Volume : 2 | Issue : 5 | May 2013 • ISSN No 2277 - 8160

Research Paper | Computer Science

Privacy Preserving Data Mining

Prof. Dhaval Jadhav Research Scholar, CMJ University - Shillong- Meghalaya

Dr. Harshad Bhadka Head, MCA Department, C .U Shah college of MCA, Wadhwan. Gujarat

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,Pareto frontier, Data

1 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.

I address the privacy/utility tradeoff problem by considering the privacy and algorithmic requirements simultaneously. I study this problem in the context of two privacy models that attracted a lot of attention in recent years, *k*-anonymity and differential privacy. My analysis and experimental evaluations confirm that algorithmic decisions made with privacy considerations in mind may have a profound impact on the accuracy of the resulting data mining models.

2. Literature Review

2.1Data 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.

2.2 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.

One way to guarantee k-anonymity of a data mining model is to build it from a k-anonymized table. However, this poses two main problems: First, the performance cost of the anonymization process may be very high, especially for large and sparse databases. Infact, the cost of anonymization can exceed the cost of mining the data. Second, the process of anonymization may inadvertently delete features that are critical for the success of datamining and leave out those that are useless; thus, it would make more sense to perform datamining first and anonymization later.

2.3 Differential Privacy

Differential privacy is a recent privacy definition that guarantees the outcome of a calculation to be insensitive to any particular record in the data set.

Definition 8 *We say a randomized computation M provides ϵ-differential privacy if for any datasets A and B with symmetric difference $A\Delta B = 1$ (A and B are treated as multisets), and any set of possible outcomes $S \subseteq Range(M)$,*

$$\Pr[M(A) \in S] \leq \Pr[M(B) \in S] \times e^{\epsilon}.$$

The parameter ϵ allows us to control the level of privacy. Lower values of ϵ mean stronger privacy, as they limit further the influence of a record on the outcome of a calculation.

The parameter ϵ allows us to control the level of privacy. Lower values of ϵ mean stronger privacy, as they limit further the influence of a record on the outcome of a calculation. The values typically considered for ϵ are smaller than 1, e.g., 0.01 or 0.1 (for small values we have $e^{\epsilon} \approx 1 + \epsilon$). The definition of differential privacy maintains a composability property: when consecutive queries are executed and each maintains differential privacy, their ϵ parameters can be accumulated to provide a differential privacy bound over all the queries.

Typically, differential privacy is achieved by adding noise to the outcome of a query. One way to do so is by calibrating the magnitude of noise required to obtain ϵ differential privacy according to the sensitivity of a function. The sensitivity of a real-valued function expresses the maximal possible change in its value due to the addition or removal of a single record.

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 threeo bjective 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 1.1.

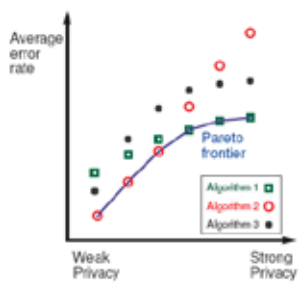

Figure 1.1: Example of a Pareto frontier. Given a number of learning samples, what are the privacy and accuracy tradeoffs?

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) isPareto 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 furtherextend the Pareto frontier, allowing for better privacy and accuracy tradeoffs.

5. 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 varyconsiderably 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 showen how privacy considerations can be interleaved withinthe execution of a data mining algorithm, allowing to switch rapidly between utilityoriented decisions and privacy-oriented decisions. For example, when inducing decisiontrees, a splitting criterion (utility) is used to pick an attribute to split a node. If thiswould result in a breach of k-anonymity, the attribute is generalized (privacy) and thealgorithm will re-evaluate (utility) the candidate attributes to make a new decision.This kind of interaction between utility and privacy considerations is not possible whenthe anonymization and data mining processes are distinct. For differentialprivacy wedemonstrated that it is not only important what the calculated functionality is, butalso 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 effectivenessof the resulting algorithm. In addition, when choosing the data mining algorithm, thedata miner should balance utility considerations with privacy considerations. Functionalities that are comparable in terms of utility may have a very different privacyimpact. For example, the Max, Information Gain and Gini Index criteria for choosingan attribute to split a decision tree node provide decision trees with comparable accuracy when no privacy considerations are involved. However, in a privacy preservingalgorithm, 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.

Algorithm-oriented approach

Different data mining algorithms take different approaches to data analysis and therefore present different utility considerations. Therefore, I claim that to improve the tradeoff between privacy and utility, a \one size fits all" solution would not be optimal. For example, using a classification metric to guide an anonymization algorithm could provide anonymized data that can be reasonably used for classification tasks. However, to reach better accuracy, a better approach would be to adapt a specific data mining algorithm such that it provides a k-anonymous outcome. Similarly, a framework such as PINQ has the huge advantage that it makes privacy-preserving data analysis approachable for programmers who have no expertise in privacy research. However, to achieve a good tradeoff between privacy and accuracy, it is better to develop a differentially-private version of the desired data mining algorithm.

Inducing a Larger Tree

We conducted numerous experiments, creating and learning trees of depths 3, 5 and 10,and setting the attributes and class to have 2, 3 or 5 distinct values. We used B = 1:0over sample sizes ranging from 1000 to 50,000 instances, setting 50,000 as the bound on the dataset size for determining information gain sensitivity. For each tested combination of values we generated 10 trees, executed 20 runs on each tree (each on a newly generated training set), and averaged the results over all the runs. In general, the results exhibit behaviour similar to that seen in the previous set of experiments. The variance in accuracy,albeit smaller than that observed for a single split, was apparent also when inducing deeper trees. For example, the typical standard deviation for the accuracy results presented in Figure 2.1 was around _5% and even lower than that for the results presented in Figure.

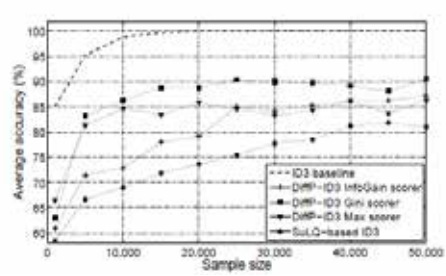

Figure 2.1 Inducing a tree of depth 5 with 10 binary attributes and a binary class, B = 1:0, pleaf = 0:3,pnoise = 0:1

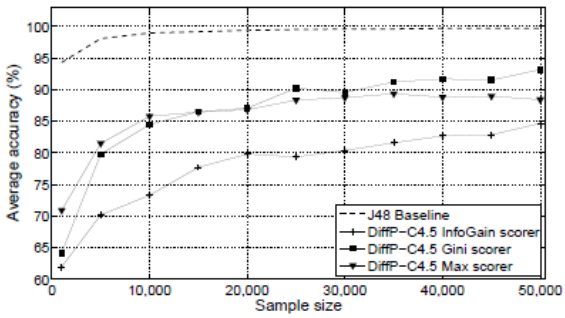

Figure 2.2: Inducing a tree of depth 5 with 7 binary attributes, 3 numeric attributes and a binary class, B = 1:0,pleaf = 0:3,pnoise = 0:1

When inducing shallower trees or using attributes with fewer distinct values, we observed an interesting pattern, which is illustrated in Figures 5.5 and 5.6. When the size of the dataset is small, algorithms that make e_cient use of the privacy budget are superior. This result is similar to the results observed in the previous experiments. However, as the number of available samples increases, the restrictions set by the privacy budget have less inuence on the accuracy of the resulting classi_er. When that happens, the depth constraint becomes more dominant; the Max scorer, which with no privacy restrictions usually produces deeper trees than those obtained with InfoGain or Gini scorers, provides inferior results with respect to the other methods when the depth constraint is present.

4. Real Dataset

We conducted experiments on the Adult dataset from the UCI Machine Learning Repository, which contains census data. The data set has 6 continuous attributes and 8 nominal attributes. The class attribute is income level, with two possible values, _50K or >50K.

After removing records with missing values and merging the training and test sets, the dataset contains 45,222 learning samples (we set 50,000 as the bound on the dataset size for determining information gain sensitivity). We induced decision trees of depth up to 5 with varying privacy budgets. Note that although the values considered for B should typically be smaller than 1, we used larger values to compensate for the relatively small dataset size3 (the privacy budget toll incurred by the attributes is heavier than it was in the synthetic dataset experiments). We executed 10 runs of 10-fold cross-validation to evaluate the different scorers. Statistical significance was determined using corrected paired t-test with confidence 0.05. Figure 5.7 summarizes the results. In most of the runs, the Max scorer did significantly better than the Gini scorer, and both did significantly better than the Info Gain scorer. In addition, all scorers showed significant improvement in accuracy as the allocated privacy budget was increased. The typical measured standard deviation in the allocated privacy budget was increased. The typical measured standard deviation in accuracy was _0:5.

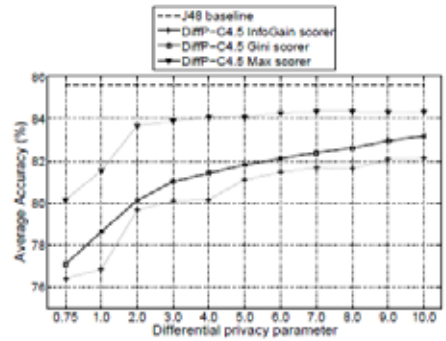

Figure 2.3: Accuracy vs. privacy (B) in the Adult dataset

Measuring privacy defining what privacy means in the context of data mining is a difficult task, and at large it is still an open research question. In this work we focused on two definitions of privacy that have drawn significant attention from the research community in the last few years: k-anonymity and differential privacy. k-anonymity and derived approaches rely on a syntactic definition to ensure that one released record is indistinguishable from a set of other records. The size k of the in distinguishability.set is used as the privacy measure in k-anonymity. We extended the definition of k- anonymity to the realm of data mining and used the extended definition to generate data mining models that conform to k-anonymity. While k-anonymity is relatively easy to understand and verify, recent research has pointed out that it may not offer suitable protection when an adversary has access to abundant auxiliary information or even other k-anonymized data sets. Differential privacy, on the other hand, provides a semantic definition of privacy with worst-case guarantees that hold regardless of the adversary's resources or available auxiliary information. A privacy parameter controls how much information can be leaked by a differentially private mechanism. Unfortunately, the strong guarantees provided by differential privacy come at a cost. Differential privacy poses harsh restrictions on data use {interactive mechanisms for differential privacy manage a privacy budget and block access to data once it is ex- hausted; Non-interactive mechanisms should be tailored to a predetermined set of queries, limiting future possible uses. These limitations may inhibit wide adoption of differential privacy unless breakthroughs in differential privacy research manage to overcome them.

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]RakeshAgrawal and RamakrishnanSrikant. 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. |