppressed domain value to maximize some criterion taking into account both information gained and privacy lost. We evaluate this method on real life data sets. Several features make this approach practically useful:

☞*No taxonomy required*. Suppression replaces a domain value with *?* without requiring
a taxonomy of values. This is a useful feature because most data do not have an
associated taxonomy, though taxonomies may exist in certain specialized domains.

☞ *Preserving the truthfulness of values*. The special value *?* represents the union," a
less precise but truthful representation, of suppressed domain values. This truthfulness
is useful for reasoning and explaining the classification model.

☞*Subjective notion of privacy*. The data publisher has the exibility to her own notion of privacy using templates for sensitive inferences.

☞ *Excient computation*. It operates on simple but effective data structures to reduce
the need for accessing raw data records.

☞ *Anytime solution*. At any time, the user (the data publisher) can terminate the com-
putation and have a table satisfying the privacy goal.

☞*Extendibility*. Though we focus on categorical attributes and classification analysis,
this work can be easily extended to continuous attributes and other information utility
criteria.

CONCLUSIONS

Due to the wide use of the Internet and the trends of enterprise integration, one-stop ser-
vice, simultaneous cooperation and competition, and outsourcing in both public and private sectors, data publishing has become a daily and routine activity of individuals, companies, organizations, government agencies. Privacy-preserving data publishing is a promising approach for data publishing without compromising individual privacy or disclosing sensitive information. In this thesis, we studied different types of linking attacks in the data publishing scenarios of single release, sequential release, and secure data integration . Our contributions can be summarized as follows:

☞ *Preserving Privacy and Information.* We considered the problem of protecting in-
dividual privacy while releasing person-specific data for classification modeling. We
chose classification analysis as the information requirement because the data quality
and usefulness can be objectively measured. Our proposed framework can easily adopt
other information requirement with a different selection criterion.

☞*A Uni¯ed Privacy Notion.* We demeaned a new privacy notion, called privacy template in the form of *hX; Y; ki*, that unique anonymity template and confidentiality template.This unified notion is applicable to all data publishing scenarios studied in this thesis.

☞*A Framework of Anonymization Algorithm.* Despite the data publishing scenarios are
very di®erent, we presented a framework of anonymization algorithm, Top-Down Re-
¯nement (TDR), to iteratively specialize the data from a general state into a special
state, guided by maximizing the information utility and minimizing the privacy speci-

¯city. This top-down approach serves a natural and e±cient structure for handling
categorical and continuous attributes and multiple privacy templates. Experiments
suggested that our TDR framework effectively preserves both information utility and
individual privacy and scales well for large data sets in different data publishing sce-
narios.

☞ *Extended Data Publishing Scenarios.* Most existing works considered the simplest data publishing scenario, that is, a single release from a single publisher. Such mechanisms are insufficient because they only protect the data up to the release or the cipient. Therefore, we also extended the privacy notion and anonymization framework
to other real life data publishing scenarios, including sequential release publishing and
Secure data integration.

REFERENCES:-

A. Benjamin C. M. Fung On PRIVACY-PRESERVING DATA PUBLISHING. Benjamin C. M. Fung 2007 SIMON FRASER UNIVERSITY Summer 2007.

[1] C. C. Aggarwal. On *k*-anonymity and the curse of dimensionality. In *Proc. of the 31st*
International Conference on Very Large Data Bases (VLDB), pages 901{909, Trond-
heim, Norway, 2005.

[2] C. C. Aggarwal, J. Pei, and B. Zhang. On privacy preservation against adversarial data
mining. In *Proc. of the 12th ACM SIGKDD International Conference on Knowledge*
Discovery and Data Mining, Philadelphia, PA, August 2006.

[3] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and
A. Zhu. Anonymizing tables. In *Proc. of the 10th International Conference on Database*
Theory (ICDT), pages 246{258, Edinburgh, UK, January 2005.

[4] R. Agrawal, A. Ev¯mievski, and R. Srikant. Information sharing across private
databases. In *Proc. of the 2003 ACM SIGMOD International Conference on Man-*
agement of Data, San Diego, CA, 2003.

[5] R. Agrawal, T. Imielinski, and A. N. Swami. Mining association rules between sets of
items in large datasets. In *Proc. of the 1993 ACM SIGMOD*, pages 207{216, 1993.

[6] R. Agrawal and R. Srikant. Privacy preserving data mining. In *Proc. of the 2000 ACM*
SIGMOD International Conference on Management of Data, pages 439{450, Dallas,
Texas, May 2000.

[7] S. Agrawal and J. R. Haritsa. A framework for high-accuracy privacy-preserving mining.
In *Proc. of the 21st IEEE International Conference on Data Engineering (ICDE)*, pages
193{204, Tokyo, Japan, 2005.

[8] R. J. Bayardo and R. Agrawal. Data privacy through optimal k-anonymization. In
Proc. of the 21st IEEE International Conference on Data Engineering (ICDE), pages
217{228, Tokyo, Japan, 2005.

[9] L. Burnett, K. Barlow-Stewart, A. Pros, and H. Aizenberg. The gene trustee: A uni-
versal identi¯cation system that ensures privacy and con¯dentiality for human genetic
databases. *Journal of Law and Medicine*, 10:506{513, 2003.

[10] Business for Social Responsibility. BSR Report on Privacy, 1999. http://www.bsr.org/.
123
BIBLIOGRAPHY 124

[11] D. Chaum. Untraceable electronic mail, return addresses, and digital pseudonyms.
Communications of the ACM, 24(2):84{88, 1981.

[12] S. Chawathe, H. G. Molina, J. Hammer, K. Ireland, Y. Papakonstantinou, J. Ull-
man, and J. Widom. The TSIMMIS Project: Integration of heterogeneous information
sources. In *16th Meeting of the Information Processing Society of Japan*, pages 7{18,
1994.

[13] C. Clifton. Using sample size to limit exposure to data mining. *Journal of Computer*
Security, 8(4):281{307, 2000.

[14] C. Clifton, M. Kantarcioglu, J. Vaidya, X. Lin, and M. Y. Zhu. Tools for privacy
preserving distributed data mining. *SIGKDD Explorations*, 4(2), December 2002.

[15] L. H. Cox. Suppression methodology and statistical disclosure control. *Journal of the*
American Statistics Association, Theory and Method Section, 75:377{385, 1980.

[16] T. Dalenius. Finding a needle in a haystack - or identifying anonymous census record.
Journal of O±cial Statistics, 2(3):329{336, 1986.

[17] U. Dayal and H. Y. Hwang. View de¯nition and generalization for database integration
in a multidatabase systems. *IEEE Transactions on Software Engineering*, 10(6):628{
645, 1984.
[18] A. Deutsch and Y. Papakonstantinou. Privacy in database publishing. In *ICDT*, 2005.
[19] W. Du, Y. S. Han, and S. Chen. Privacy-preserving multivariate statistical analysis:
Linear regression and classi¯cation. In *Proc. of the SIAM International Conference on
Data Mining (SDM)*, Florida, 2004.
[20] W. Du and Z. Zhan. Building decision tree classi¯er on private data. In *Workshop
on Privacy, Security, and Data Mining at the 2002 IEEE International Conference on
Data Mining*, Maebashi City, Japan, December 2002.