Workload-Aware Anonymization

Kristen LeFevre[1] David J. DeWitt[1] Raghu Ramakrishnan[1,2]

[1]University of Wisconsin - Madison, 1210 West Dayton St., Madison, WI 53706
[2]Yahoo! Research, 701 First Ave., Sunnyvale, CA 94089

ABSTRACT

Protecting data privacy is an important problem in microdata distribution. Anonymization algorithms typically aim to protect individual privacy, with minimal impact on the quality of the resulting data. While the bulk of previous work has measured quality through one-size-fits-all measures, we argue that quality is best judged with respect to the workload for which the data will ultimately be used.

This paper provides a suite of anonymization algorithms that produce an anonymous view based on a target class of workloads, consisting of one or more data mining tasks, as well as selection predicates. An extensive experimental evaluation indicates that this approach is often more effective than previous anonymization techniques.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications

General Terms

Algorithms, Experimentation, Security

Keywords

Privacy, Anonymity, Data Recoding, Predictive Modeling

1. INTRODUCTION

k-Anonymity [22, 23] and l-diversity [18] have been studied widely as mechanisms for preventing *re-identification* attacks in microdata release. Of course, subject to the given anonymity constraints, the data should remain as useful as possible. Unfortunately, there is often a tension between these two goals.

It is our position that the best way of measuring quality is based on the task for which the data will ultimately be used. This paper provides anonymization techniques that incorporate a target workload of selections and mining tasks.

1.1 Motivating Example

Suppose that a trusted agency compiles a database of disease information for several million hospital patients. However, the agency is prohibited by law from distributing this data without taking precautions to ensure individual privacy. For example, the agency should take steps to guarantee that the released data does not reveal any individual's HIV status.

Alice is an external researcher who is directing two separate studies, each of which could benefit from using the data in the central database. As part of the first study, Alice wants to build a classification model that uses age, smoking history, and HIV status to predict life expectancy. In the second study, she would like to find combinations of variables that are useful for predicting elevated cholesterol and obesity in males over 40.

In this situation, it is desirable to distribute anonymized microdata to individuals like Alice (the *data recipients*).[1] One might consider a simpler protocol, in which Alice requests a specific model, constructed entirely by the agency. However, there are two downsides to this approach. First, the simple model-distribution protocol assumes that the tasks are fully-specified at the time of the initial request. However, in our example, Alice's second study involves an entire class of models, each constructed using a subset of the data (attributes and records). Indeed, workloads like this arise naturally in certain types of exploratory data analysis [9].

Also, the inference implications of releasing one or more models constructed on the agency's unmodified data are not well-understood. Each such model reveals something about the distributional characteristics of the agency's data, and in certain cases, the revealed information might constitute a breach of privacy. However, in the case of a single released view, there are well-defined notions of anonymity, and the best Alice can do is to approximate the distribution in the (sanitized) data she is given.

The work presented in this paper is motivated by this type of scenario, where the goal is to create a single view of the database that respects all given anonymity constraints, but that remains useful for carrying out the tasks in a target class of workloads.

1.2 Paper Overview and Contributions

We begin by reviewing the problems of anonymity, classification, and regression in Section 2. Because previous defin-

[1] We assume that Alice only receives one version of any given data set and that she does not collude with others receiving data distributions from the same source.

KDD'06, August 20–23, 2006, Philadelphia, Pennsylvania, USA.
Copyright 2006 ACM 1-59593-339-5/06/0008 ...$5.00.

itions of anonymity with respect to a sensitive attribute (i.e., l-diversity [18]) have assumed that the sensitive attribute is nominally-valued, we also propose a novel diversity requirement for numeric attributes.

Our first main contribution, described in Section 3, is a suite of algorithms for generating an anonymous data snapshot, while preserving the utility of the data with respect to a target class of workloads. While previous work has considered incorporating a single classifier (constructed over the entire released data set) [13, 15, 24], we incorporate the following expressive workload characteristics:

- **Classification & Regression** We incorporate models predicting both categorical and numeric attributes.
- **Multiple Target Models** Often, the data recipient will want to build separate models to predict multiple different attributes.
- **Selection & Projection** Frequently, one or more of the mining tasks will involve only a subset of the data (e.g., males over 40). In this case, it is important to guarantee that this data can be precisely and accurately selected from the released snapshot. Similarly, it is important to guarantee that the data remains useful when only a subset of the released attributes is used for a particular task.

Our second main contribution is an extensive experimental evaluation, described in Section 4. The results show that our anonymization algorithms are often more effective than previous algorithms in producing high-quality data, as judged by a variety of workloads.

Much of the previous work on k-anonymity has measured data quality or optimality using simple measures based on equivalence class size or the total number of generalizations/suppressions [2, 5, 17, 19, 22, 23]. Not surprisingly, our experiments also show that one-size-fits-all measures are not necessarily indicative of quality with respect to a particular workload.

In order to assess the impact of anonymization on subsequent analysis techniques, we first had to address some additional problems. Because standard learning algorithms use point data for training, rather than the region data produced by multidimensional recoding, Section 4.2 proposes a pre-processing step for converting regions to points. Following pre-processing, standard learning algorithms can be applied *without modification.*

The paper concludes with discussions of related and future work in Sections 5 and 6.

2. PRELIMINARIES

K-anonymity [22, 23] and l-diversity [18] were proposed to limit re-identification risk in microdata publishing. Consider a single relation T. In defining anonymity, each attribute in T is characterized by at most one of the following types:

- **Unique Identifiers** A *unique identifier* is any attribute that identifies individuals (e.g., SS#). Known identifiers are typically removed entirely from released microdata.
- **Quasi-identifier** $(Q_1, ..., Q_d)$ A *quasi-identifier* is a minimal set of attributes that can be joined with external information to re-identify individual records. We assume that a quasi-identifier is recognized based on knowledge of the domain.
- **Sensitive attributes (S)** An attribute is considered *sensitive* if an adversary should not be permitted to uniquely associate its value with a unique identifier. For example, the *HIV Status* field in released medical data would likely be considered sensitive. Previous work assumed a single, nominally-valued, sensitive attribute [18]; we also propose an extension to a numeric sensitive attribute.

The k-anonymity requirement is quite simple. Intuitively, it stipulates that no individual record should be uniquely identifiable from a group of k on the basis of its quasi-identifier values. We will refer to each group of tuples in T with identical quasi-identifier values as an *equivalence class.*

K-Anonymity [22, 23] A table T is k*-anonymous* with respect to quasi-identifier set $Q_1, ..., Q_d$ if every unique tuple $\langle q_1, ..., q_d \rangle$ in the (multiset) projection of T on $Q_1, ..., Q_d$ occurs at least k times.

l-Diversity [18] provides a natural extension, incorporating a nominal sensitive attribute S. The l-diversity principle requires that each equivalence class (as defined by k-anonymity) also contain at least l "well-represented" distinct values for S. This principle can be instantiated in various ways. The strictest proposal formulates l-diversity in terms of entropy. Because entropy is concave, entropy l-diversity requires that the full database have entropy at least $log(l)$. D_S denotes the (finite) domain of attribute S.

Entropy l-Diversity (Nominal S) [18] A table T is entropy l-diverse with respect to quasi-identifier set $Q_1, ..., Q_d$ and sensitive attribute S if, for every equivalence class E in T, $\sum_{s \in D_S} -p(s|E) log\ p(s|E) \geq log(l)$, where $p(s|E)$ is the fraction of tuples in E with $S = s$.

For numeric sensitive attributes, diversity is more subtle. For example, if $S = Salary$, an equivalence class containing salaries $\{100K, 101K, 102K\}$ is considered 3-diverse, but intuitively does not protect privacy as well as an equivalence class containing salaries $\{1K, 50K, 500K\}$. For this reason, we define a new diversity requirement that guarantees a certain level of dispersion within each equivalence class:

Squared-Error Diversity (Numeric S) Table T is squared-error diverse with respect to quasi-identifier set $Q_1, ..., Q_d$ and sensitive attribute S if, for every equivalence class E in T, $\sum_{i \in E}(s_i - \overline{s}(E))^2 \geq error$, where $\overline{s}(E)$ is the mean value of S in E, and $error$ is the diversity parameter.

2.1 Classification & Regression

In classification/regression, attributes are typically characterized by at most one of the following types:

- **Target attribute (C or R)** The goal of classification is to build a model that accurately predicts the value of a nominal class label (C). Regression aims to predict a numeric attribute (R).
- **Predictor attributes** Some set of (discrete or continuous) predictor attributes (also commonly called *features*) are used to predict the target attribute.

When a target classification or regression model is considered in conjunction with anonymity, each attribute has two

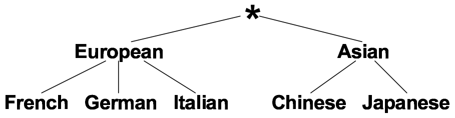


Figure 1: A possible value generalization hierarchy for the Nationality domain

characterizations. In the remainder of this paper, we will assume that the set of predictor attributes is a quasi-identifier. Under this assumption, it is contradictory to categorize an attribute as both target and sensitive, and we disallow this categorization.

2.2 Recoding

Numerous *recoding* techniques have been proposed for sanitizing microdata to satisfy an anonymity constraint. In a relational database, each attribute X has a domain of values D_X. A *global recoding* achieves anonymity by mapping the quasi-identifier domains to ranges or coarsened values.

Global recoding can be broken down into two sub-classes [16, 17]. If the quasi-identifier consists of d attributes (Q_1, ..., Q_d), a *single-dimensional global recoding* is defined by a set of functions $\phi_1, ..., \phi_d$ such that each $\phi_i : D_{Q_i} \to D'$. An *anonymous view* V of T is obtained by applying each ϕ_i to the value of Q_i in each tuple of T.

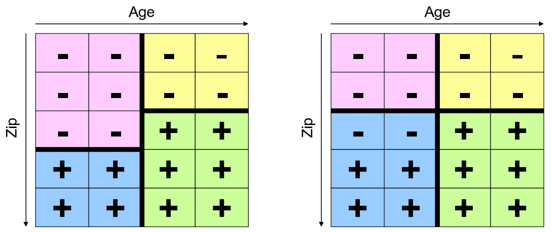
On the other hand, a *multidimensional global recoding* is defined by a *single* function $\phi : D_{Q_1} \times ... \times D_{Q_d} \to D'$, which is used to recode the domain of unique *vectors* associated with the quasi-identifier. In this case, V is obtained by applying ϕ to the vector of quasi-identifier values in each tuple of T.

For attributes with continuous or ordinal (ordered categorical) domains, it is convenient to think of each vector of quasi-identifier values $\langle q_1, ..., q_d \rangle$ as a point in a d-dimensional space. A class of multidimensional recoding models partitions the domain space into non-overlapping d-dimensional rectangular regions [17]. Recoding function ϕ is defined by mapping each point to the region in which it is contained. Thus, each region corresponds to an equivalence class in anonymous view V.[2]

When the domain of a quasi-identifier attribute is nominal, this partitioning may be further constrained by a user-defined *value generalization hierarchy*, or partial order, as described by Samarati and Sweeney [22, 23]. For example, Figure 1 shows a possible hierarchy for the Nationality domain; the domain values are found at the leaves. The notation *French* $\preceq$ *European* indicates that *French* is descended from *European* in the hierarchy.

The hierarchy can be used in several ways to constrain the set of possible recodings [16]. In this paper, *within a particular d-dimensional region*, we require that if ϕ maps a leaf value v to some ancestor a, then all leaves that are descended from a must also be mapped to a.

Every single-dimensional recoding can be equivalently expressed as a multidimensional recoding, but the reverse is frequently not true [17]. Depending on the distribution of the data, this can affect data quality. For example, consider a dataset with exactly two predictors/quasi-identifiers (Age and Zip). Suppose the distribution of class labels $(+, -)$ is as shown in Figure 2, and that $k = 3$. In this case, there is a k-anonymous multidimensional recoding that groups together only records with like labels, but this cannot be accomplished with single-dimensional recoding, which requires that the values of each attribute be recoded uniformly.



(a) Multidimensional (b) Single-Dimensional

Figure 2: Comparing multidimensional and single-dimensional recoding in two dimensions

3. WORKLOAD-AWARE ANONYMIZATION

This section proposes several algorithms for creating a single snapshot of a given data set that respects a given anonymity constraint, but remains useful for executing a particular class of workloads. The target class of workloads is specified by the following parameters:

1. A set of predictor attributes $(Q_1, ..., Q_d)$
2. Either a set of one or more nominal target class labels $(C_1, ..., C_m)$, or numeric target attributes $(R_1, ..., R_m)$
3. Optionally, a set of selection predicates $(PR_1, ..., PR_n)$

The anonymity constraint is k-anonymity, optionally extended by l-diversity or squared-error diversity. Also, we assume that the predictor attributes are a quasi-identifier.

In the simplest case, when the target workload consists of one classification or regression model, without selection predicates, the heuristics used by our algorithms implement entropy l-diversity and squared-error diversity in reverse.

3.1 Single Target Classification Model

The Mondrian algorithm was recently proposed for k-anonymization using multidimensional recoding [17]. The algorithm is based on a greedy recursive partitioning of the (multidimensional) quasi-identifier domain space (see Figure 3). In order to obtain approximately uniform partition occupancy, [17] suggests recursively choosing the split attribute with the largest normalized range of values, and (for continuous or ordinal attributes) partitioning the data around the median value of the split attribute. This process is repeated until no *allowable* split remains, meaning that a particular region cannot be further divided without violating the anonymity constraint, or constraints imposed by value generalization hierarchies. We refer to this algorithm as **Median Mondrian**.

When the (set of) target mining model(s) is known, we can improve this heuristic. First consider a single target classification model, with predictor attributes $Q_1, ..., Q_d$ (also the quasi-identifier) and class label C. In this case, we propose a heuristic partitioning scheme based on information gain, which is reminiscent of decision tree construction. Intuitively, the goal of this greedy criterion is to produce homogeneous partitions of class labels.

At each recursive step, we choose the split that minimizes the weighted entropy over the set of resulting partitions

[2] Hyper-rectangular regions are easily expressed in tabular form using range values (e.g., Age = [20-35]).

(without violating the anonymity constraint). P denotes the current (recursive) tuple set, and *partitions* P' denotes the set of partitions resulting from the candidate split. $p(c|P')$ is the fraction of tuples in P' with class label $C = c$. We refer to this algorithm as **InfoGain Mondrian**.

$$Entropy(P, C) = \sum_{partitions\ P'} \frac{|P'|}{|P|} \sum_{c \in D_C} -p(c|P') log p(c|P') \tag{1}$$

InfoGain Mondrian handles continuous quasi-identifier values as they are typically handled by decision-trees, partitioning around the *threshold* value with smallest entropy (see [12]). The data is first sorted with respect to the split attribute. Then the data is scanned, and each time there is a change in class label, this *candidate threshold* is checked with respect to anonymity and entropy. In the event that no candidate threshold satisfies the anonymity constraint, the median is also checked as a default.

InfoGain Mondrian scales to large data sets through a straightforward adaptation of an existing scalable decision-tree induction scheme, such as RainForest [14].

3.2 Single Target Regression Model

Similar greedy heuristics can be used when the target attribute is numeric. Specifically, we use the *mean squared error (MSE)* to measure the impurity of target attribute R within a candidate partition P'. A heuristic inspired by the CART algorithm for regression trees [7] recursively chooses the split that minimizes the weighted sum of MSEs over the set of resulting partitions. $\overline{r}(P')$ denotes the mean value of R in P'.

$$\begin{aligned} MSE(P') &= \frac{1}{|P'|} \sum_{i \in P'} (r_i - \overline{r}(P'))^2 \\ Weighted\ MSE &= \sum_{Partitions\ P'} \frac{|P'|}{|P|} (MSE(P')) \\ &= \frac{1}{|P|} \sum_{Partitions\ P'} \sum_{i \in P'} (r_i - \overline{r}(P'))^2 \end{aligned}$$

Because $|P|$ is constant for all candidate splits, the algorithm chooses the split that minimizes the following expression (without violating anonymity). We call this **Least Squared Deviance (LSD) Mondrian**. This algorithm handles continuous attributes through discretization.

$$Error^2(P, R) = \sum_{Partitions\ P'} \sum_{i \in P'} (r_i - \overline{r}(P'))^2 \tag{2}$$

3.3 Multiple Target Models

In certain cases, we would like to allow the data recipient to build several models, to accurately predict the *marginal distributions* of several class labels $(C_1, ..., C_m)$ or regression attributes $(R_1, ..., R_m)$. InfoGain Mondrian and LSD Mondrian can be extended to handle multiple discrete and numeric target attributes, respectively.

For classification, there are two ways to make this extension. In the first approach, the data recipient would build a single model to predict the *vector* of class labels, $\langle C_1, ..., C_m \rangle$, which has domain $D_{C_1} \times ... \times D_{C_m}$. A greedy split criterion would minimize entropy with respect to this single variable.

However, in this simple approach, the size of the domain grows exponentially with the number of target attributes. To avoid potential problems due to data sparsity, we instead simplify the problem by assuming independence among target attributes. This is a reasonable assumption because we are ultimately only concerned about the marginal distribution of each target attribute. Under the independence assumption, a greedy split criterion minimizes the sum of weighted entropies:

```
Anonymize($tuples$, $attrs$)
  if (no allowable split for $tuples$)
    return $\phi : t \in tuples \rightarrow$ bounding region(tuples)
  else
    $best \leftarrow$ Choose_Attribute($attrs$, $tuples$)
    if continuous($best$) or ordinal($best$)
      $threshold \leftarrow$ Choose_Threshold($best$)
      $lhs \leftarrow \{t \in tuples : t.best \leq threshold\}$
      $rhs \leftarrow \{t \in tuples : t.best > threshold\}$
      return Anonymize($rhs$,$attrs$) $\cup$ Anonymize($lhs$,$attrs$)
    else if nominal($best$)
      $recodings \leftarrow \{\}$
      for each child $v_i$ of $root(best.hierarchy)$
        $tuples_i \leftarrow \{t \in tuples : t.best \preceq v_i\}$
        $attrs' \leftarrow$ replace $root(best.hierarchy)$ with $v_i$ in $attrs$
        $recodings \leftarrow recodings \cup$ Anonymize($tuples_i$, $attrs'$)
      return $recodings$
```

Figure 3: Basic Mondrian algorithm

$$\sum_{i=1}^{m} Entropy(P, C_i) \tag{3}$$

In regression (the squared error split criterion in particular), there is no analogous distinction between treating the set of target attributes as a single variable and assuming independence. For example, if we have two target attributes, R_1 and R_2, the joint error is the distance between an observed point (r_1, r_2) and the centroid $(\overline{r_1}(P), \overline{r_2}(P))$ in 2-dimensional space. The squared joint error is just the sum of individual squared errors, $(r_1 - \overline{r_1}(P))^2 + (r_1 - \overline{r_2}(P))^2$. For this reason, the greedy split criterion minimizes the sum of squared error:

$$\sum_{i=1}^{m} Error^2(P, R_i) \tag{4}$$

3.4 Incorporating Selection

Sometimes one or more of the tasks in the target workload will use only a subset of the released data, and it is important that this data can be selected precisely, despite recoding. For example, a researcher may want to build a model using only males over 40, but this is difficult if the ages of some men are recoded to the range $[30 - 50]$. This problem was originally described in [17].

Consider a set of selection predicates $(PR_1, ..., PR_m)$ defined by boolean functions of the quasi-identifier attributes $(Q_1, ..., Q_d)$. Conceptually, each PR_i defines a *query region* R_i in the domain space such that $R_i = \{p \in D_{Q_1} \times ... \times D_{Q_d} : PR_i(p) = true\}$. For the purposes of this work, we only consider selections for which the query region can be expressed as a hyper-rectangle. (Some additional selections can be decomposed into two or more hyper-rectangles, and incorporated as separate queries.)

A multidimensional recoding function ϕ divides the domain space into non-overlapping regions $P_1, ..., P_n$. Formally, the *recoding region* $P_i = \{p \in D_{Q_1} \times ... \times D_{Q_d} : \phi(p) = p'_i\}$, where p'_i is a particular generalization of the

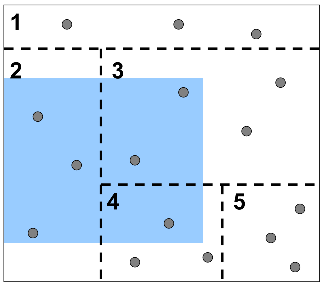


Figure 4: Selection example

quasi-identifier vector. When evaluating PR_i over the recoded view V, it may be that no subset of the recoding regions can be combined to produce query region R_i. Instead, it is intuitive to return the tuples from V that are contained in any recoding region *overlapping* R_i. More formally,

$$\begin{aligned} Overlap(R_i) &= \cup\{P_i : P_i \cap R_i \neq \phi\} \\ PR_i(V) &= \{\phi(p) : \phi(p) \in V \wedge p \in Overlap(R_i)\} \end{aligned}$$

Notice that this will often produce a larger result set than evaluating PR_i over the original table T; the *imprecision* is the difference in size between these two result sets.

$$imprecision(PR_i, \{P_1, ..., P_n\}) = |PR_i(V)| - |PR_i(T)| \quad (5)$$

For example, Figure 4 shows a 2-dimensional domain space. The shaded area represents a query region, and the tuples of T are represented by points. The recoding regions are bounded by dotted lines and numbered. Recoding regions 2, 3, and 4 overlap the query region. If we evaluated this query using the original data, the result set would include 6 tuples. However, evaluating the query using the recoded data yields 10 tuples, an imprecision of 4.

Ideally, the goal of selection-oriented anonymization is to divide the domain space into a set of (anonymous) recoding regions that minimize imprecision for the set of target predicates. We incorporate this goal into the Mondrian algorithm through a new greedy splitting heuristic. Specifically, at each recursive step, when partitioning a recursive region P, we choose the split that minimizes the total imprecision for the set of resulting regions $\{P'_1, ..., P'_n\}$:

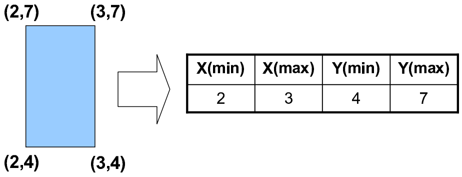
$$\sum_{i=1}^{m} imprecision(PR_i, \{P'_1, ..., P'_n\}) \quad (6)$$

The algorithm proceeds until there is no allowable split that reduces the imprecision of the current partition P, and continuous attributes are handled through discretization. We will call this algorithm **Selection Mondrian**.

In practice, we expect this technique to be used most often for simple selections, such as breaking down health data by state. After incorporating selections, we continue to anonymize each resulting partition independently, using the appropriate classification- or regression-oriented algorithm.

4. EXPERIMENTAL EVALUATION

Our experimental evaluation has several goals, the first of which is to provide some insight about quality evaluation methodology. We describe an experimental protocol for evaluating an anonymization algorithm with respect to a target data mining workload, and we compare the results to those obtained using some simpler quality measures.



X(min)	X(max)	Y(min)	Y(max)
2	3	4	7

Figure 5: Mapping a d-dimensional rectangular region to $2 * d$ attributes

The second goal is to evaluate the algorithms described in Section 3. In particular, we assess the impact of incorporating a set of target classification or regression models into the anonymization, and multidimensional recoding. Also, we evaluate the effectiveness of our algorithms with respect to selections, projections, and multiple target models.

4.1 Methodology

Given a target classification or regression workload, the most direct way to evaluate the quality of an anonymization is by training each target model using the anonymized data, and evaluating the resulting models using *predictive accuracy* (classification), *mean absolute error* (regression), or similar measures. We will call this methodology *model evaluation*. All of our model evaluation experiments follow a common protocol:

1. The data is first divided into training and testing sets (or 10-fold cross-validation sets), T_{train} and T_{test}.
2. The anonymization algorithm determines recoding function ϕ using only the *training set* T_{train}. Anonymous view V_{train} is obtained by applying ϕ to T_{train}.
3. The same recoding function ϕ is then applied to the *testing set* (T_{test}), yielding V_{test}.
4. The classification or regression model is trained using V_{train}, and tested using V_{test}.

This experimental design is different from the setup used by Fung et al. [13] for an important reason. In [13], the *combined* training and testing sets were anonymized using a single-dimensional recoding algorithm based on information gain. Following this step, the data was separated into training and testing sets. In our opinion, this setup is inappropriate for evaluating the anonymization algorithm because incorporating the test set when choosing a recoding is tantamount to looking at the test set while doing feature selection. Instead, all of our experiments hold out the test set during both the anonymization and training phases.

We used k-anonymity as the anonymity constraint, and we used the implementations of the following learning algorithms provided by the Weka software package [25]:

- **Decision Tree (J48)** Default settings were used.
- **Naive Bayes** Supervised discretization was used for continuous attributes; otherwise all default settings were used.
- **Random Forests** Each classifier was comprised of 40 random trees, and all other default settings were used.
- **Support Vector Machine (SMO)** Default settings were used, including a linear kernel function.
- **Linear Regression** Default settings were used.
- **Regression Tree (M5)** Default settings were used.

Attribute	Distribution	Generalize
salary	Uniform in [20,000, 150,000]	continuous
commission	If salary $\geq$ 75,000, then 0 Else Uniform in [10,000, 75,000]	continuous
age	Uniform integer in [20,80]	continuous
elevel	Uniform integer in [0, 4]	hierarchy
car	Uniform integer in [1, 20]	hierarchy
zipcode	Uniform integer in [0, 9]	continuous
hvalue	zipcode * h * 100,000 where h uniform in [0.5, 1.5]	continuous
hyears	Uniform integer in [1, 30]	continuous
loan	Uniform in [0, 500,000]	continuous

Function	**Group A**
2	$((age < 40) \wedge (50K \leq salary \leq 100K)) \vee$ $((40 \leq age < 60) \wedge (75K \leq salary \leq 125K)) \vee$ $((age \geq 60) \wedge (25K \leq salary \leq 75K))$
4	$((age < 40) \wedge$ $(((elevel \in \{0,1\})?(25K \leq salary \leq 75K))$ $: (50K \leq salary \leq 100K)))) \vee$ $((40 \leq age < 60) \wedge$ $(((elevel \in \{1,2,3\})?(50K \leq salary \leq 100K))$ $: (75K \leq salary \leq 125K)))) \vee$ $((age \geq 60) \wedge$ $(((elevel \in \{2,3,4\})?(50K \leq salary \leq 100K))$ $: (25K \leq salary \leq 75K))))$
6	$((age < 40) \wedge$ $(50K \leq (salary + commission) \leq 100K)) \vee$ $((40 \leq age < 60) \wedge$ $(75K \leq (salary + commission) \leq 125K)) \vee$ $((age \geq 60) \wedge$ $(25K \leq (salary + commission) \leq 75K))$
7	$disposable = .67 \times (salary + commission)$ $-.2 \times loan - 20K$ $disposable > 0$

Figure 6: Synthetic predictor/quasi-identifier attributes and class label functions

In addition to model evaluation, we also measured certain characteristics of the anonymized training data to see if there was any correlation between these simpler measures and the results of the model evaluation. Specifically, we measured the *average equivalence class size*, and for classification tasks, we measured the *conditional entropy* of the class label given the partitioning:

$$H(C|P) = \sum_{partitions\ p} p(p) \sum_{classes\ c} -p(c|p)\ log\ p(c|p) \qquad (7)$$

4.2 Learning from Regions

When single-dimensional recoding is used, standard learning algorithms can be applied directly to the resulting point data, notwithstanding the "coarseness" of some points [13]. Although multidimensional recoding techniques are more flexible, using the resulting hyper-rectangular data to train standard data mining models poses an additional challenge.

To address this problem, we make a simple observation. Because we restrict the recoding regions to include only d-dimensional hyper-rectangles, each region can be uniquely represented as a point in $(2 * d)$-dimensional space. For example, Figure 5 shows a 2-dimensional rectangle, and its unique representation as a 4-tuple. This assumes a total order on the values of each attribute, similar to the assumption made by support vector machines.

Following this observation, we adopt a simple pre-processing technique for learning from regions. Specifically, we extend the recoding function ϕ to map data points to d-dimensional regions, and in turn, to map these regions to their unique representations as points in $(2 * d)$-dimensional space.

Our primary goal in developing this technique is to establish the utility of our anonymization algorithms. There are many possible approaches to the general problem of learning from regions. For example, Zhang and Honavar proposed an algorithm for learning decision trees from attribute values at various levels of a taxonomy tree [26]. However, a full comparison is beyond the scope of this paper.

Census Database

Attribute	**Dist. Vals**	**Generalization**
Region	57	hierarchy
Age	77	continuous
Citizenship	5	hierarchy
Marital Status	5	hierarchy
Education (years)	17	continuous
Sex	2	hierarchy
Hours per week	93	continuous
Disability	2	hierarchy
Race	9	hierarchy
Salary	2/continuous	target

Contraceptives Database

Attribute	**Dist. Vals**	**Generalization**
Wife's age	34	continuous
Wife's education	4	hierarchy
Husband's education	4	hierarchy
Children	15	continuous
Wife's religion	2	hierarchy
Wife working	2	hierarchy
Husband's Occupation	4	hierarchy
Std. of Living	4	continuous
Media Exposure	2	hierarchy
Contraceptive	3	target

Figure 7: Summary of real-world data sets

4.3 Experimental Data

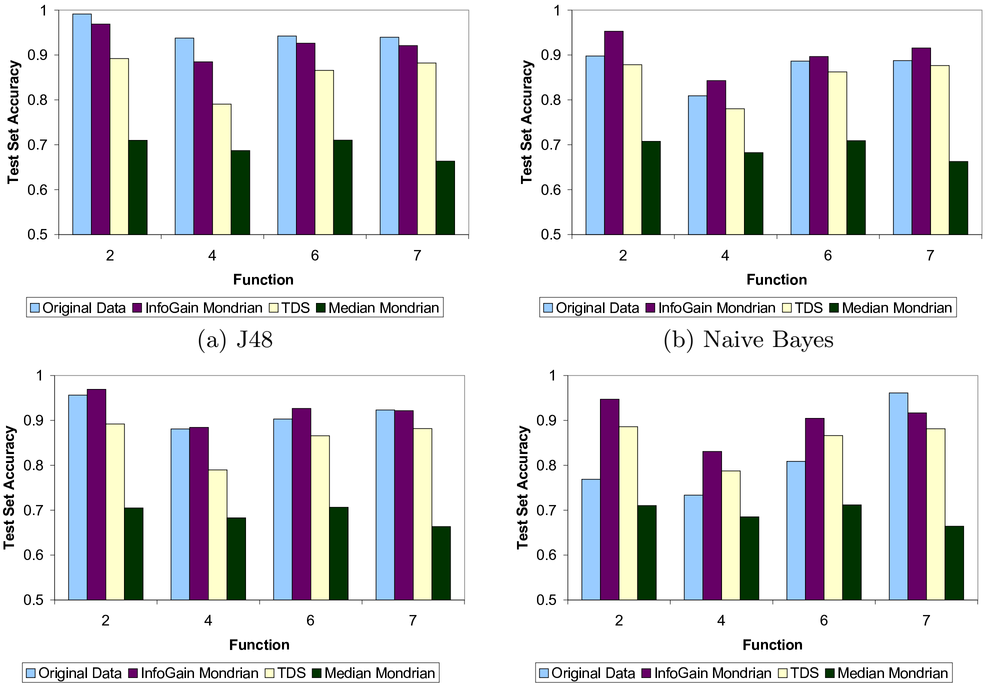
Our first set of experiments used **synthetic data** based on the classification generator introduced by Agrawal et al. [3]. Predictor/quasi-identifier attributes were generated according to the distributions described in Figure 6, and class labels were generated as a function of the predictor values. We present results for four representative label functions, chosen from the original ten (functions 2,4,6,7). To simplify the evaluation, we applied the labeling functions deterministically, without injecting noise.

Notice that the basic labeling functions in Figure 6 include a number of constants (e.g., $75K$). In order to get a more robust understanding of the behavior of the various anonymization algorithms, for functions 2, 4, and 6, we instead generated many independent data sets, varying the function constants independently at random over the range of the attribute.

Figure 6 notes, for each predictor/quasi-identifier attribute, whether it was treated as continuous or nominal (with an associated generalization hierarchy) during anonymization.

In addition to the synthetic data, we also used several **real-world data sets**. The first was derived from a sample of the 2003 Public Use Microdata, distributed by the United States Census American Community Survey[3], with target attribute Salary. This data was used for both classification and regression, and contained 49,657 records. For classification, we replaced the numeric Salary with a Salary class ($< 30K$ or $\geq 30K$); approximately 56% of the data records had Salary $< 30K$. For classification, this is similar to the

[3] http://www.census.gov/acs/www/index.html



(a) J48

(b) Naive Bayes

(c) Random Forests

(d) SVM

Figure 8: Average predictive accuracy for models trained using anonymized synthetic data (k=25)

Adult database from the UCI Machine Learning Repository [6], which has been used in numerous k-anonymity evaluations. However, we chose to compile a new data set that can be used for both classification and regression.

The second real data set is the smaller Contraceptives database from the UCI Repository, which contained 1,473 records after removing those with missing values. This data includes nine socio-economic indicators, which are used to predict the choice of contraceptive method (*long-term*, *short-term*, or *none*) among sampled Indonesian women. Summaries of both real data sets are provided in Figure 7.

4.4 Comparison with Previous Algorithms

InfoGain Mondrian and LSD Mondrian combine multidimensional recoding with classification- and regression-oriented splitting heuristics. In this section, we evaluate the effects of these two components through a comparison with two previous anonymization algorithms. All of the experiments in this section consider a single target model, constructed over the entire anonymized training set.

Several previous algorithms have incorporated a single target classification model while choosing a single-dimensional recoding [13, 15, 24]. To gage the impact of multidimensional recoding, we compared InfoGain Mondrian and the greedy **Top-Down Specialization (TDS)** algorithm [13]. Also, multidimensional recoding was used in **Median Mondrian** [17], without regard to workload. We compare this to InfoGain Mondrian and LSD Mondrian to gage the effects of incorporating a target model.

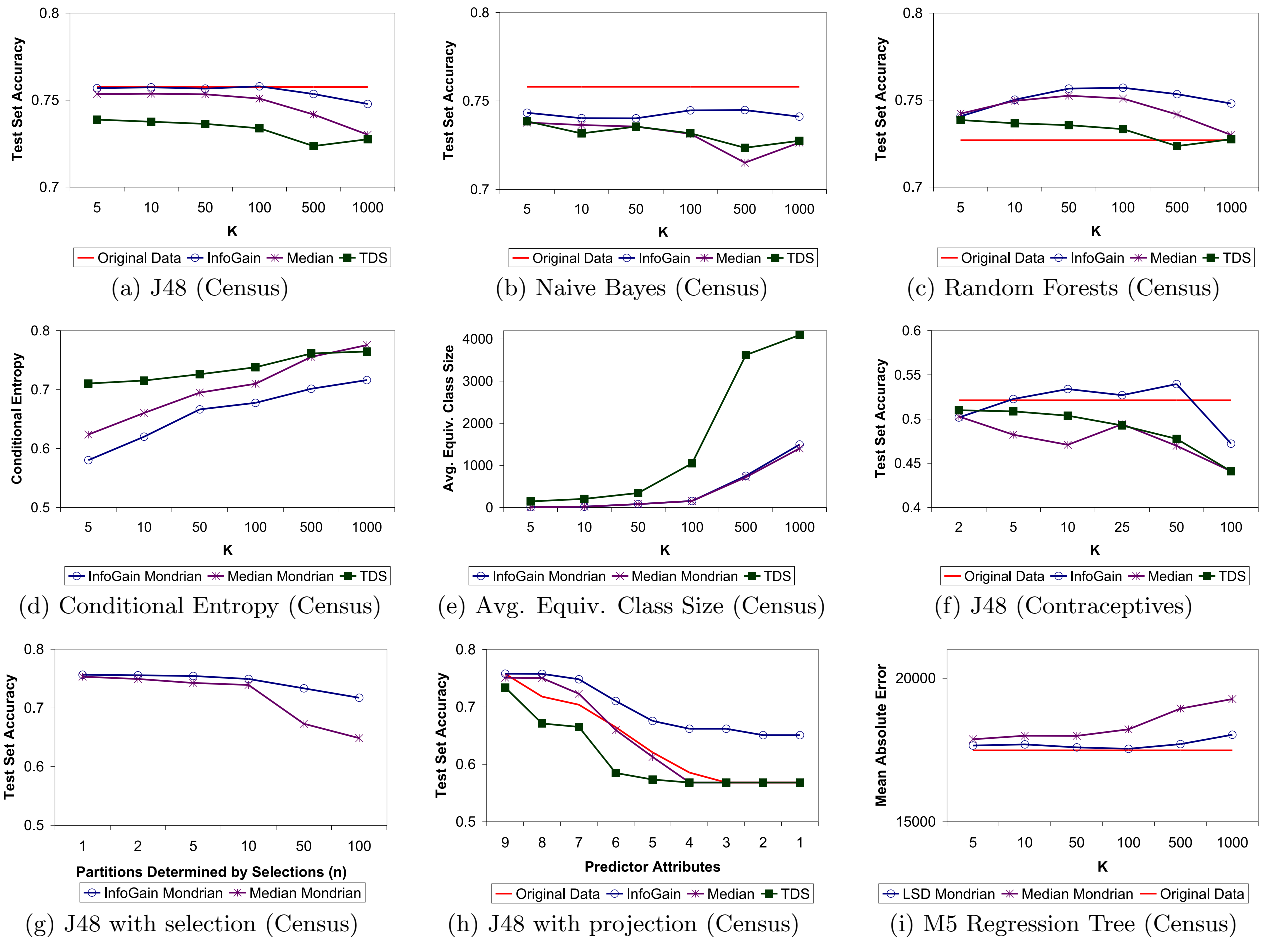
Using the synthetic data, Figure 8 compares the predictive accuracy of classifiers trained on data produced by the different anonymization algorithms. In these experiments, we generated 100 independent training and testing sets, each containing 1000 records, and we fixed $k = 25$. The results are averaged across these 100 trials. For comparison, we also include the accuracies of classifiers trained on the (not anonymized) original data.

InfoGain Mondrian consistently outperforms both TDS and Median Mondrian, a result that is overwhelmingly significant based on a series of paired t-tests. It is important to note that the pre-processing step used to convert regions to points (Section 4.2) is only used for the multidimensional recodings; the classification algorithms run unmodified on the single-dimensional recodings produced by TDS [13]. Thus, should a better technique be developed for learning from regions, this would improve the results for InfoGain Mondrian, but it would not affect TDS.[4]

We performed a similar set of experiments using the real-world data. Figures 9(a,b,c) show results for the Census classification data, for increasing k. The graphs show test set accuracy (averaged across 10 folds) for three learning algorithms. The variance across the folds was quite low, and the differences between InfoGain Mondrian and TDS, and between InfoGain Mondrian and Median Mondrian, were highly significant based on paired t-tests.

It is important to point out that in certain cases, notably Random Forests, the learning algorithm overfits the model when trained using the original data. For example, the model for the original data in Figure 9(c) gets 97% accuracy on the training set, but only 73% accuracy on the test set. When overfitting occurs, it is not surprising that the models trained on anonymized data obtain higher accuracy because anonymization acts as a form of feature selection/construction. Interestingly, we also tried applying a traditional form of feature selection (ranked feature selection based on information gain) to the original data,

[4] Note that by mapping to $2 * d$ dimensions, we effectively expand the hypothesis space considered by the linear SVM. Thus, it is not surprising that this improves accuracy for the non-linear class label functions (Figure 8(d)).



(a) J48 (Census)

(b) Naive Bayes (Census)

(c) Random Forests (Census)

(d) Conditional Entropy (Census)

(e) Avg. Equiv. Class Size (Census)

(f) J48 (Contraceptives)

(g) J48 with selection (Census)

(h) J48 with projection (Census)

(i) M5 Regression Tree (Census)

Figure 9: Comparing anonymization techniques using real-world data

and this did not improve the accuracy of random forests for any number of chosen attributes. We suspect that this discrepancy is due to the flexibility of the recoding techniques. Single-dimensional recoding (TDS) is more flexible than traditional feature selection because it can incorporate attributes at varying levels of granularity. Multidimensional recoding is more flexible still because it (conditionally) incorporates different attributes for different data subsets.

Next, Figures 9(d,e) show conditional entropy and average equivalence class size measurements, averaged across the ten anonymized training folds of the Census classification data. Average equivalence class size, which does not take into account any characteristics of the workload, is not a very good indicator of model accuracy. Conditional entropy, which incorporates the target class label, is a lot better; low conditional entropy generally indicates higher accuracy.

We performed the same set of experiments using the Contraceptives database, and observed similar behavior. InfoGain Mondrian yielded higher accuracy than TDS or Median Mondrian. Results for J48 are shown in Figure 9(f). The remaining results are omitted due to space constraints.

For regression, we found that LSD Mondrian generally led to better models than Median Mondrian. Figure 9(i) shows the mean absolute test set error for the M5 regression tree, using the Census regression data. A similar relative comparison was observed for linear regression, but the overall error was higher because Salary is non-linear.

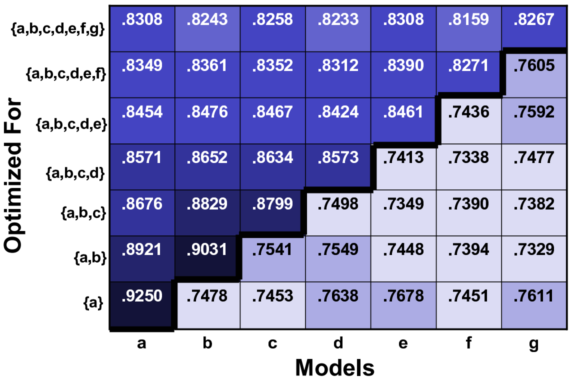
4.5 Multiple Target Models

In Section 3.3 we described a simple adaptation to the basic InfoGain Mondrian algorithm that allowed us to incorporate more than one target attribute, expanding the set of models for which a particular anonymization is "optimized." To evaluate this technique, we performed a set of experiments using the synthetic classification data, increasing the number of class labels.

Figure 10 shows average test set accuracies for J48. We first generated 100 independent training and testing sets, containing 1000 records each. We used synthetic labeling functions 2-6,7, and 9 from the Agrawal generator [3], randomly varying the constants in functions 2-6 as described in Section 4.3.

Each column in the figure (models A-G) represents the average of 25 random permutations of the synthetic functions. The anonymizations (rows in the figure) are "optimized" for an increasing number of target models. (For example, the anonymization in the bottom row is optimized exclusively for model A.) There are two important things to note from the chart, and similar behavior was observed for the other classification algorithms.

- Looking at each model (column) individually, when the model is included in the anonymization (above the bold line), test set accuracy is higher than when the model is not included (below the line).



Optimized For	a	b	c	d	e	f	g
{a,b,c,d,e,f,g}	.8308	.8243	.8258	.8233	.8308	.8159	.8267
{a,b,c,d,e,f}	.8349	.8361	.8352	.8312	.8390	.8271	.7605
{a,b,c,d,e}	.8454	.8476	.8467	.8424	.8461	.7436	.7592
{a,b,c,d}	.8571	.8652	.8634	.8573	.7413	.7338	.7477
{a,b,c}	.8676	.8829	.8799	.7498	.7349	.7390	.7382
{a,b}	.8921	.9031	.7541	.7549	.7448	.7394	.7329
{a}	.9250	.7478	.7453	.7638	.7678	.7451	.7611

Figure 10: Average test set accuracy for multiple incorporated target models (J48, k=25)

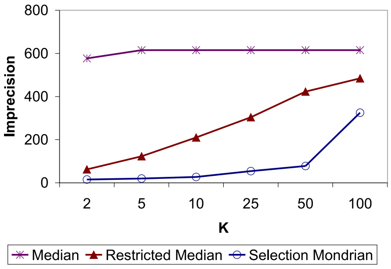


Figure 11: Imprecision for synthetic Function 2

- As we increase the number of included models (moving upward above the line within each column), the test set accuracy tends to decrease. This is because the quality of the anonymization with respect to each individual model is "diluted" by incorporating additional models.

4.6 Selection

In Section 3.4, we discussed the importance of preserving selections, and described an algorithm for incorporating rectangular selection predicates into an anonymization. We conducted an experiment using the synthetic data (1,000 generated records), but treating synthetic Function 2 as a selection predicate. Figure 11 shows the imprecision of this selection when evaluated using the recoded data. The figure shows results for data recoded using three different anonymization algorithms. The first algorithm is Median Mondrian, with greedy recursive splits chosen from amongst all of the quasi-identifier attributes. It also shows a restricted variation of Median Mondrian, where splits are made with respect to only Age and Salary. Finally, it shows the results of Selection Mondrian, incorporating Function 2 as three separate rectangular query regions. It is intuitive that imprecision increases with k, and that imprecision is reduced by incorporating the selection into the anonymization.

Incorporating selections can also affect model quality. In the absence of selections, InfoGain and LSD Mondrian choose recursive splits using a greedy criterion driven by the target model(s). When selections are included, the resulting partitions may not be the same as those that would be chosen based on the target model(s). In the worst case, there may be a selection on an attribute that is uncorrelated with the target attribute.

To test this intuition, we performed an experiment using the Census classification data. To simulate the effect of selections that are uncorrelated with the target model, we first assigned each training tuple to one of n groups, chosen uniformly at random. (We assume $\frac{|Data|}{n} \geq k$.) This mimics the behavior of Selection Mondrian for a set of equality selections on a new attribute, Group number, which takes values $1, ..., n$. We then anonymized each group independently, using either InfoGain Mondrian or Median Mondrian. Once recodings were determined for each training group, we randomly assigned each test tuple to one of the n groups, and recoded the tuple using the recoding function for that group. Finally, we trained a single classification model using the full recoded training set (union of all training groups), and tested using the full recoded test set. This process was repeated for each of ten folds.

The results of this experiment for J48 are shown in Figure 9(g), for increasing n and $k = 50$. As expected, accuracy decreases slightly as the number of selections (n) increases. However, several selections can be incorporated without large negative effects. Similar results were observed for the other classification algorithms.

4.7 Projection

Sometimes not every model constructed by the data recipient will use the full set of predictor attributes; rather, they will use a projected attribute subset. We conducted an experiment to compare anonymization algorithms when only a subset of the released predictor attributes is actually used. First, we ranked the attributes using the original data and a greedy information gain criterion. Then we removed the attributes in order, from most to least predictive, and constructed classification models using the remaining attributes. We fixed $k = 100$.

As expected, test set accuracy decreases as the most predictive attributes are dropped. However, the rate of this decline varies depending on the anonymization algorithm used. Figure 9(h) shows the observed accuracies for J48 using the Census database. Because of the single-dimensional recoding pattern, which is known to preserve fewer attributes over non-uniform quasi-identifier distributions [17], this rate of decay is the most precipitous for TDS.

The results were similar for the other classification algorithms and the Contraceptives data.

5. RELATED WORK

The most closely-related work includes several algorithms that have incorporated a *single* classification model (constructed over the full data set) while choosing a k-anonymous *single-dimensional* recoding. The proposed algorithms include top-down [13] and bottom-up [24] greedy heuristic searches, and genetic algorithms [15]. Each of these papers used the target classification model to evaluate the recoding. Additionally, other recent work suggested using a workload of aggregate queries as a tool for evaluating the quality of anonymizations [17].

Numerous other k-anonymization algorithms have been proposed [2, 5, 16, 19, 22, 23]. However, much of the previous work has sought to optimize simple general-purpose measures of quality, such as the size of equivalence classes, or the total number of generalizations/suppressions.

Aside from k-anonymity, a variety of other methods have been proposed for protecting individual privacy while allowing certain data mining tasks. One widely-studied approach is based on the randomized response paradigm [4, 11, 21]. The main advantage of generalization is that the released data is "truthful," though at a coarsened level of granularity. This allows additional workloads to be carried out using the data, including selection. Generalization also has similar advantages as compared to data swapping [20].

Several cluster-based techniques have also been proposed that are similar in spirit to k-anonymity. The condensation approach first divides the data into "condensation groups" with required minimal occupancy, and then generates point data based on the aggregate statistical properties of each group [1]. Microaggregation first clusters the data into (ideally homogeneous) groups of required minimal occupancy, and then publishes the centroid of each group [10]. However, neither of these approaches requires that the resulting groups be hyper-rectangular, nor do they handle categorical attributes with hierarchical generalization constraints.

Finally, privacy-preserving histogram sanitization was proposed with the similar goal of guaranteeing that individuals blend into a crowd, based on some suitable distance measure [8]. However, the probabilistic privacy definition does not capture situations where the identification of even a single individual would be considered a breach, and the proof of privacy is highly dependent on the original data distribution.

6. CONCLUSION AND FUTURE WORK

k-Anonymity and l-diversity are widely-studied techniques for protecting individual privacy in microdata release. Subject to the anonymity requirement, the data should remain as useful as possible with respect to the workload for which it will ultimately be used.

This paper provided algorithms for incorporating a class of target workloads, consisting of classification or regression models, as well as selection predicates, when generating an anonymous data recoding. An extensive experimental study validated the effectiveness of these algorithms with respect to a variety of workloads. Additionally, our results show that simple quality measures are not always indicative of data quality with respect to a particular workload.

This work also brought to light several interesting opportunities for future work. As described in Section 4.4, anonymization sometimes behaves as a form of feature selection or construction. This has some interesting implications because multidimensional recoding naturally leads to a form of feature selection where different attributes are conditionally retained (at varying levels of granularity) for different data subsets. In the future, it will be valuable to characterize the situations under which this approach leads to better predictive accuracy than traditional feature selection.

Additionally, our selection-oriented anonymization algorithm (Section 3.4) currently only supports selections that can be expressed as rectangular regions. Although we expect simple queries to be the most common, we are working to extend this algorithm to a more expressive class of queries.

Finally, a full study of the learning from regions problem is the topic of future research.

Acknowledgments

Our thanks to Bee-Chung Chen, Hector Corrada Bravo, Ted Wild, and Jude Shavlik for insightful conversations, to Jesse Davis for comments on an earlier draft of this paper, and to Benjamin Fung for providing an implementation of the TDS algorithm.

This work was supported by an IBM Ph.D. fellowship and National Science Foundation Grant IIS-0524671.

7. REFERENCES

[1] C. Aggarwal and P. Yu. A condensation approach to privacy-preserving data mining. In *EDBT*, 2004.

[2] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu. Anonymizing tables. In *ICDT*, 2005.

[3] R. Agrawal, S. Ghosh, T. Imielinski, and A. Swami. Database mining: A performance perspective. In *IEEE Transactions on Knowledge and Data Engineering*, volume 5, 1993.

[4] R. Agrawal and R. Srikant. Privacy-preserving data mining. In *SIGMOD*, 2000.

[5] R. Bayardo and R. Agrawal. Data privacy through optimal k-anonymization. In *ICDE*, 2005.

[6] C. Blake and C. Merz. UCI repository of machine learning databases, 1998.

[7] L. Breiman, J. Freidman, R. Olshen, and C. Stone. *Classification and Regression Trees.* Wadsworth International Group, Belmont, CA, 1984.

[8] S. Chawla, C. Dwork, F. McSherry, and K. Talwar. On the utility of privacy-preserving histograms. In *Uncertainty in Artificial Intelligence*, 2005.

[9] B. Chen, L. Chen, Y. Lin, and R. Ramakrishnan. Prediction cubes. In *VLDB*, 2005.

[10] J. Domingo-Ferrer and J. Mateo-Sanz. Practical data-oriented microaggregation for statistical disclosure control. *IEEE Transactions on Knowledge and Data Engineering*, 4(1), 2002.

[11] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke. Privacy preserving mining of association rules. In *SIGKDD*, 2002.

[12] U. M. Fayyad and K. Irani. On the handling of continuous-valued attributes in decision tree generation. *Machine Learning*, 8:87–102, 1992.

[13] B. Fung, K. Wang, and P. Yu. Top-down specialization for information and privacy preservation. In *ICDE*, 2005.

[14] J. Gehrke, R. Ramakrishnan, and V. Ganti. RainForest: A framework for fast decision tree construction of large datasets. In *VLDB*, 1998.

[15] V. Iyengar. Transforming data to satisfy privacy constraints. In *ACM SIGKDD*, 2002.

[16] K. LeFevre, D.DeWitt, and R. Ramakrishnan. Incognito: Efficient full-domain k-anonymity. In *ACM SIGMOD*, 2005.

[17] K. LeFevre, D. DeWitt, and R. Ramakrishnan. Mondrian multidimensional k-anonymity. In *ICDE*, 2006.

[18] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. l-Diversity: Privacy beyond k-anonymity. In *ICDE*, 2006.

[19] A. Meyerson and R. Williams. On the complexity of optimal k-anonymity. In *PODS*, 2004.

[20] S. Reiss. Practical data-swapping: The first steps. *ACM Transactions on Database Systems*, 9:20–37, 1984.

[21] S. Rizvi and J. R. Haritsa. Maintaining data privacy in association rule mining. In *VLDB*, 2002.

[22] P. Samarati. Protecting respondents' identities in microdata release. *IEEE Trans. on Knowledge and Data Engineering*, 13(6), 2001.

[23] L. Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. *Int'l Journal on Uncertainty, Fuzziness, and Knowledge-based Systems*, 10(5):571–588, 2002.

[24] K. Wang, P. Yu, and S. Chakraborty. Bottom-up generalization: A data mining solution to privacy protection. In *ICDM*, 2004.

[25] I. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques.* Morgan Kaufmann, San Francisco, 2nd edition, 2005.

[26] J. Zhang and V. Honavar. Learning decision tree classifiers from attribute value taxonomies and partially specified data. In *ICML*, 2003.