A Privacy Protection Technique for Publishing Data Mining Models and Research Data

YU FU, ZHIYUAN CHEN, GUNES KORU, and ARYYA GANGOPADHYAY
University of Maryland Baltimore County

Data mining techniques have been widely used in many research disciplines such as medicine, life sciences, and social sciences to extract useful knowledge (such as mining models) from research data. Research data often needs to be published along with the data mining model for verification or reanalysis. However, the privacy of the published data needs to be protected because otherwise the published data is subject to misuse such as linking attacks. Therefore, employing various privacy protection methods becomes necessary. However, these methods only consider privacy protection and do not guarantee that the same mining models can be built from sanitized data. Thus the published models cannot be verified using the sanitized data. This article proposes a technique that not only protects privacy, but also guarantees that the same model, in the form of decision trees or regression trees, can be built from the sanitized data. We have also experimentally shown that other mining techniques can be used to reanalyze the sanitized data. This technique can be used to promote sharing of research data.

Categories and Subject Descriptors: I.2.7 [**Artificial Intelligence**]: Natural Language Processing; H.2.8 [**Database Management**]: Database Applications—*Data mining*

General Terms: Algorithms, Security

Additional Key Words and Phrases: Preserving data mining

ACM Reference Format:

Fu, Y., Chen, Z., Koru, G., and Gangopadhyay, A. 2010. A Privacy protection technique for publishing data mining models and research data. ACM Trans. Manag. Inform. Syst. 1, 1, Article 7 (December 2010), 20 pages. DOI 10.1145/1877725.1877732. http://doi.acm.org/10.1145/1877725.1877732

1. INTRODUCTION

Data mining techniques have been widely used in many research disciplines such as medicine, life sciences, and social sciences to extract useful knowledge from research data in the form of data mining models [Grossman et al. 2001].

Y. Fu, Z. Chen and A. Gangopadhyay were supported in part by the National Science Foundation under Grant number IIS-0713345

Authors' addresses: Y. Fu, Z. Chen (corresponding author), G. Koru, and A. Gangopadhyay, Department of Information Systems, University of Maryland, Baltimore County, 1000 Hilltop Circle, Baltimore, MD 21250; email: zhchen@umbc.edu.

DOI 10.1145/1877725.1877732 http://doi.acm.org/10.1145/1877725.1877732

These models can then be published and used by others. For example, decision trees have been used to predict adverse drug reactions using clinical trial data [Hammann et al. 2010]. These trees can be used as guidelines for doctors to decide whether to prescribe a drug to a patient.

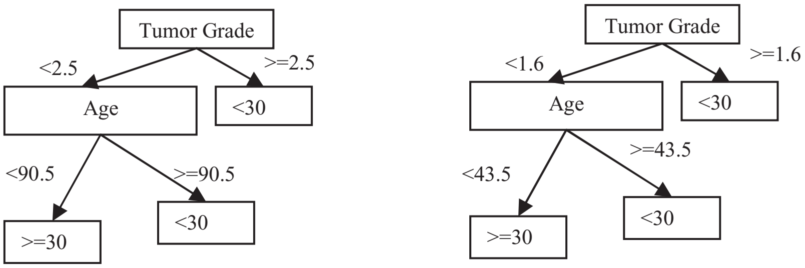
In addition to publishing mining results, researchers often need to publish research data that is used to create these models. There are important reasons to publish the underlying research data.

First, research data, if published, can be used by other researchers to verify the published research results. This can significantly add credibility to the results and alleviate some of the problems about scientific misconduct and research fraud. In a survey participated by 1389 researchers in the European Union [Kuipers and Hoeven 2009], around 90% of the participants considered that publishing research data was very important or important for validation of research results.

There has been an increasing trend of fraud in medical research recently [Black April 18, 2006]. Another recent example is “climate gate,” where a group of researchers working on climate change was accused of scientific misconduct. An independent panel cleared most accusations but pointed out that the researchers should have made data more accessible to the public and used better statistical tools in their analysis [House of Commons Science and Technology Committee Parliament of the United Kingdom 2010]. Clearly, publishing research data would reduce the chance of potential frauds or controversies.

Second, other researchers may conduct *secondary analysis* over the published research data in their own research. This has been widely used in disciplines such as social science and medical research. In these disciplines, data collection is often very expensive and secondary analysis saves resources that would otherwise be spent on collecting data. For example, in the medicine field, the cost of clinical trials conducted in the U.S. for new drugs was $25 billion at 2006 (Fee March 01, 2007). Social science studies also often use census data, which is impossible to be collected by individuals. Thus researchers often conduct secondary analysis on existing data if they are made available. For example, secondary analysis was used to discover the causes of some diseases from medical records that were not collected with the intention of detecting such a causal relation [Dale et al. 1988]. As of April 30, 2009, a search of the phrase “secondary analysis” in PubMed (a commonly used citation database of medical research) returns 3601 research articles. In the European Union survey mentioned before [Kuipers and Hoeven 2009], 91% of participants considered that publishing research data was very important or important for reanalysis of existing data.

Secondary analysis can be also divided into two categories: (1) reanalysis, which is the analysis of the data on the *same* research problem and (2) analysis that is used to solve a *different* research problem. For example, decision trees have been used to predict adverse drug reactions [Hammann et al. 2010]. Suppose that a researcher publishes the research data along with the decision tree. Other researchers may try the same or different mining methods for various educational or research purposes. This would be an example of first category because the research problem, that is, to predict adverse drug reactions of the



(a) original data, accuracy=0.73

(b) Sanitized by (LeFevre et al. 2006b), accuracy=0.59

Fig. 1. Decision trees built from original cancer data and sanitized data.

drug, remains the same. A third group of researchers may want to use the published data to discover behavioral patterns (e.g., smoking) among patients with a certain type of disease, regardless of whether they have adverse reactions to the drug. This is the case of the second category because the research problem is different from the problem in the original research.

Third, many people argue that if a research project is publicly funded, the results of the study including the research data should become public property; therefore, the data should be made accessible to the public. For example, the National Program of Cancer Registries (NPCR) is funded by the U.S. government. It collects data on the occurrence of cancer as well as initial treatment [National Center for Chronic Disease Prevention and Health Promotion 2010]. The data is available to the public and has been used for numerous research purposes such as for discovering patterns of cancer in different population groups [McDavid et al. 2004]. According to the European Union survey mentioned earlier [Kuipers and Hoeven 2009], 87% of participants consider that publishing research data is very important or important due to public funding.

Despite its numerous benefits, sharing research data is difficult because such data often contains privacy-sensitive information such as patients' medical conditions. Therefore, it becomes necessary to prevent or reduce privacy risks before releasing research data. Otherwise, an adversary can link the quasi-identifiers in the published datasets with other data sources available in order to reveal patients' identity. At the current state of the practice, privacy concerns often prevent many companies or researchers from sharing raw data. According to the EU survey [Kuipers and Hoeven 2009], only 25% of researchers publish their data. The survey also shows that the major barriers of sharing research data are legal issues (41% of responses) and potential misuse of data (41% of responses), both of which are closely related to privacy concerns.

There has been a rich body of work on privacy protection techniques [Vaidya et al. 2005]. The existing techniques only consider privacy protection and usually do not guarantee that the same mining model can be built from sanitized data. Thus, other researchers cannot verify the published models using the sanitized data.

For example, Figure 1(a) shows a decision tree built from a dataset about cancer patients to predict whether a patient will survive more than 30 months.

There are three attributes in the dataset: tumor grade, age, and survival status (whether the patient survives more than 30 months). We used a sanitization method proposed in LeFevre et al. [2006b]. Figure 1(b) shows the decision tree built from the sanitized data. Clearly, the tree in Figure 1(b) is different from the original one in Figure 1(a). Thus, other researchers cannot use the sanitized data to verify the published original decision tree.

The decision tree model is not preserved because the decision tree building algorithm needs to compute information gain (or other splitting criteria) for all possible splits in the data. Since data values have been distorted during the sanitization process, the information gain computed in the sanitized data is often different from that in the original data. Thus different splits are selected in the sanitized data. For example, in the original data, the split on tumor grade = 2.5 generates the highest information gain. However, after sanitization, the values of tumor grade attribute have been distorted and the best split becomes tumor grade = 1.6.

In addition, the accuracy (using 10-fold cross-validation) for the original tree is 0.73 while the accuracy of the tree built from sanitized data is only 0.59. Thus, the sanitized data also leads to inferior models being built. This also has negative impact on researchers who want to reanalyze the published research data.

In this article, we focus on two primary uses of research data: (1) verification of the published mining models; (2) reanalysis of the research data to solve the same research problem (i.e., the first category of secondary analysis). This work makes the following contributions:

- We propose a technique that both protects privacy and guarantees that the decision tree and regression tree, two popular data mining models, will be preserved, that is, the decision tree or regression tree built from the sanitized data will be exactly the same as that built from the original data.
- We conducted comprehensive experiments to compare our approach with existing privacy protection techniques. The results favor our approach.

The rest of the article is organized as follows. Section 2 reviews the related work. Section 3 describes our approach. Section 4 shows the experimental results and Section 5 concludes the article.

2. RELATED WORK

In this section, we first discuss the risks to privacy and the general categories for the privacy models. Next we discuss privacy protection techniques and briefly mention the work about hiding sensitive patterns. Finally, we place our study in the context of the related work.

Privacy risks and models. There are two types of privacy risks.

- Identity disclosure when the identity of a specific person in the dataset is revealed.
- Value disclosure when the values of some sensitive attribute values are revealed.

The two most popular privacy models are K-anonymity [Sweeney 2002] and L-diversity [Machanavajjhala et al. 2007]. K-anonymity prevents identity disclosure caused by linking attacks, which link attributes (called quasi-identifiers) such as birth date, gender, and ZIP code with publicly available datasets. This can be done by generalization, that is, replacing specific values with more general ones. For example, the exact age of a patient can be replaced with a range. Records with the same quasi-identifier values form an equivalence class. K-anonymity ensures that there are at least K people with the same quasi-identifier such that the risk of identity disclosure is reduced to $1/K$.

L-diversity prevents value disclosure by further requiring that the people with the same quasi-identifier contain at least L well-represented sensitive values such that attackers cannot discover the values of sensitive attributes easily. A more advanced model called t-closeness tries to make sure the distribution of sensitive attributes in each equivalence class is similar to the global distribution [Li et al. 2007].

Privacy protection techniques. There has been a rich body of work to enforce privacy protection models [Aggarwal et al. 2008]. These techniques can be divided into random perturbation [Agrawal and Srikant 2000], generalization, or suppression [LeFevre et al. 2006a], random permutation (e.g., randomly permute the values of sensitive attributes) [Xiao and Tao 2006], and synthetic data generation [Aggarwal and Yu 2004]. There also exists work on secure multiparty computation [Vaidya et al. 2005], which is useful for the distributed mining case. In this article we will only consider the case when the research data is published along with mining models. Next we will discuss several techniques related to data publication.

An additive perturbation technique was proposed in [Agrawal and Srikant 2000]. A reconstruction technique was also proposed to reconstruct the marginal distribution from perturbed data. A tree-based approach was proposed to sanitize data [Li and Sarkar 2006b]. The proposed approach used a KD-tree to divide data into groups and then generalize data in each group. A workload-aware anonymization approach was proposed in LeFevre et al. [2006b], where the anonymization process is optimized for specific mining tasks. For example, the anonymization tries to maximize information gain (which is used in decision tree building) for classification. Another perturbation approach was proposed in Li and Sarkar [2006a] for categorical data. This approach randomly swaps sensitive attribute values in records that have high disclosure risks and at the same time tries to preserve both the marginal distribution of the sensitive attribute and the correlation between nonsensitive attributes and the sensitive attribute.

However, such techniques do not provide any guarantee on model preservation. Next we discuss two of them. Let us first consider the information-gain-based method in LeFevre et al. [2006b]. As shown in Figure 1, this method does not preserve decision trees because the information gain computed over the sanitized data is often different from the information gain computed over the original data.

Now let us consider the method proposed in Li and Sarkar [2006a]. This method tries to preserve the marginal distribution of sensitive attributes and the correlation between sensitive and nonsensitive attributes. This method has two problems. First, there is no guarantee that the aforesaid statistical information will be 100% preserved (the solution only tries to preserve it as much as possible). Second, preserving such statistical information may not be sufficient to preserve the decision tree patterns. For example, in Figure 1, the class label (survival status) cannot be sensitive because otherwise attackers can simply use the published decision tree to predict its values. Suppose the sensitive attribute is "age" and other attributes are nonsensitive. The method proposed in Li and Sarkar [2006a] tries to preserve the correlation between age and tumor grade as well as the correlation between age and survival status. However, the decision tree also depends on the correlation between tumor grade and survival status, which may not be preserved by the preceding method.

The only work we are aware of that preserves data mining models is to publish the contingency table for naïve Bayesian classifiers [Mozafari et al. 2009]. However, it is unclear how that method can be applied to other mining models such as decision trees because decision trees use more information than the contingency table.

There also exists work on hiding sensitive patterns such as frequent item sets in data. Several approaches were proposed in Menon and Sarkar [2007] and Menon et al. [2005] to hide sensitive frequent item sets and at the same time minimize information loss.

Comparison of our approach with the related work. All the existing work on privacy protection does not guarantee that the decision tree or regression tree models are preserved. The approach proposed in this article preserves these two tree models and at the same time protects data privacy. This article is also an extended version of our preliminary work [Fu et al. 2009a; 2009b]. The extensions include: (1) more comprehensive experiments, (2) extension of our approach to satisfy given privacy requirements, (3) efficiency improvement of our approach.

3. OUR APPROACH

In Section 3.1, we first specifically formulate the problem tackled in this article. Then, we briefly describe the decision tree and regression tree building algorithms. In Section 3.2 we prove a theorem that describes the conditions under which a tree model can be preserved. Finally, in Section 3.3, we present a method that preserves both privacy and the decision or regression tree model.

3.1 Background

Problem definition. Let T be a data table with attributes A_1 to A_m. These attributes can be divided into sensitive attributes (whose values need to be protected) and nonsensitive attributes. We also assume that all nonsensitive attributes are quasi-identifier attributes. We assume that attribute A_m is the response variable (which needs to be predicted). Let K and L be two integers, and B be a decision tree or regression tree building algorithm.

The goal is to create a sanitized table T' such that T' satisfies K-anonymity and L-diversity, and at the same time, B can build the same decision tree or regression tree P from T' or T to predict the value of A_m.

Decision tree and regression tree building algorithms. The structure of a decision tree or a regression tree is as follows. Each internal node of a decision tree or regression tree contains a test condition and several branches representing test outcomes. For example, in the root node of the tree seen in Figure 1(a), the patients with a tumor grade less than 2.5 are assigned to the left child, and those with a tumor grade greater than or equal to 2.5 are assigned to the right child. A leaf of a decision tree predicts a class label; a leaf of a regression tree predicts a numerical outcome.

Most existing tree building algorithms create the tree in a top-down fashion [Han and Kamber 2000]. They start with a single node that represents the whole training dataset. Next, they recursively expand the current tree by partitioning the records in the tree nodes. At each step, an attribute A_i other than the response variable A_m is chosen to optimize a splitting criterion. If A_i is numerical, a value v (typically as the average of two consecutive A_i values) is selected such that the records with A_i values less than v go to left child, and those with values greater than or equal to v go to right child. If A_i is categorical, either a binary split is used where A_i's values will be divided into two disjoint sets, or a multiway split is used, where each value of A_i will become a child node. These algorithms stop when a certain stopping criterion is met during the successive splitting actions.

There are three commonly used splitting criteria for decision trees: information gain, gain ratio, and Gini index. Here we just describe information gain while our approach also applies to the other two. Let S be the set of records at an internal node in P, t be the number of child nodes, $S_j (1 \le j \le t)$ be the set of records in child j, A_i be the split attribute, v be the split value, and $C_1, \ldots, C_q$ be the q classes. Let $f(C_i,S)$ be the frequency of class C_i in S. The information gain equals

$$\begin{aligned} \mathit{InfoGain}(S, v) \;=\; & \sum_{i=1}^{q} \frac{f(C_i, S)}{|S|} \log_2 \frac{|S|}{f(C_i, S)} \\ & - \sum_{j=1}^{t} \frac{|S_j|}{|S|} \sum_{i=1}^{q} \left(\frac{f(C_i, S_j)}{|S_j|} \log_2 \frac{|S_j|}{f(C_i, S_j)} \right). \end{aligned} \tag{1}$$

For regression trees, the commonly used splitting criterion is the reduction of deviance of response variable. Let A_m be the response variable. The reduction of deviance equals

$$\mathit{DevGain}(S, v) = \mathit{Var}(S, A_m) - \sum_{j=1}^{t} \frac{|S_j|}{|S|} \mathit{Var}(S_j, A_m), \tag{2}$$

where $\mathit{Var}(S, A_m)$ is the variance of response variable A_m in set S.

3.2 Conditions for Preserving Tree Models

In this section, we will describe conditions under which tree models will be preserved. Note that the first sum in Eq. (1) is constant for all splits, thus the splitting criterion only depends on $|S_j|$ and $f(C_i, S_j)$, that is, the size of each child node and the distribution of class labels in each child node. This property also holds for gain ratio and Gini index. Similarly, in Eq. (2), the first term is constant for all possible splits, so the splitting criterion only depends on $|S_j|$ and $Var(S_j, A_m)$, that is, the size of each child node and the variance of response variable in each child node.

THEOREM 1. *If the privacy protection algorithm satisfies the following three conditions, the decision tree or regression tree generated from the sanitized data will be the same as that generated from the original data.*

(1) *It leaves the values of response variable A_m unchanged.*[1]

(2) *Let A_i be a categorical attribute that appears in the tree.*
   (a) *If a multiway split is used, A_i cannot be generalized.*
   (b) *If a two-way split is used, let VS_1 and VS_2 be the sets of A_i values in the child nodes S_1 and S_2. Let $VS'(v)$ be the set of values (including v) that will be generalized to the same value v'. Then all values in $VS'(v)$ must belong to the same branch as v. That is, if $v \in VS_1$ (or VS_2), then $VS'(v) \subseteq VS_1$ (or VS_2).*

(3) *Let A_i be a numerical attribute that appears in the tree. Both of the following two conditions need to be satisfied.*
   (a) *The order of values of A_i is preserved, that is, if $v_1 \leq v_2$ in original data, $v'_1 \leq v'_2$ in the sanitized data where $v_j (j = 1, 2)$ is a value of A_i and v'_j is the sanitized value of v_j.*
   (b) *Let v be a split value and $v_1(v_2)$ be the maximal (minimal) value of A_i in the left (right) child after the split, respectively. The sanitized values v'_1 and v'_2 must satisfy that $v_1 + v_2 = v'_1 + v'_2$.*

PROOF. For decision trees, these conditions ensure that both the child record sets (S_j) and the distribution of class labels in each S_j for all possible splits remain unchanged in the sanitized data. According to Eq. (1), keeping the child record sets and the distribution of class labels the same ensures that the decision tree model will be preserved. Similarly, for regression trees, these conditions will ensure that S_j and variance of response variable in each S_j remain unchanged. Thus, according to Eq. (2), the regression tree model will be also preserved. Next, we illustrate the case for decision trees. The same reasoning applies for regression trees. □

Condition (2) ensures that, when the sanitization process generalizes a categorical attribute, the generalization will preserve the set of records in each branch. When a multiway split is used, each value of A_i forms a branch so A_i cannot be generalized because otherwise some branches in the decision tree will be merged after sanitization (Condition (2a)). When two-way split

[1] Normalization of A_m is allowed.

is used, we use an example to illustrate Condition (2b). Suppose that an attribute called *education level* contains five values: high school, 2-year college, Bachelor, Master, and PhD. Suppose a split puts the records with high school and 2-year college into the left child and the remaining records into the right child. In the sanitization process, we can generalize high school and 2-year college into at-most-2-year-College, and generalize Bachelor, Master, and PhD into at-least-4-year-college. Clearly, if a record belongs to left (or right) child in the original data, its generalized value still belongs to the left (or right) child.

Condition (3a) ensures that the order for a numerical split attribute A_i is preserved. The decision tree algorithm will check all possible splits on A_i, thus preserving the order of A_i will ensure that the same set of child record sets (S_j) will be generated. Since condition (1) also preserves the class labels, the distribution of class labels in S_j remains unchanged. Thus, the decision tree algorithm will select the same best split as in the original data.

Example 1. For example, suppose there are six records $r_1, \ldots,$ and r_6. r_1, r_3, and r$_5$ in class C_1 and r_2, r_4, and r_6 in class C_2. Let A_i be a numerical attribute. Suppose that in the original data, the sorted order of A_i is r_1, r_2, r_3, r_4, r_5, and r_6. The class labels in the order of A_i are thus C_1, C_2, C_1, C_2, C_1, and C_2. The best split in the original data generates two S_j: $\{r_1(C_1), r_2(C_2), r_3(C_1)\}$ and $\{r_4(C_2), r_5(C_1), r_6(C_2)\}$. If the order on A_i is preserved in the sanitized data, the class label in the order of A_i is still C_1, C_2, C_1, C_2, C_1, and C_2. Thus, the best split will remain unchanged (between r_3 and r_4).

However, suppose the order of A_i in sanitized data is changed to r_1, r_3, r_2, r_4, r_5, *and* r_6, the class label in the order of A_i becomes C_1, C_1, C_2, C_2, C_1, and C_2. The best split in the sanitized data becomes $\{r_1(C_1), r_3(C_1)\}$ and $\{r_2(C_2), r_4(C_2), r_5(C_1), r_6(C_2)\}$, which is different from the best split in the original data.

Condition (3b) further ensures that for a numerical split attribute A_i, the split value in the transformed data equals the split value in the original data, which equals the average of the maximal value of A_i in left right and the minimal value of A_i in the right child.

3.3 Proposed Data Sanitization Procedure

Figure 2 describes the Tree-Pattern-Preserving Algorithm (TPP). The input of the algorithm includes original data T, a decision tree or regression tree building algorithm B, and privacy parameters K and L. The output is a tree model P and a sanitized dataset T' that satisfies both K-anonymity and L-diversity. The same tree P can be built from T' as well.

Step 1 of the algorithm builds a decision tree with one node. Steps 2 to 4 try to sanitize the data. We will show shortly that these steps satisfy all conditions in Theorem 1 and thus preserve the current decision tree or regression tree. Step 5 will check whether privacy requirements are satisfied. If so, we will repeatedly expand the decision tree or regression tree and rerun steps 2 to 5 to sanitize the data. Otherwise, we return the latest tree that satisfies the privacy requirements along with the sanitized data.

```
1. Run tree building algorithm $B$ to generate a tree $P$ with only one node.
2. For each attribute $A_i$ that is not the response variable and is not used in $P$, replace its value with a single value (for categorical attribute, use ALL; for numerical attribute, use mean of $A_i$).
3. For each numerical attribute $A_i$ that appears in the tree, do the following:
   a) For each node x in $P$ that uses $A_i$ as split attribute, collect boundary values as the maximal $A_i$ value in the left child and the minimal $A_i$ value in the right child.
   b) Sort values of $A_i$ and divide them into intervals using boundary values collected in step 3a)
   c) If an interval contains two boundary values, split it into two equal size intervals such that each contains only one boundary value. Compute the mean of each new interval, let them be $u_1$, $u_2$, ...
   d) For each node $x$ in $P$ that uses $A_i$ as split attribute, let $v_1$ ($v_2$) be the maximal (minimal) $A_i$ value in the left (right) child of $x$. Let $I_1$ ($I_2$) be the intervals with $v_1$ ($v_2$) as the right (left) boundary. Compute $d = min\{v_1 - u_1, u_2 - v_2\}$. Replace values in $I_1$ with $v_1$-$d$, and values in $I_2$ with $v_2$+$d$.
4. For each categorical attribute $A_i$ that appears in $P$, if two-way split is used, divide values of $A_i$ into groups such that the values in the same group appear in the same branches in $P$ (this can be done by sorting values on the branches they appear). Replace values of $A_i$ in the same group with the same generalized value.
5. Group all records on quasi-identifier attributes. For each group, check whether it satisfies K-anonymity and L-diversity.
6. If privacy requirements are satisfied, call tree building algorithm $B$ to expand $P$ once more and rerun step 2 to 5 until the stopping condition of $B$ is met.
7. Otherwise, return the last tree that satisfies the privacy requirements and the data sanitized based on that tree.
```

Fig. 2. Tree-Pattern-Preserving algorithm (TPP).

Next, we show how steps 2 to 4 satisfy conditions in Theorem 1. First, these steps do not change the values of response variables. Thus, Condition (1) is satisfied. Step 4 sanitizes categorical attributes and it is easy to verify it satisfies Condition (2). Step 3 sanitizes numerical attributes. We will use an example to show how it satisfies Condition (3).

Figure 3 shows how step 3 works for Example 1. Suppose the tree building algorithm selects a numerical attribute A_i as the split attribute. The best split in original data is between r_3 and r_4. Thus step (3a) will pick the A_i values of r_3 and r_4 as boundaries (let them be v_1 and v_2, respectively). In step (3b), two intervals get created: I_1 containing r_1 to r_3 and I_2 containing r_4 to r_6. Each interval only contains one boundary value. Step (3c) computes the mean of each interval. Step (3d) computes the gap between v_1 (v_2) and the mean of I_1 (I_2). Let δ be the smaller of these two gaps. It then generalizes the values in I_1 to $v_1 - d$, and values in I_2 to $v_2 + d$. Clearly, the new split value in the sanitized data $(v_1 - d + v_2 + d)/2$ is the same as the old split value $(v_1 + v_2)/2$. Thus Condition (3b) is satisfied. The order is also preserved because $v_1 \leq v_2$ and $v_1 - d \leq v_2 + d$. Thus Condition (3a) is satisfied.

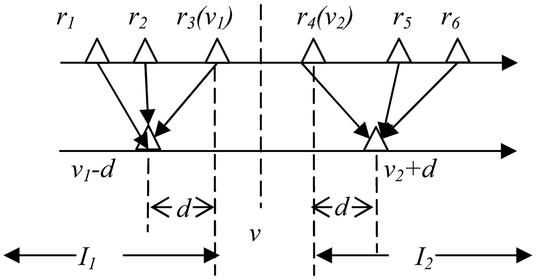


Fig. 3. Sanitizing numerical attribute.

Privacy protection. The algorithm will ensure that the sanitized data satisfies K-anonymity and L-diversity. One possible attack is when the attacker knows the order of a numerical attribute (e.g., knowing P_1 has the smallest value). TPP preserves the order for this attribute. However, all values in the same interval are generalized to the same value (e.g., P_1 to P_3 all have the same A_i values after sanitization). Thus, the attacker can only locate the group of rows with smallest A_i values, but cannot decide which one in the group is P_1. Since there are at least K people in each group, the probability of identifying P_1 is at most *1*/*K*.

Complexity analysis. Let n be the number of rows and m be the number of attributes. Let | P | be the number of nodes in the final tree. Step (3a) costs O(| P |) because it needs to check every node in | P | . Step (3b) needs to sort values of each numerical attribute and costs *O(m n log n)*. Step 2, (3c), (3d), and step 4 transform all data values and cost O(mn). Step 5 can be implemented by sorting data on quasi-identifier attributes and costs O(*mn log n*). Step 6 needs to rerun step 2 to 5 for an expanded tree. Thus the overall cost of TPP is O(*m n log n* | P |) because the tree can be expanded at most O(| P |) times.

Efficiency improvement. Next we describe several ways to improve the efficiency of the TPP algorithm.

First, we can delay the replacement of data values in steps 2, (3d), and 4. Instead, we can keep for each numerical attribute the set of intervals generated in step (3c), and keep for each categorical attribute the set of generalized values (see step 4). We can then replace data values only for the final tree P.

Second, we can further reduce the cost of sorting in step 5. This sorting step put records with the same values on quasi-identifier attributes into the same group. We can presort all records on numerical attributes. Note that our algorithm preserves the order of numerical attributes (Condition (3a) in Theorem 1). Thus, records in the same group must be consecutive records after presorting. Therefore, step 5 just needs to check whether consecutive records belong to the same group and there is no need to resort data. As a result, the cost of step 5 is reduced from O(*mn log n* | P |) to O(*mn log n*).

Finally, the cost of rerunning steps 3 and 4 for an expanded tree P can be reduced. The key observation is that the expanded tree has the same structure as the original tree except the expanded nodes. Thus we only need to consider the expanded part of tree in steps 3 and 4. Here we just illustrate the cost for step 3. Step (3a) needs to collect boundary nodes in the expanded part of the tree. Since the values of A_i are already sorted in previous round, there is no need to resort them in step (3b). Instead, we just need to check whether new intervals are generated by these additional boundary nodes. For example, suppose in Example 1 the expanded tree has a new split between r_2 and r_3. We just need to collect two new boundary nodes: r_2 and r_3. The old interval I_1 (which contains r_1 to r_3) is now split into two new intervals: $\{r_{1,}\ r_2\}$ and $\{r_3\}$. Thus the total cost of step 3 and 4 are also reduced from O($mn\ log\ n \mid P \mid$) to *O(mn log n)* because there is no need to resort data for an expanded tree. Hence the overall cost of the algorithm is reduced to O($mn\ log\ n$).

3.4 Extension to Preserve Multiple Tree Models

The TPP algorithm preserves one decision or regression tree model. However, in practice researchers may use different tree building algorithms or change the parameters of tree building algorithms to generate multiple tree models. TPP can be extended to preserve multiple tree models. Suppose there are $B_1, B_2, \ldots, B_u$ tree building algorithms (it could be the same algorithm with different parameter settings), and they generate tree models $P_1, P_2, \ldots, P_u$. Figure 4 shows the extended algorithm that preserves all these tree models. The extended algorithm is very similar to the original TPP algorithm except that it tries to preserve the conditions in Theorem 1 (which guarantees preservation of tree models) for all the tree models. For example, at step (3a) of the extended algorithm, we will collect boundary values in all the tree models rather than in a single tree model for a numerical attribute A_i. At steps (3b) and (3c) intervals of A_i will be formed using all these boundary values. Later values of A_i are transformed in step (3d), which is in the same way as in TPP. This will ensure that the split on A_i in all these tree models will be preserved.

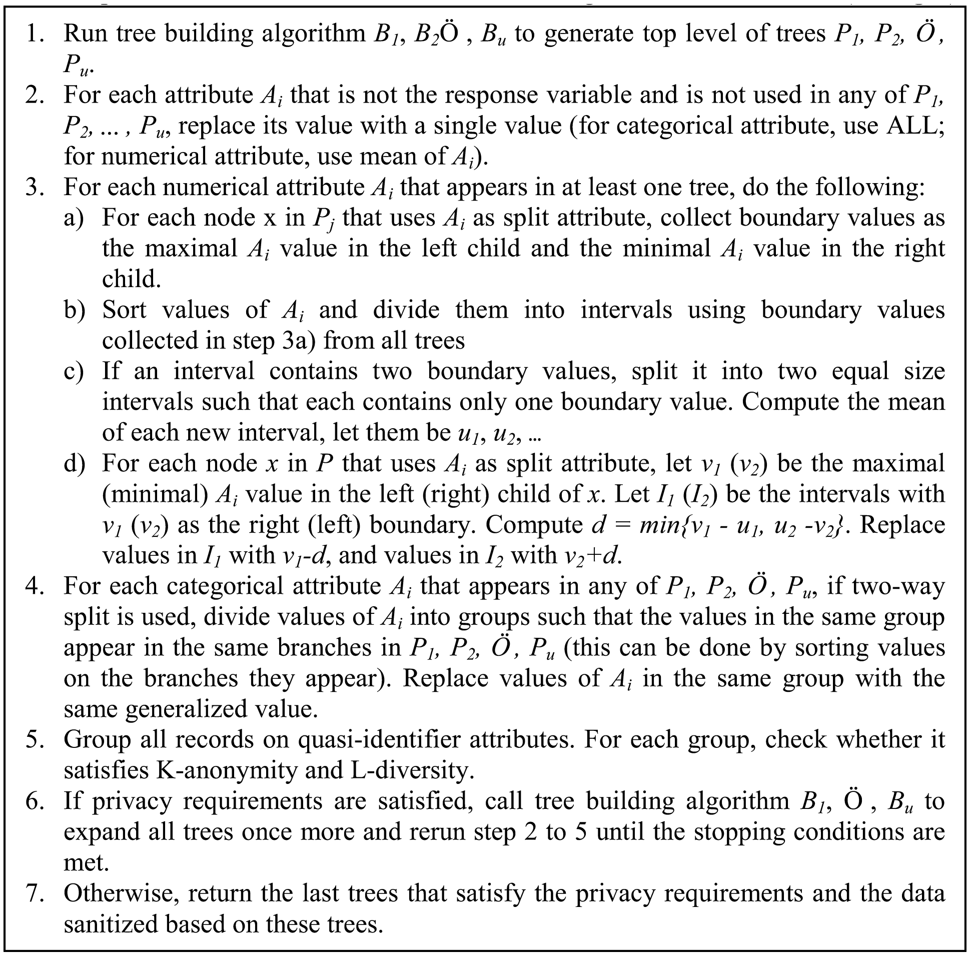
For example, consider the 6 points in Figure 3. Suppose tree P_1 splits between r_3 and r_4 and tree P_2 splits between r_2 and r_3. Thus the extended TPP algorithm will generate 3 intervals: r_1 to r_2, r_3, and r_4 to r_6. Clearly, both the split between r_2 and r_3 and the split between r_3 and r_4 will be preserved.

The complexity of the extended algorithm is at most u (the number of tree models) times the complexity of the TPP algorithm.

4. EXPERIMENTAL EVALUATION

Data. We used two real-life datasets: the Adult dataset from UCI Repository of Machine Learning datasets [Hettich et al. 1998] and the Cancer dataset obtained from University of Kentucky Cancer Research Center. The Adult data contains census data and is also the de facto benchmark in the literature. It contains 30717 records, 5 numerical attributes, and 7 categorical attributes. We used "occupation" as the sensitive attribute and the rest as quasi-identifiers. The Cancer dataset contains 3537 records. It has 3 numerical attributes and 3 categorical attributes. We used "histology" as the sensitive attribute. Our method was implemented in R. The experiment was run on a desktop PC with 3.2G HZ CPU and 2GB RAM, running Windows XP.

Methods. For the Adult dataset, we built a decision tree to predict whether the annual household income is over 50K. For the Cancer dataset, we built



```
1. Run tree building algorithm $B_1$, $B_2$Ö , $B_u$ to generate top level of trees $P_1, P_2,$ Ö, $P_u$.
2. For each attribute $A_i$ that is not the response variable and is not used in any of $P_1$, $P_2$, ... , $P_u$, replace its value with a single value (for categorical attribute, use ALL; for numerical attribute, use mean of $A_i$).
3. For each numerical attribute $A_i$ that appears in at least one tree, do the following:
   a) For each node x in $P_j$ that uses $A_i$ as split attribute, collect boundary values as the maximal $A_i$ value in the left child and the minimal $A_i$ value in the right child.
   b) Sort values of $A_i$ and divide them into intervals using boundary values collected in step 3a) from all trees
   c) If an interval contains two boundary values, split it into two equal size intervals such that each contains only one boundary value. Compute the mean of each new interval, let them be $u_1$, $u_2$, ...
   d) For each node $x$ in $P$ that uses $A_i$ as split attribute, let $v_1$ ($v_2$) be the maximal (minimal) $A_i$ value in the left (right) child of $x$. Let $I_1$ ($I_2$) be the intervals with $v_1$ ($v_2$) as the right (left) boundary. Compute $d = min\{v_1 - u_1, u_2 - v_2\}$. Replace values in $I_1$ with $v_1$-$d$, and values in $I_2$ with $v_2$+$d$.
4. For each categorical attribute $A_i$ that appears in any of $P_1, P_2,$ Ö, $P_u$, if two-way split is used, divide values of $A_i$ into groups such that the values in the same group appear in the same branches in $P_1, P_2,$ Ö, $P_u$ (this can be done by sorting values on the branches they appear). Replace values of $A_i$ in the same group with the same generalized value.
5. Group all records on quasi-identifier attributes. For each group, check whether it satisfies K-anonymity and L-diversity.
6. If privacy requirements are satisfied, call tree building algorithm $B_1$, Ö , $B_u$ to expand all trees once more and rerun step 2 to 5 until the stopping conditions are met.
7. Otherwise, return the last trees that satisfy the privacy requirements and the data sanitized based on these trees.
```

Fig. 4. Tree-Pattern-Preserving algorithm extended to preserve multiple trees.

a regression tree to predict the number of years a patient will survive after diagnosis of cancer. We compare our method (TPP) to the InfoGain method in LeFevre et al. [2006b] because it has the best prediction accuracy among existing methods. InfoGain partitions data into groups such that information gain is maximized. It then generalizes quasi-identifier attributes in each group. It does not satisfy Condition 3 in Theorem 1 (preserving order and split values for numerical attributes), thus it does not preserve decision trees or regression trees.

Metrics. We reported the accuracy of mining models built from the sanitized data using 10-fold cross-validation. We used K-anonymity and L-diversity to measure the degree of privacy protection. Larger K and L mean more protection. In terms of L-diversity, the sensitive attributes in both datasets are not used in the decision tree or regression tree model and are thus suppressed by both TPP and InfoGain. This is the best a privacy protection method can do. The best

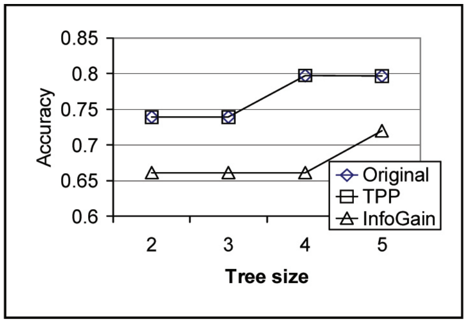


Fig. 5. Accuracy of decision trees on Adult data.

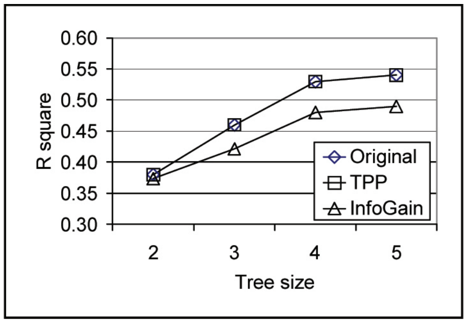


Fig. 6. R square of regression trees on Cancer data.

strategy for attackers is to assume that the sensitive attribute always has the most frequent value, assuming that attackers know the most frequent value of the sensitive attribute. We use the strong form of L-diversity where the fraction of the most frequent values in each equivalence class must be less than $1/L$ [Xiao and Tao 2006]. Thus the maximal probability of privacy breach is $1/L$.

Accuracy of tree models. Since both the prediction accuracy and the degree of privacy protection vary with the size of trees, we varied tree size (as the number of leaf nodes) in our experiments. Figure 5 reports the accuracy of decision trees built from sanitized data. The accuracy for trees built from the original data is also reported as the baseline. The results show that the trees built from data sanitized by TPP have higher accuracy than the trees of the same size but built from data sanitized by InfoGain. More importantly, TPP always preserves the decision tree model while InfoGain never preserves the model in all experiments. The accuracy using data sanitized by TPP is the same as that using original data because TPP preserves decision trees.

Figure 6 reports the R square of regression trees built from sanitized data. Again, the trees built from TPP have the same mining quality (in terms of R square) as the trees built from the original data. InfoGain does not preserve regression trees and also leads to lower R square.

Privacy results. Figures 7 and 8 report K-anonymity results for the two datasets, respectively. K decreases as the tree becomes larger because as

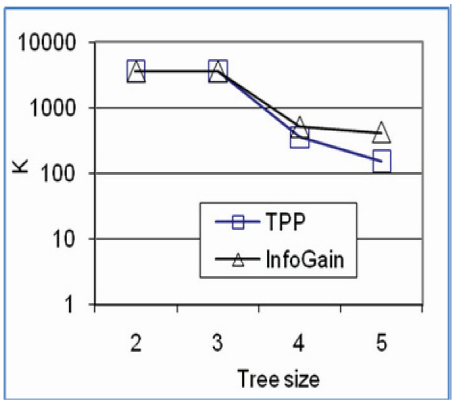


Fig. 7. K-anonymity on Adult data.

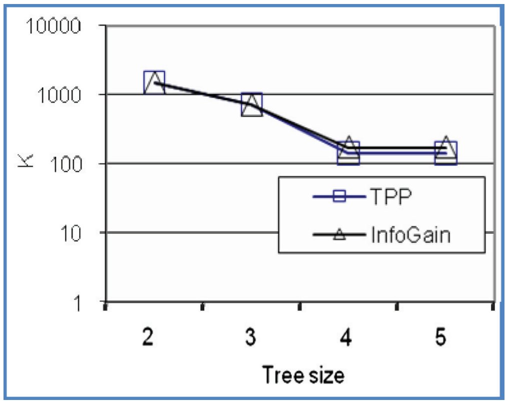


Fig. 8. K-anonymity on Cancer data.

the tree grows, more intervals will be generated by TPP and the degree of generalization becomes less. The K values for TPP are slightly worse than those of InfoGain for trees with 4 or 5 leaves, because TPP preserves the tree model and thus does less generalization. This is the price we pay for preserving mining models.

Figures 9 and 10 report L-values for the two datasets, respectively. In all experiments, the values of L for both methods are quite close and are in the range of 2 to 3 for the Adult data and are in the range of 3 to 4 for the Cancer data. This means that the probability of privacy breach is at most 1/2 for Adult and 1/3 for Cancer, assuming that the attackers know the most frequent values of sensitive attributes. These probabilities are relatively high, largely due to the skewed distribution of sensitive attributes. For example, the probability of most frequent value of the sensitive attributes is 0.135 in the Adult dataset and is 0.24 in the Cancer dataset. Thus even if we completely distort the dataset (by making all nonsensitive attribute values identical), the probability of privacy breach is still 0.135 in Adult and 0.24 in Cancer.

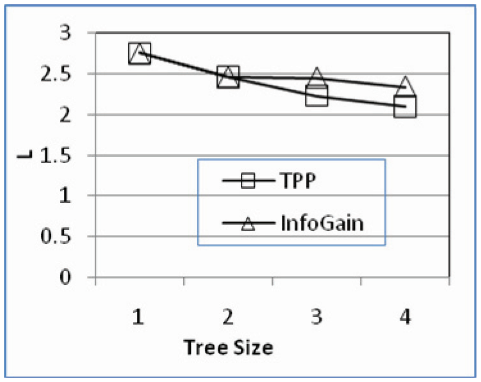


Fig. 9. L-diversity on Adult data.

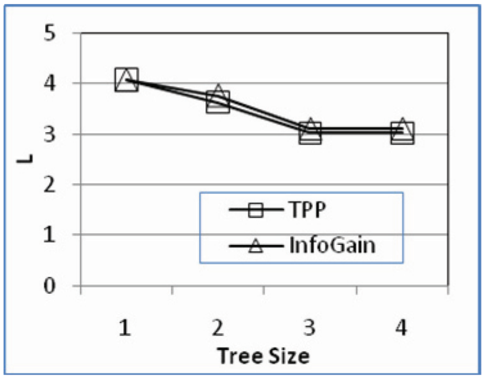


Fig. 10. L-diversity on Cancer data.

Tree size can be decided by considering the trade-off between prediction accuracy and the degree of privacy protection. For example, for the Adult dataset, the tree with 4 leaves seems to give the best trade-off. A smaller tree size is usually preferred because it is less likely to overfit the data. It also leads to better privacy protection. Thus, a rule of thumb is to use the smallest tree that provides sufficient accuracy.

Accuracy of other mining methods. We also conducted experiments simulating the cases when other researchers reanalyze the published data using different mining methods. We assumed that other researchers would use two other classification methods: naïve bayesian and support vector machine for the Adult dataset. They would also use another prediction method: linear regression for the Cancer dataset. Figure 11 reports the accuracy of naïve Bayesian using the Adult dataset sanitized by TPP or InfoGain as training data. Figure 12 reports the results for SVM. Figure 13 reports the R square of linear regression on the Cancer dataset sanitized by TPP or InfoGain. All results show that TPP leads to better mining quality than the InfoGain method. Further, the mining quality using data sanitized by TPP is also quite close to that of the original data when tree size reaches 4.

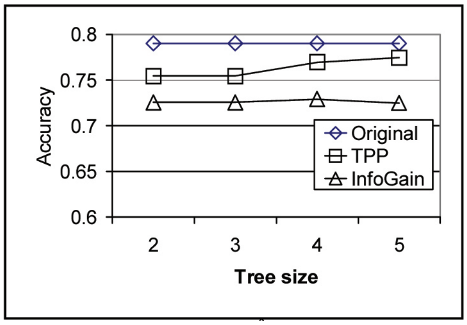


Fig. 11. Accuracy of naïve Bayesian on Adult data.

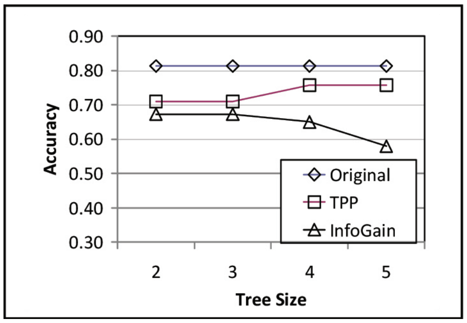


Fig. 12. Accuracy of SVM on Adult data.

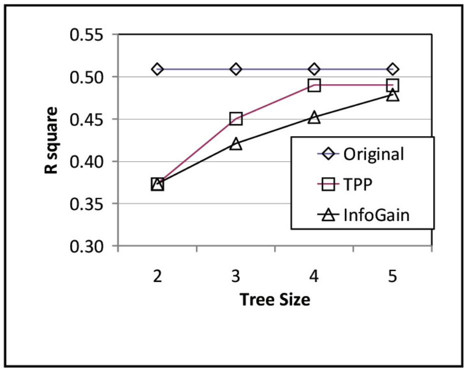


Fig. 13. R square of linear regression on Cancer data.

A possible explanation is that when different mining methods are used for the same mining task (e.g., to predict the class label), they all rely on similar patterns in the original data. For example, decision trees, naïve Bayesian, and SVM all rely on the correlation between other attributes and the class label attribute to predict class labels. By preserving decision tree models, our approach already largely preserves such correlations. Thus our approach also

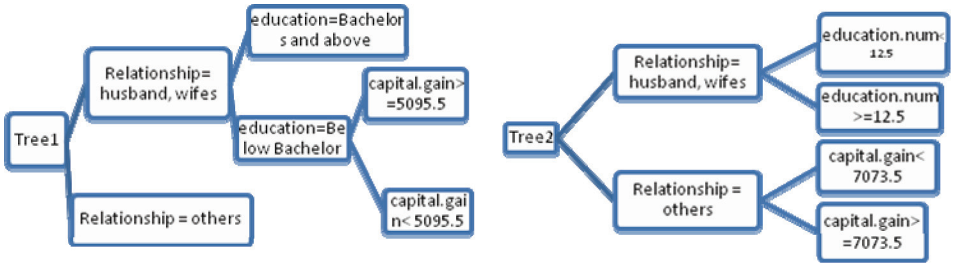


Fig. 14. Two tree models built from Adult dataset.

Table I. Accuracy and K Values when Preserving Two Tree Models

	K-value	Accuracy
Only Tree 1 preserved	154	0.8051568
Only Tree 2 preserved	120	0.7471107
Both Tree 1 and Tree2 preserved	64	0.8051568 (for tree1) 0. 7471107 (for tree2)

leads to good mining results for naïve Bayesian and SVM. Therefore, when researchers try a different mining method on the sanitized data, our proposed method (TPP) is still very likely to give better mining results than existing methods.

Preservation of two tree models. Figure 14 shows two trees we want to preserve in the sanitized Adult dataset. These trees are built using two different tree building algorithms in R: “class” and “anova”. Though both trees try to predict the same attribute “income”, they have different tree structure since the second level of splits.

We used the extended TPP algorithm (described in Section 3.4) to produce the sanitized data based on two trees. Table I shows the accuracy as well as K values for three cases: (1) when the first tree was preserved, (2) when the second tree was preserved, (3) when both trees were preserved. The results show that our method does preserve both trees as well as the accuracy of both tree models. However, the degree of privacy protection (i.e., value of K) decreases due to extra sanitization that is needed to preserve both trees.

Execution time. The execution time of our method was less than 2.5 seconds in all experiments. We also ran an experiment to examine the scalability of our algorithm. We generated 10 datasets containing 6.25%, 12.5%, 25%, 50%, 100% of the records in the Adult dataset. Figure 15 reports the execution time of TPP over these datasets. We also generated 10 datasets containing 3, 4, 5,.., to all 12 attributes. Figure 16 reports the execution time of TPP over these datasets. The results show that TPP scales almost linearly with both number of records and number of attributes.

5. CONCLUSION

This article proposes a privacy protection technique that preserves decision tree and regression tree models and at the same time protects privacy. We first identify conditions that a privacy protection method must satisfy to preserve

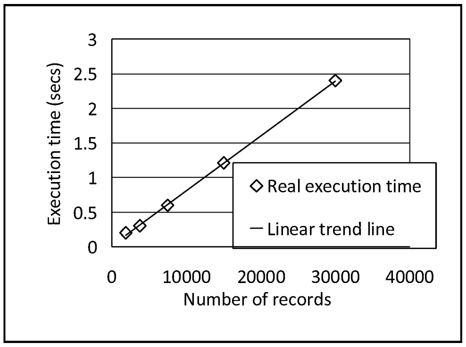


Fig. 15. Execution time when varying number of records.

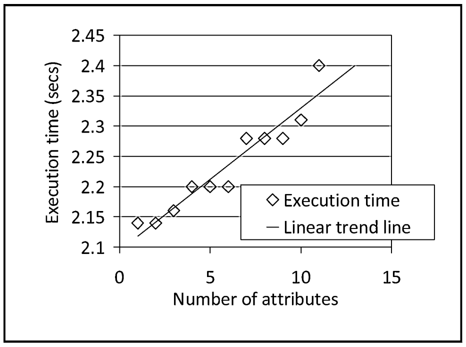


Fig. 16. Execution time when varying number of attributes.

the mining models and then design an efficient algorithm that satisfies these conditions.

Experimental results show that our approach not only preserves decision tree and regression tree models, but also leads to better mining quality for several popular mining methods over the sanitized data.

Researchers can use our approach to sanitize their research data and then publish the sanitized data along with mining models. Other researchers can verify the published models using the published data. They can also try other mining methods on sanitized data to solve the same research problem. Application of our approach may potentially reduce both research fraud and encourage sharing of research data.

As future work, we will investigate whether our approach can be extended to preserve other types of data mining models. Further, it will be interesting to study whether a privacy protection method can preserve the relative order of performance of different mining models. For example, suppose on the original dataset a mining model A (e.g., a decision tree model) is superior to a different mining model B (e.g., a naïve Bayesian model), it will be desirable if model A is still better than model B in the sanitized data.

REFERENCES

AGGARWAL, C. C. AND YU, P. S. 2004. A condensation approach to privacy preserving data mining. In *Proceedings of the International Conference on Extending Database Technology (EDBT'04).*

AGRAWAL, R. AND SRIKANT, R. 2000. Privacy preserving data mining. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*. 439–450.

BLACK, A. 2006. Fraud in medical research: A frightening, all-too-common trend on the rise. *Natur. News Apr. 18.*

DALE, A., ARBER, S., AND PROCTER, M. 1988. *Doing Secondary Analysis.* Contemporary Social Research Series No. 17. Unwin Hyman, London.

FEE, R. 2007. The cost of clinical trials. *Drug Discov. Devel. 10*, 3, 32.

FU, Y., CHEN, Z., KORU, A. G., AND GANGOPADHYAY, A. 2009a. A privacy protection technique for publishing data mining models and supporting data. In *Proceedings of the Workshop on Information Technologies and Systems Meetings (WITS'09).*

FU, Y., KORU, A. G., CHEN, Z., AND EMAM, K. E. 2009b. A tree-based approach to preserve privacy of software engineering data and predictive models. In *Proceedings of the International Conference on Predictor Models in Software Engineering.*

GROSSMAN, R. L., KAMATH, C., KEGELMEYER, P., KUMAR, V., AND NAMBURU, R. (EDS.). 2001. *Data Mining for Scientific and Engineering Applications.* Kluwer Academic Publishers, Norwell, MA.

HAMMANN, F., GUTMANN, H., VOGT, N., HELMA, C., AND DREWE, J. 2010. Prediction of adverse drug reactions using decision tree modeling. *Clinic. Pharmacol. Therapeut.* To appear.

HAN, J. AND KAMBER, M. 2000. *Data Mining: Concepts and Techniques.* Morgan Kaufmann Publishers.

HETTICH, S., BLAKE, C. L., AND MERZ, C. J. 1998. UCI repository of machine learning databases. http://www.ics.uci.edu/simmlearn/MLRepository.html

HOUSE OF COMMONS SCIENCE AND TECHNOLOGY COMMITTEE PARLIAMENT OF THE UNITED KINGDOM. 2010. The disclosure of climate data from the climatic research unit at the University of East Anglia. http://www.publications.parliament.uk/pa/cm200910/cmselect/cmsctech/387/387i.pdf

KUIPERS, T. AND HOEVEN, J. V. D. 2009. Insight into digital preservation of research output in Europe. http://www.parse-insight.eu/publications.php#d3-4

LEFEVRE, K., DEWITT, D. J., AND RAMAKRISHNAN, R. 2006a. Mondrian multidimensional K-anonymity. In *Proceedings of the International Conference on Data Engineering (ICDE'06).* 25.

LEFEVRE, K., DEWITT, D. J., AND RAMAKRISHNAN, R. 2006b. Workload-Aware anonymization. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* 277–286.

LI, N., LI, T., AND VENKATASUBRAMANIAN, S. 2007. t-Closeness: Privacy beyond k-anonymity and l-diversity. In *Proceedings of the International Conference on Data Engineering (ICDE'07).*

LI, X.-B. AND SARKAR, S. 2006a. Privacy protection in data mining: A perturbation approach for categorical data. *Inform. Syst. Res. 17*, 3, 254–270.

LI, X.-B. AND SARKAR, S. 2006b. A tree-based data perturbation approach for privacy-preserving data mining. *IEEE Trans. Knowl. Data Engin. 18*, 9, 1278–1283.

MCDAVID, K., SCHYMURA, M. J., ARMSTRONG, L., SANTILLI, L., SCHMIDT, B., BYERS, T., STEELE, C. B., O'CONNOR, L., SCHLAG, N. C., ROSHALA, W., DARCY, D., MATANOSKI, G., SHEN, T., AND BOLICK-ALDRICH, S. 2004. Rationale and design of the National Program of Cancer Registries' breast, colon, and prostate patterns of care study. *Cancer Causes Control 15*, 10, 1057–1066.

MENON, S. AND SARKAR, S. 2007. Minimizing information loss and preserving privacy. *Manag. Sci. 53*, 1, 101–116.

MENON, S., SARKAR, S., AND MUKHERJEE, S. 2005. Maximizing accuracy of shared databases when concealing sensitive patterns. *Inform. Syst. Res. 16,* 3, 256–270.

NATIONAL PROGRAM OF CANCER REGISTRIES (NPCR). 2010. National center for chronic disease prevention and health promotion. http://www.cdc.gov/cancer/npcr/about.htm

VAIDYA, J., CLIFTON, C., AND ZHU, M. 2005. *Privacy Preserving Data Mining.* Springer.

XIAO, X. AND TAO, Y. 2006. Anatomy: Simple and effective privacy preservation. In *Proceedings of the International Conference on Very Large Databases (VLDB'06).* 139–150.

Received April 2010; revised October 2010; accepted October 2010