Volume : 2 | Issue : 8 | Aug 2013 • ISSN No 2277 - 8160

Research Paper | Computer Science

Security And Privacy Challenges in Data Mining

Prof. Aashish H Kacha

Suraj Muchhala Polytechnic

ABSTRACT

given the rising privacy concerns, the data mining community has faced a new challenge. Having shown how effective its tools are in revealing the knowledge locked within huge databases, it is now required to develop methods that restrain the power of these tools to protect the privacy of individuals. The question how these two contrasting goals, mining new knowledge while protecting individuals' privacy, can be reconciled, is the focus of this research. We seek ways to improve the tradeoff between privacy and utility when mining data.

KEYWORDS: K-Anonymity, Binary class, Mining,

Introduction

In recent years, privacy preserving data mining has emerged as a very active research area. This field of research studies how knowledge can be extracted from large data stores while maintaining commercial or legislative privacy constraints. Quite often, these constraints pertain to individuals represented in the data stores. While data collectors strive to derive new insights that would allow them to improve customer service and increase their sales, consumers are concerned about the vast quantities of information collected about them and how this information is put to use. Privacy preserving data mining aims to settle these conflicting interests. The question how these two contrasting goals, mining new knowledge while protecting individuals' privacy, can be reconciled, is the focus of this research. We seek ways to improve the tradeoff between privacy and utility when mining data.

Data, Information, and Knowledge

Data are any facts, numbers, or text that can be processed by a computer. Today, organizations are accumulating vast and growing amounts of data in different formats and different databases. This includes:

- operational or transactional data such as, sales, cost, inventory, payroll, and accounting
- nonoperational data, such as industry sales, forecast data, and macro economic data
- meta data - data about the data itself, such as logical database design or data dictionary definitions

Information

The patterns, associations, or relationships among all this *data* can provide *information*. For example, analysis of retail point of sale transaction data can yield information on which products are selling and when.

Knowledge

Information can be converted into *knowledge* about historical patterns and future trends. For example, summary information on retail supermarket sales can be analyzed in light of promotional efforts to provide knowledge of consumer buying behavior.

Literature Review

Data Mining

The primary ingredient of any Data Mining exercise is the database. A *database* is an organized and typically large collection of detailed facts concerning some domain in the outside world. The aim of Data Mining is to examine this database for regularities that may lead to a better understanding of the domain described by the database. In Data Mining we generally assume that the database consists of a collection of *individuals*. Depending on the domain, individuals can be anything from customers of a bank to predict the behavior of new individuals. Consider, for example, a sample of customers of a bank and how they responded to a certain offer. We can build a model describing how the response depends on different characteristics of the customers, with the aim of predicting how other customers will respond to the offer. A lot of time and effort can thus be saved by only approaching customers with a predicted interest.

K-Anonymity

One definition of privacy which has received a lot of attention in the past decade is that of k-anonymity. The guarantee given by k-anonymity is that no information can be linked to groups of less than k individuals. The k-anonymity model of privacy was studied intensively in the context of public data releases, when the database owner wishes to ensure that no one will be able to link information gleaned from the database to individuals from whom the data has been collected. To be of any practical value, the definition of privacy must satisfy the needs of users of a reasonable application. Two examples of such applications are (1) a credit giver, whose clientele consists of numerous shops and small businesses, and who wants to provide them with a classifier that will distinguish credit-worthy from credit-risky clients, and (2) a medical company that wishes to publish a study identifying clusters of patients who respond differently to a course of treatment. These data owners wish to release data mining output, but still be assured that they are not giving away the identity of their clients. If it could be verified that the released output withstands limitations similar to those set by k-anonymity, then the credit giver could release a k-anonymous classifier and reliably claim that the privacy of individuals is protected. Likewise, the authors of a medical study quoting k-anonymous cluster centroids could be sure that they comply with HIPAA privacy standards, which forbid the release of individually identifiable health information.

The past two decades has seen a dramatic increase in the amount of information or data being stored in electronic format. This accumulation of data has taken place at an explosive rate. It has been estimated that the amount of information in the world doubles every 20 months and the size and number of databases are increasing even faster. The increase in use of electronic data gathering devices such as point-of-sale or remote sensing devices has contributed to this explosion of available data. Figure 1 from the Red Brick company illustrates the data explosion.

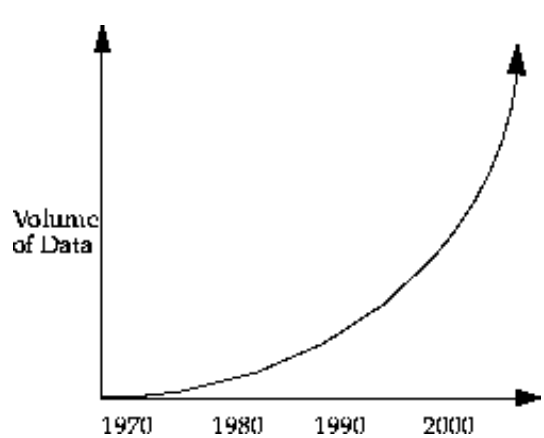

Figure 1: The Growing Base of Data

3. Need of Study

Given the rising privacy concerns, the data mining community has

faced a new challenge. Having shown how effective its tools are in revealing the knowledge locked within huge databases, it is now required to develop methods that restrain the power of these tools to protect the privacy of individuals. The question how these two contrasting goals, mining new knowledge while protecting individuals' privacy, can be reconciled, is the focus of this research. We seek ways to improve the tradeoff between privacy and utility when mining data.

To illustrate this problem, I present it in terms of Pareto efficiency. Consider three objective functions: the accuracy of the data mining model (e.g., the expected accuracy of a resulting classifier, estimated by its performance on test samples), the size of the mined database (number of training samples), and the privacy requirement, represented by a privacy parameter. In a given situation, one or more of these factors may be fixed: a client may present a lower acceptance bound for the accuracy of a classifier, the database may contain a limited number of samples, or a regulator may pose privacy restrictions. Within the given constraints, I wish to improve the objective functions: achieve better accuracy with fewer learning examples and better privacy guarantees. However, these objective functions are often in conflict. For example, applying stronger privacy guarantees could reduce accuracy or require a larger dataset to maintain the same level of accuracy. Instead, we should settle for some tradeoff. With this perception in mind, I can evaluate the performance of data mining algorithms. Consider, for example, three hypothetical algorithms that produce a classifier. Assume that their performance was evaluated on datasets with 50,000 records, with the results illustrated in Figure 2.

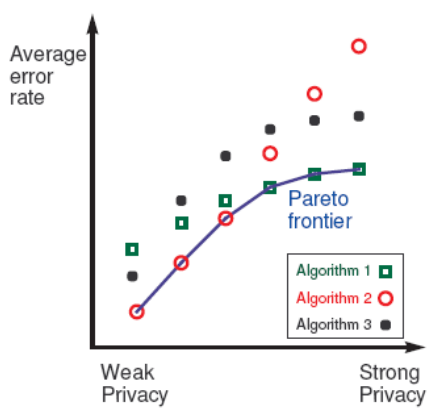

Figure 2: Example of a Pareto frontier. Given a number of learning samples, what are the privacy and accuracy tradeoffs?

We can see that when the privacy settings are high, algorithm 1 obtains on average a lower error rate than the other algorithms, while algorithm 2 does better when the privacy settings are low. A Pareto improvement is a change that improves one of the objective functions without harming the others. Algorithm3 is dominated by the other algorithms: for any setting, we can make a Pareto improvement by switching to one of the other algorithms. A given situation (a point in the graph) esparto efficient when no further Pareto improvements can be made. The Pareto frontier is given by all the Pareto efficient points. My goal is to investigate algorithms that can further extend the Pareto frontier, allowing for better privacy and accuracy tradeoffs.

Implementation

Considering privacy and utility

In many k-anonymity works the anonymization process is guided by utility metrics, regardless of the actual data mining algorithm to be executed on theanonymized data. In the context of differential privacy, the PINQ framework was suggested as a programming interface that provides access to data while enforcing privacy constraints. In theory, PINQ should allow a programmer to write privacy preserving algorithms without requiring privacy expert knowledge. The PINQ layer enforces differential privacy, and the programmer gains a considerable amount of exibilityin designing privacy preserving algorithms. Unfortunately, a data mining algorithm can be implemented in several ways on top of this interface, and accuracy may vary considerably between these implementations.

In contrast, we argue that to improve the tradeoff between privacy and utility, these two goals should be considered simultaneously within a single process. In the context of k-anonymity, It will be showmen how privacy considerations can be interleaved within the execution of a data mining algorithm, allowing to switch rapidly between utility oriented decisions and privacy-oriented decisions. For example, when inducing decision trees, a splitting criterion (utility) is used to pick an attribute to split a node. If this would result in a breach of k-anonymity, the attribute is generalized (privacy) and the algorithm will re-evaluate (utility) the candidate attributes to make a new decision. This kind of interaction between utility and privacy considerations is not possible when the anonymization and data mining processes are distinct. For differential privacy we demonstrated that it is not only important what the calculated functionality is, but also how it is calculated. For example, a splitting criterion for decision tree induction, such as information gain, can be evaluated in several ways on top of a privacy preserving data interface, and choosing a good implementation is crucial to the effectiveness of the resulting algorithm. In addition, when choosing the data mining algorithm, the data miner should balance utility considerations with privacy considerations. Functionalities that are comparable in terms of utility may have a very different privacy impact. For example, the Max, Information Gain and Gini Index criteria for choosing an attribute to split a decision tree node provide decision trees with comparable accuracy when no privacy considerations are involved. However, in a privacy preserving algorithm, privacy considerations should be taken into account as well. The Max criterion for choosing an attribute has low sensitivity, so it has an advantage over the other criteria, especially when working on small data sets or with a small privacy budget. On the other end, Gini Index and Information Gain tend to generate shallower trees than the Max criterion, so given depth constraints on the induced decision tree; they may outperform a differentially private decision tree generated with the Max criterion. Hence the utility and privacy considerations should both be taken into account to obtain the best tradeoff.

How does data mining work?

While large-scale information technology has been evolving separate transaction and analytical systems, data mining provides the link between the two. Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries. Several types of analytical software are available: statistical, machine learning, and neural networks. Generally, any of four types of relationships are sought:

- Classes: Stored data is used to locate data in predetermined groups. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order. This information could be used to increase traffic by having daily specials.
- Clusters: Data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to identify market segments or consumer affinities.
- Associations: Data can be mined to identify associations. The beer-diaper example is an example of associative mining.
- Sequential patterns: Data is mined to anticipate behavior patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.

Data Mining: Issues

One of the key issues raised by data mining technology is not a business or technological one, but a social one. It is the issue of individual privacy. Data mining makes it possible to analyze routine business transactions and glean a significant amount of information about individuals buying habits and preferences.

Another issue is that of data integrity. Clearly, data analysis can only be as good as the data that is being analyzed. A key implementation challenge is integrating conflicting or redundant data from different sources. For example, a bank may maintain credit cards accounts on several different databases. The addresses (or even the names) of a single cardholder may be different in each. Software must translate data from one system to another and select the address most recently entered.

REFERENCES

1. Charu C. Aggarwal. On k-anonymity and the curse of dimensionality. In VLDB,pages 901{909, 2005. | 2. Charu C. Aggarwal and Philip S. Yu. A condensation approach to privacy preserving data mining. In EDBT, pages 183{199, 2004. | 3. Charu C. Aggarwal and Philip S. Yu. Privacy-Preserving Data Mining: Models and Algorithms. Springer Publishing Company, Incorporated, July 2008. | 4. GaganAggarwal, Tom_asFeder, KrishnaramKenthapadi, Samir Khuller, Rina Panigrahy, Dilys Thomas, and An Zhu. Achieving anonymity via clustering. In PODS '06: Proceedings of the Twenty-_fth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, New York, NY, USA, 2006. ACM Press. | 5. GaganAggarwal, Tom_asFeder, KrishnaramKenthapadi, Rajeev Motwani, Rina Panigrahy, Dilys Thomas, and An Zhu. Approximation algorithms for k- anonymity. In Journal of Privacy Technology (JOPT), 2005. | 6. Rakesh Agrawal and Rama Krishnan Srikant. Privacy-preserving data mining.InProc.of the ACM SIGMOD Conference on Management of Data, pages 439{450. ACM Press, May 2000. | 7. Maurizio Atzori, Francesco Bonchi, FoscaGiannotti, and Dino Pedreschi. Blocking anonymity threats raised by frequent itemset mining. In ICDM, pages 561{564, 2005. | 8. Maurizio Atzori, Francesco Bonchi, FoscaGiannotti, and Dino Pedreschi.k-anonymous patterns. In PKDD '05: Proceedings of the 9th European Conference on Principles and Practice of Knowledge Discovery in Databases, pages 10{21,2005. | 9. Boaz Barak, KamalikaChaudhuri, Cynthia Dwork, Satyen Kale, Frank McSherry,andKunalTalwar. Privacy, accuracy, and consistency too: a holistic solution to contingency table release. In Proc. of PODS, pages 273{282, New York, NY, 2007. | 10. http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm | 11. http://zakki.dosen.narotama.ac.id/files/2012/02/Data-mining1.doc |