Information based data anonymization for classification utility

Jiuyong Li[1], Jixue Liu, Muzammil Baig

School of Computer & Information Science,University of South Australia

Raymond Chi-Wing Wong

Department of Computer Science & Engineering, Hong Kong University of Science and Technology

Abstract

Anonymization is a practical approach to protect privacy in data. The major objective of privacy preserving data publishing is to protect private information in data whereas data is still useful for some intended applications, such as building classification models. In this paper, we argue that data generalization in anonymization should be determined by the classification capability of data rather than the privacy requirement. We make use of mutual information for measuring classification capability for generalization, and propose two k-anonymity algorithms to produce anonymized tables for building accurate classification models. The algorithms generalize attributes to maximize the classification capability, and then suppress values by a privacy requirement k (IACk) or distributional constraints (IACc). Experimental results show that algorithm IACk supports more accurate classification models and is faster than a benchmark utility-aware data anonymization algorithm.

Keywords: Privacy, anonymization, k-anonymity, classification, mutual information, Kullback-Leibler divergence.

1. introduction

1.1. Data privacy and anonymization

Privacy preservation has become a major issue in many data mining applications. Various organisations, such as hospitals, medical administrations and insurance companies, have collected a large amount of data over years. However, these organisations are reluctant to publish the data because of privacy concern. It is necessary to ensure privacy protection when data is published.

Anonymization is a major technique for protecting privacy in data publishing. For example, k-anonymity [19] protects data privacy by ensuring that the probability for

[1] Mawson Lakes, SA 5095, Australia, tel: +61 8 83023898, fax: +61 8 83023381,
Email: FirstName.LastName@unisa.edu.au and raywong@cse.ust.hk

identifying an individual in a published data set is at most $1/k$. A common way to achieve k-anonymity is to generalize values within the person identifiable attributes in a table, called *quasi-identifier*. For example, if the following information, “gender = male, age = 45, postcode = 5011”, is too specific in a data set, e.g. fewer than k men of age 45 live in the suburb of postcode 5011. These people are potentially identifiable. If the record is generalized as ”gender = male, age = 45 - 55, postcode = 5009 - 5012”, more than k people will have the same person identifiable information in the data and hence their privacy is better preserved. The higher the privacy protection requirement is, the more the generalization will be. The most generalized form of a record is the “*, *, *”. The replacement of values with “*”s is called *suppression*. This is equivalent to that nothing is published for identity-related attributes.

Much research work has been conducted to enhance the protection level of the k-anonymity model, such as, l-diversity [15], (α, k)-anonymity [23], and t-closeness [13]. These models impose further protective requirements on the published data. They block attribute inference channels from identity-related attributes to sensitive values in the data. These models provide strong privacy protection, but are not good for some data mining tasks, such as classification, because associations between some attributes and classes have been purposely hidden. Models preventing attribute inference, such as, l-diversity and t-closeness, impose an upper bound for classification accuracies [14].

The goal for publishing a data set is to make it useful rather than to lock it in owner’s safe case. Building classification models is a major utility. For example, the hospital data is released to public for modeling causes of diseases. Normally, it is not an obligation for a data owner to build models but it is an obligation for a data owner to keep data privacy when the data is released. In many circumstances, k anonymity provides sufficient protection. For example, every released medical record has been authorized by a patient, and there is no privacy concern in the data itself However k-anonymization is necessary for preventing the medical data set from being linked to other patient sensitive information such as DNA sequences.

Most research on data utility has focused on value precision. The purpose is to minimise value generalisations in an anonymized table. Criteria such as distortion [12], uncertainty [24], query accuracy [11], information loss [20], and information utility [25] capture this information directly. The smaller modifications are made to a data set, the better the anonymization is. To increase value precision of anonymized tables, many anonymization techniques, such as, multidimensional [10] and local recoding [12, 24] methods, have been proposed. These methods reduce uncertainty or distortions of the anonymized data.

1.2. Related work and motivations

When the anonymized data is for building classification models, the utility requirement is quite different. Generalization is not a problem for building many interpretable classification models, such as decision trees and naive Bayes models. However, domain consistency is important and many precision based anonymization methods are not applicable [12, 24]. For example, values generalized to overlapping intervals, such as (10 - 14), (11 - 12), and (13 - 17), are not good for classification model building. generalized values mixed with different levels of an attribute taxonomy hierarchy, such as, 11th, 12th and senior secondary (which is a generalization of 11th and 12th) of Education attribute, are not good for building classification models either. Other methods are required for data anonymization for classification utility.

# of	Quasi-identifier		Other attributes		
repeating rows	Gender	Age	Blood pressure	...	Problem
5	male	60	high	...	yes
5	male	70	high	...	yes
1	male	30	normal	...	yes
1	female	70	high	...	no
5	female	30	normal	...	no
5	female	40	normal	...	no

(a) Raw table

# of	Quasi-identifier		Other attributes		
repeating rows	Gender	Age	Blood pressure	...	Problem
5	male	*	high	...	yes
5	male	*	high	...	yes
1	male	*	normal	...	yes
1	female	*	high	...	no
5	female	*	normal	...	no
5	female	*	normal	...	no

(b) anonymized by the current classification-aware anonymization methods

# of	Quasi-identifier		Other attributes		
repeating rows	Gender	Age	Blood pressure	...	Problem
5	male	60-70	high	...	yes
5	male	60-70	high	...	yes
1	*	*	normal	...	yes
1	*	*	high	...	no
5	female	30-40	normal	...	no
5	female	30-40	normal	...	no

(c) anonymized by the proposed algorithm

Table 1: An illustration of various classification-aware anonymization methods (k=10)

Many previous studies aiming at classification utility have been done. Iyengar [6] has firstly proposed an optimization approach to minimise class impurity in data generalization. The optimization has been shown impractical for medium and large data sets. Wang et al. have proposed a bottom up anonymization method for classification utility [21], which only handles categorical values. An improved method, called TDS (top-down specialization method), from the same research group has been proposed [4]. TDS makes use of the single dimensional generalization approach. It is efficient and keeps good classification capability in the anonymized data. A further improvement of TDS is called TDR [5] (Top-Down Refinement). TDR improves functionalities of TDS greatly. It handles both categorical and numerical values with and without generalization taxonomy trees. It also handles data with multiple quasi-identifiers. Recently, Kisilevich el al. [7] have proposed a multi-dimensional suppression approach, called kACTUS, for classification-aware anonymization. kACTUS makes use of a decision tree, i.e. C4.5 [17], as a base for deciding multi-dimensional regions to be suppressed. The pioneering work in multi-dimensional generalization has been proposed by Lefevre et al. [10], called Mondrian. Mondrian has then been extended to InfoGain Mondrian for various utilities including classification [11]. InfoGain Mondrian has been shown to achieve better classification accuracy than the TDS, and is a benchmark algorithm for classification based anonymization. Other privacy classification work has been done on the publishing classification models without violating the k-anonymity constraint. Friedman et al. have proposed a method for building k-anonymous decision trees [3]. Sharkey et al. [18] have also proposed a method for publishing decision trees along with a pseudo data set generated by the tree model. The release of a model lacks great flexibilities to users in comparison to the release of data. Firstly, there are many different types of classification models. A data owner won't know which model that a user is interested in. Secondly, for the same type of models, many adjustable parameters will lead to different models. For example, some users are interested in the specificity and some are interested in the precision. Their required models are quite different. Therefore, in this paper, we consider data publishing instead of model publishing.

Let us look at three most recent and closely related methods in data publishing for classification utility: InfoGain Mondrian [11], TDR [5], and kACTUS [7]. Interestingly, they produce the same 10-anonymous table for Table 1(a) following very different pathes. InfoGain Mondrian is a multi-dimensional generalization method, and it partitions the data space into a number of disjoined (hyper) rectangular regions by attributes in the quasi-identifier. The smallest partitioned regions (not optimized because the optimal solution is intractable), each of which contains at least k data points, are used for attribute value generalization. In this example, attribute Gender is partitioned along male and female. Any partition in attribute Age will lead to a region which has data points fewer than 10. Therefore, the Age attribute is kept at the top level "*". TDR starts with a table with all values suppressed. TDR then tests attributes Gender and Age to find out which will lead to better trade off between information gain and anonymity loss. Attribute Gender wins and attribute Gender is refined, and as a result values of males and females are shown. Note that classes (problems) have been well separated by attribute Gender. Values in Age are kept suppressed because the release of values in Age does not improve classification performance. kACTUS firstly builds a decision tree on Gender and Age attributes. The decision tree contains one node with the test that "whether gender = male or not". Both outcome branches contain 11 data points each and hence they comply with 10-anonymity. Attribute Age has not been referenced in the decision tree and hence all values are suppressed.

A disadvantage of the anonymized table in Table 1(b) is that it suppresses too many values when the data set has satisfied the privacy requirement. It is true that no other anonymization method is able to improve the classification accuracy in this 10-anonymous table. However, when other attributes are taken into account, it will be useful to have relationships between attributes Age and Blood Pressure. In other large data sets, such relationships potentially help classification. Normally the quasi-identifier is only a part of all the attributes, and we should not assume that a classification model is built on the quasi-identifier only. The combination of the quasi-identifier and other attributes potentially results in a more accurate classification model than that from the quasi-identifier alone. Furthermore, the assessment of classification capability depends on a criterion employed. For example, a criterion that is good for decision trees may not be good for logistic regression. We are unable to have a criterion that is good for all classification models, but it will be beneficial to maintain as many data values as possible. In contrast, both TDR and kACTUS have been designed with an implicit principle that only releases a small number of values to support a good classification model while the anonymity requirement is satisfied. As a result, their anonymized data is potentially not good for all classification models.

InfoGain Mondrian [11] utilises a principle that makes the minimal changes (generalizations) to a data set to achieve k-anonymity with the consideration of classification utility in the generalization process. The selection of an attribute to partition a data set aims to minimise weighted entropy to the classification advantage. This principle is very reasonable because it leaves as much information as possible in the anonymized data set. Such information will be helpful for building various classification models with different classification methods. However, it does not produce a better result in this data set in comparison to the two previous methods using an opposite principle. The reason lies in the generalization process. Region (male, 60-70) contains 10 data points, and region (male, 30-40) contains only 1 data point. It is better to suppress one data point in region (Male, 30-40) than to generalize Age 30-40 and 60-70 to 30-70 (or * in this case). A major benefit of the suppression is that the Age attribute is relatively complete and it has potential for other classification utilities, such as building classification models by using attributes Age and Blood Pressure.

We have seen the benefit of combining generalization with suppression. However, we should not extend InfoGain Mondrian [11] to produce classification-aware anonymized data sets by incorporating suppression. Two disadvantages of multi-dimensional generalization limit its applications in classification. One is the inconsistent domain and the other is possible overfitting to the quasi-identifier. For example, region (male, 21-40) and (female, 11-30) are legitimate partitions in the multi-dimensional generalization. Such partition leads to overlapping intervals 11-30 and 21-40 in the Age attribute and this causes problems for many classification methods. Let us assume that given male, it is good for classification to partition Age into 21-40 and 41-70. However, such a partition might not be good for another attribute, say Blood pressure, which is not in the quasi-identifier. When multiple attributes are conjunctively generalized within the quasi-identifier for classification, this reduces the chance for an attribute in the quasi-identifier to join another attribute for classification. Multi-dimensional generalization overfits anonymized data to the quasi-identifier, and the overfitting potentially damages the overall classification capability of anonymized data sets.

Based on the above discussions, we propose a new anonymization method for classification utility, which combines global generalization and local suppression. Levels of generalization are determined by the data distribution instead of the privacy requirement to better preserve data classification capability. Suppression is then used to remove detailed information locally to achieve data anonymization. The information loss in the suppression can be measured by Kullback-Leibler divergence and mutual information change. These measures give an objective assessment on quality of data after the suppression. The proposed algorithm has been shown to support more accu-

Quasi-identifier			Other	
Gender	Age	Postcode	attributes	Problem
male	35	4350	. . .	yes
male	40	4351	. . .	no
male	45	4350	. . .	no
female	43	4352	. . .	yes
female	62	4353	. . .	yes
female	68	4352	. . .	no

(a) Raw table

Quasi-identifier			Other	
Gender	Age	Postcode	attributes	Problem
male	[31, 40]	435*	. . .	yes
male	[31, 40]	435*	. . .	no
*	[41, 50]	435*	. . .	no
*	[41, 50]	435*	. . .	yes
female	[61, 70]	435*	. . .	yes
female	[61, 70]	435*	. . .	no

(b) 2-anonymous table

Table 2: A raw table and a 2-anonymous table

rate models, and be faster than the benchmark utility-aware anonymization algorithm, InfoGain Mondrian.

2. Problem description

2.1. Problem definition

The problem to be resolved in this paper is given as the following. Given a data set, generalization hierarchies for attributes in the quasi-identifier, and a k (or utility constraints), how to produce an anonymized data set which is good for building classification models?

2.2. Revisiting basic concepts of k-anonymization

The objective of k-anonymization is to make every tuple in identity-related attributes of a published table identical to at least $(k - 1)$ other tuples. Identity-related attributes are those which potentially identify individuals in a table. For example, all tuples in columns $\{$Gender, Age, Postcode$\}$ of Table 2 are unique, and hence problems of these individuals may be revealed if the table is published. To preserve their privacy, we may generalize Gender and Postcode attribute values such that each tuple in attribute set $\{$Gender, Age, Postcode$\}$ has at least two occurrences. A view after this generalization is given in Table 2(b).

Since various countries use different postcode schemes, in this paper, we adopt a simplified postcode scheme, where its hierarchy $\{$4201, 420*, 42**, 4***, *$\}$ corresponds to $\{$suburb, city, region, state, unknown$\}$, respectively. A tuple for an attribute set in a record is an ordered list of values corresponding to the attribute set in the record.

Definition 1 (Quasi-identifier) *A* quasi-identifier *(QID) is a set of attributes in a table that potentially identify individuals in the table.*

For example, attribute set {Gender, Age, Postcode} in Table 2(a) is a quasi-identifier. Table 2(a) potentially reveals private information of patients. Normally, a quasi-identifier is specified by domain experts.

Definition 2 (Equivalence class) *An* equivalence class *of a table with respect to an attribute set is the set of all tuples in the table containing identical values for the attribute set.*

For example, tuples 1 and 2 in Table 2(b) form an equivalence class with respect to attribute set {Gender, Age, Postcode}. Their corresponding values are identical.

Definition 3 (k-anonymity property) *A table is k-*anonymous *with respect to a quasi-identifier if the size of every equivalence class with respect to the attribute set is k or more.*

k-anonymity requires that every tuple occurrence for a given quasi-identifier has a frequency of at least k. For example, Table 2(a) does not satisfy 2-anonymity property since all tuples are unique.

Definition 4 (k-anonymization) k-anonymization *is a process to modify a table to another table that satisfies the k-anonymity property with respect to the quasi-identifier.*

For example, Table 2(b) is a 2-anonymous view of Table 2(a) since the size of all equivalence classes with respect to the quasi-identifier {Gender, Age, Postcode} is at least 2.

Generalization is a common way for anonymization. Generalization maps a value to a range or a coarsened value, for example, age of 35 is generalized to 31 - 40 and nationality Chinese is generalized to Asian. In the process of generalization, the value precision of the original values is lost partially. When values are generalized to the highest level *, this generalization is called suppression. Suppression is a generalization from any level to the highest level. Suppression removes a value from a table and replace it by *. Suppression looks rough but is very useful in practice. It hides information effectively without affecting values from irrelevant tuples. Suppressed values are equivalent to missing values which most classification methods can handle.

2.3. Why k-anonymization?

Some readers may wonder why we still study k-anonymization given so many new enhanced privacy protective models, such as, l-diversity [15], $(\alpha, k)-$anonymity [23] and t-closeness [13], have been presented. We believe that k-anonymization has its unique application potential that could not be replaced by other enhanced models. Firstly, these enhanced models require more generalizations than the k-anonymity model and reduces the utility (especially model building utility) greatly. Secondly, the k-anonymity model is mainly designed for preventing linking attack. The risk of linking attack is much higher than the risk of other attacks, such as homogeneity and background knowledge attacks [15]. In many cases, we only need to prevent linking attack. For example, given a medical data set that itself does not contain sensitive information (Date of Birth, Gender, Postcode, Diagnosis). The publication of diagnosis results is consented by patients for research purpose. Another DNA sequence data set is very sensitive, (DNAsequence). A DNA sequence contains gene signatures of many diseases, and so some diagnoses can be used to link to DNA sequences [16]. Therefore, an individual in a medical data is potentially linked to his/her DNA sequence. The DNA sequence contains more information than the gene signature of the disease that is consented by a patient for disclosure. For example, it contains family vulnerability of some diseases which should be private for the whole family. Therefore, the medical data set should be k-anonymized to prevent such linking attacks.

2.4. Utility of anonymous tables

Utility is a subjective criterion. It is difficult to give a general fits-all definition for utility. The closeness of values in an anonymized data set to values in the original data set has been a criterion, for example, distortions [12], uncertainty [24] and query accuracy. Other criteria [11], such as discernability metric [1] and normalized average equivalence class size [10] are indirect measurement of value precision. A major objective of anonymization is to enable building quality models on the anonymized data. An anonymized data set does not have to be close to the original data set at the value level, but a model built on the anonymized data should be as good as a model built on the original data.

Some may argue that we may release a model instead of a table. Firstly, there are many model building methods and we do not know which model a user is interested in. Secondly, for a model building method, there are many adjustable parameters. The parameters are fully dependent on the applications. One set of parameters does not suit all. A released model does not suit all users.

When the utility is for model building, value consistency is very important. For example, values from mixed attribute domains, such as values in Age attribute including 23, 10-15, 20- 30, 30-60, are not good for most classification tools. Most model building tools require attribute values from one domain. Values in overlapping intervals, such as 11 - 20 and 15 - 25, make building classification models even more difficult. A global recoding anonymization such as in [1, 4, 9] can produce an anonymized data with consistent attribute domain, but in some cases they may over generalize an attribute. For example, if two values out of 1000 values of an attribute need generalization to achieve k-anonymity. A global recoding method will generalize 1000 values and this causes over generalization. A better strategy is to suppress the two values. In this paper we employ global generalization and local suppression to keep attribute consistency and to avoid over generalization.

3. Determine the best generalization level

3.1. For single attributes

Let quasi-identifier of a data set be $Q = \{A_1, A_2, \ldots, A_m\}$, Attributes in the quasi-identifier are association with a set of attribute taxonomy hierarchies $\{T_1, T_2, \ldots, T_m\}$,

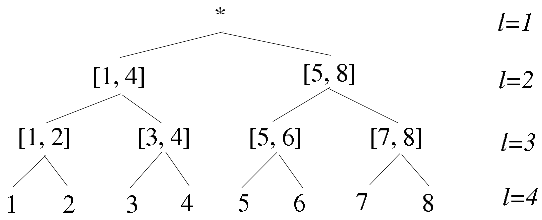


Figure 1: An examples of attribute generalization taxonomy and its height label.

which define the way for generalizing values. Let a data set $D = \{Q, O, C\}$, where O is a set of other attributes that are irrelevant to anonymization and C is a set of class labels $(c_1, c_2, \ldots, c_p)$.

Let T_i be an attribute taxonomy hierarchy. Level 1 is the most general value, representing any or all. We denote any or all as '*'. Level h is the most specific level with all values in the original data set. Levels in between are intermediate levels representing varying value specificities. For example, an attribute generalization hierarchy is given in Figure 1.

The question is which generalization level is best for classification. Mutual information has been frequently used for such a purpose.

The uncertainty associating with the set of class labels is described as the following.

$$H(C) = -\sum_{k=1}^{p} \text{freq}\,(c_k) \times \log_2 \text{freq}(c_k)$$

In classification term, entropy $H(C)$ indicates the classification uncertainty without using other attribute information. When we make use of an attribute to do the classification, the uncertainty reduction for the classification is quantified by the mutual information.

$$I(A_i(l); C) = H(C) - H(C|A_i(l))$$

$H(C|A_i(l))$ is the conditional entropy of attribute C given attribute A_i at level l. Let $A_i(l) = \{a_{i1}, a_{i2}, \ldots, a_{in_i^l}\}$ where n_i^l is the number of distinct values of attribute i at level l. $H(C|A_i(l)) = -\sum_{j=1}^{n_i^l} \sum_{k=1}^{p} \text{freq}\,(c_k|a_{ij}) \times \log_2 \text{freq}(c_k|a_{ij})$.

The mutual entropy is interpreted as that the amount of uncertainty in class label C, minus the amount of uncertainty in C which remains after attribute $A_i(l)$ is known. The mutual information can be used to indicate the classification potential of an attribute. In our problem, we use it to measure classification capability of different generalization levels of an attribute.

Mutual information varies with the number of distinct values in attribute A_i. Normally, a large number of distinct values leads to large mutual information. For example, if every value is distinct in attribute A_i, $\log_2 \text{freq}(c_k|a_{ij}(l)) = 0$ and hence $I(A_i(l); C) = H(C)$ is maximized. However, such an attribute has little predictive power. Let us assume that A_i contains student IDs and C indicates whether a student learns computer science. A_i explains C very well in the existing data, but does not

A_1	$A_2 \ldots A_m$	Class
1	$\ldots$	y
2	$\ldots$	y
3	$\ldots$	y
4	$\ldots$	y
5	$\ldots$	n
6	$\ldots$	n
7	$\ldots$	y
8	$\ldots$	n

(a)level 4

A_1	$A_2 \ldots A_m$	Class
$[1,2]$	$\ldots$	y
$[1,2]$	$\ldots$	y
$[3,4]$	$\ldots$	y
$[3,4]$	$\ldots$	y
$[5,6]$	$\ldots$	n
$[5,6]$	$\ldots$	n
$[7,8]$	$\ldots$	y
$[7,8]$	$\ldots$	n

(b) level 3

A_1	$A_2 \ldots A_m$	Class
$[1,4]$	$\ldots$	y
$[1,4]$	$\ldots$	y
$[1,4]$	$\ldots$	y
$[1,4]$	$\ldots$	y
$[5,8]$	$\ldots$	n
$[5,8]$	$\ldots$	n
$[5,8]$	$\ldots$	y
$[5,8]$	$\ldots$	n

(c) level 2

A_1	$A_2 \ldots A_m$	Class
$*$	$\ldots$	y
$*$	$\ldots$	y
$*$	$\ldots$	y
$*$	$\ldots$	y
$*$	$\ldots$	n
$*$	$\ldots$	n
$*$	$\ldots$	y
$*$	$\ldots$	n

(d) level 1

Table 3: Data at different generalization levels

predict future data since IDs do not repeat for future students. It is unreliable to make a prediction from a small number of observations. In contrast, the Gender attribute has more predictive power than the ID attribute since male students are more likely to study computer science than female students and this trend has not changed for a few years. For one attribute, generalization will increase repeated values in an attribute, but over-generalization makes the attribute less useful for classification. There is a trade off between generalization and prediction capability.

The mutual information is biased towards attributes of many values. To avoid such bias, we normalize the mutual information by using the attribute entropy. I_N denotes normalized mutual information.

$$I_N(A_i(l);C) \quad = \quad \frac{I(A_i(l);C)}{H(A_i(l))}$$

where $H(A_i(l)) = -\sum_{j=1}^{n_i^l} \text{freq}(a_{ij}(l)) \log_2 \text{freq}(a_{ij}(l))$ is the attribute entropy at level l and indicates the uncertainty of the attribute at level l.

We compare the normalized mutual information of all possible generalization levels. The one with the highest normalized mutual information is the best for the classification.

Example 1 *Table 3 lists a data set in different generalization levels. The attribute taxonomy is based on Figure 1.*

The summary of their entropy, mutual information and normalized mutual information is listed below.

l	$H(C)$	$H(C\vert A_1(l))$	$I(A_1(l);C)$	$I_N(A_1(l);C)$
4	*0.97*	*0*	*0.97*	*0.32*
3	*0.97*	*0.25*	*0.72*	*0.36*
2	*0.97*	*0.41*	*0.56*	*0.56*
1	*0.97*	*0.97*	*0*	*0*

From the above table, we see that the classification capability has been maximized at level 2.

3.2. For multiple attributes

The above process is possible to be extended to multiple attributes. For example, when we consider attribute set $\{A_1, A_2\}$, domain A_{12} should contain all value pairs of A_1 and A_2. Consider the generalization levels of different domains. Domains of different attributes at different levels form a generalization lattice. It is possible to find an optimal node in the generalization lattice that maximize the normalized mutual information. However, we do not consider such extension for the classification reasons.

Firstly, the quasi-identifier is only a small part of data attributes. Users do not intend to use the quasi-identifier only to build models. Users are more interested in the classification using attributes from both the quasi-identifier and other attributes. For example, female with high blood pressure, where Blood Pressure is a non quasi-identifier attribute. If we optimize the normalized mutual information within the quasi-identifier over multiple attributes, a classification model may fit the quasi-identifier too much and ignores other attributes.

Secondly, when multiple attributes are considered to maximize the normalized mutual information, we take advantage of classification capability of multiple attributes. The more attributes are considered, the fewer records are associating with each combined attribute value. In classification terms, the risk for overfitting increases. Overfitting is what we try to avoid in classification.

4. Measuring information loss in suppression

Previous generalization is not enough to produce tables with adequate privacy protection. Suppression is necessary to produce tables with sufficient privacy protection. We do not use further generalization since the domain consistency is important for classification. Suppression produces missing values, which can be handled by most classification methods.

k is a parameter of privacy requirement. Normally, the larger the k is the better protection for privacy. k is usually determined by users as an input parameter for a program. However, a large k may distort the data distribution and make models on the data useless. We need other criteria for measuring distributional changes in suppression for k-anonymization.

The effect of suppression on a data set should be measured by the distribution change of the data set before and after suppression. Kullback-Leibler divergence [8] is an information criterion to measure the difference between two probability distributions.

Let $D' = (A'_1, A'_2, \ldots, A'_m, O, C)$ be a table being generalized as discussed in Section 3. Note that l has been determined and fixed in the generalization process. In this section, l is a constant and hence is omitted. To distinguish attribute domains here from those in Section 3, we add a prime symbol to A_i. $(A'_1, A'_2, \ldots, A'_m)$ form the quasi-identifier, O includes other attributes, and C contains class labels. Let table $D'' = (A''_1, A''_2, \ldots, A''_m, O, C)$ be a suppressed table of table of D'. We need the value distribution in A''_i to approximate the value distribution in A'_i to reduce the adverse effect of suppression. Let $A'_i = \{a'_{i1}, a'_{i2}, \ldots, a'_{in_i}\}$ where n_i is the number of distinct values in A'_i. Similarly, $A''_i = \{a''_{i1}, a''_{i2}, \ldots, a''_{in_i}\}$. A'_i and A''_i are at the same level of attribute generalization taxonomy hierarchy, but A''_i includes suppressed values. We assume that suppressed values are also included in A'_i with a zero frequency. So both A'_i and A''_i include the same set of domain values with different frequencies. Kullback-Leibler divergence between attributes A'_i and A''_i is defined as the following.

$$\begin{aligned} & D(A'_i||A''_i) \\ = & \sum_j^{n_i} \mathrm{freq}(a'_{ij}) \log_2 \frac{\mathrm{freq}(a'_{ij})}{\mathrm{freq}(a''_{ij})} \\ = & -\sum_j^{n_i} \mathrm{freq}(a'_{ij}) \log_2(\mathrm{freq}(a''_{ij})) + \sum_j^{n_i} \mathrm{freq}(a'_{ij}) \log_2(\mathrm{freq}(a'_{ij})) \end{aligned}$$

In information theory, Kullback-Leibler divergence characterises the extra message length in bits when samples are coded based on distribution A''_i instead of A'_i. The smaller Kullback-Leibler divergence, the closer distribution A''_i is to A'_i.

Since Kullback-Leibler divergences vary from attributes to attributes, we normalize them so that they are comparable between attributes. D_N stands for normalized Kullback-Leibler divergence.

$$D_N(A'_i||A''_i) = \frac{D(A'_i||A''_i)}{H(A'_i)}$$

where $H(A'_i) = \sum_j^{n_i} \mathrm{freq}(a'_{ij}) \log_2(\mathrm{freq}(a'_{ij}))$ is the entropy of attribute A'_i.

One objective of suppression is to keep $D_N(A'_i||A''_i) \leq \delta_D$. δ_D is the maximum allowed change of Kullback-Leibler divergence, and is set as a small positive number. This is to cap the maximum distributional change of suppression.

We also need to make sure that the joint distribution of an attribute and the class attribute will not be affected by suppression. In other words, we wish that the classification capability of an attribute will not be significantly affected by suppression. Both positive and negative changes of classification capability are unfavorable. Positive changes of classification capability will lead to misleading models, and negative changes of classification capability will result in low quality models. The normalized mutual information is a proper measure for such changes.

Another objective of suppression is to keep $|I_N(A''_i; C) - I_N(A'_i; C)| \leq \delta_G$, where

A'_1	A'_2	$A'_3 \ldots A'_m$	Class
$[1,4]$	M	$\ldots$	y
$[1,4]$	M	$\ldots$	y
$[1,4]$	M	$\ldots$	y
$[1,4]$	F	$\ldots$	y
$[5,8]$	F	$\ldots$	n
$[5,8]$	F	$\ldots$	n
$[5,8]$	F	$\ldots$	y
$[5,8]$	F	$\ldots$	n

(a)

A''_1	A''_2	$A''_3 \ldots A''_m$	Class
$[1,4]$	M	$\ldots$	y
$[1,4]$	M	$\ldots$	y
$[1,4]$	M	$\ldots$	y
$*$	$*$	$\ldots$	y
$[5,8]$	F	$\ldots$	n
$[5,8]$	F	$\ldots$	n
$[5,8]$	F	$\ldots$	y
$[5,8]$	F	$\ldots$	n

(b)

Table 4: (a) A generalized data set. (b) The data set with suppression.

δ_G is the maximum normalized mutual entropy change allowed in suppression. This is to control the maximum classification capability change.

In practice, it would be difficult to determine right δ_D and δ_G. However, we can estimate them in a practical way, using fractions of maximum values. The maximum mutual information is $H(C)$ and $H(C)/H(A'_i)$ is a very big change in normalized mutual information for attribute A'_i. The maximum normalized Kullback-Leibler divergence can be estimated by $D_N(A'_i||A^*_i)$ where A^*_i keeps one tuple of every value in A'_i.

Example 2 *Table 4(a) shows a generalized table, which does not satisfy* 2*-anonymity. Table 4(b) shows a new table satisfying 3-anonymity after suppressing one record. Attribute distributions in Table 4(b) are supposed to approximate attribute distributions in Table 4(a). We use* A'_1 *as an example to show how* $D_N(A'_i||A''_i)$ *and* $I_N(A''_i;C) - I_N(A'_i;C)$ *are computed.*

We use the following summary to compute $D(A'_1||A''_1)$*.*

values	freq(a'_{1j})	freq(a''_{1j})
$[1,4]$	$4/8 = 0.5$	$3/8 = 0.375$
$[5,8]$	$4/8 = 0.5$	$4/8 = 0.5$
$*$	0	$1/8 = 0.125$

Therefore, we have the following results.

$$\begin{aligned} D(A'_1||A''_1) &= 0.5\log_2(\frac{0.5}{0.375}) + 0.5\log_2(\frac{0.5}{0.5}) = 0.21 \\ D_N(A'_1||A''_1) &= \frac{D(A'_i||A''_i)}{H(A'_1)} = 0.21 \end{aligned}$$

We know that $I_N(A'_1;C) = 0.56$ *from Example 1. We use the following summary to compute* $I(A''_1;C)$

values	freq(a''_{1j})	freq$(y|a''_{1j})$	freq$(n|a''_{1j})$
$[1,4]$	$3/7 = 0.43$	$3/3 = 1$	*0*
$[5,8]$	$4/7 = 0.57$	$1/4 = 0.25$	*3/4 = 0.75*

Algorithm 1 Information based Anonymization for Classification given k (IACk)

Input: Data set D, taxonomy hierarchies $\{T_1, T_2, \ldots, T_m\}$ of quasi-identifier attributes $\{A_1, A_2, \ldots, A_m\}$, and k

Output: a k-anonymous table of D, the largest normalized mutual information change α and the largest normalized Kullback-Leibler divergence β.

```
for each attribute $i$ in the quasi-identifier do
    compute normalized mutual information for each hierarchical level $l$ in $T_i$
    find $l'$ that maximizes the normalized mutual information
    generalize attribute $A_i$ to level $l'$ as $A'_i$
end for
sort $D'$ (after generalization) by the quasi-identifier
compute equivalence classes $E$
suppress values in equivalence classes with size $< k$
let $\alpha$ be the largest normalized mutual information change resulted by suppression
let $\beta$ be the largest normalized Kullback-Leibler divergence resulted by suppression
return $D''$ (after suppression), $\alpha$ and $\beta$
```

Note that in the above table, we discard $*$ value. This is because that the classification of y or n based on $*$ does not make sense. We wish that a classification model without using suppressed values is as good as a model using all values. We have the following results for suppression in A''_1.

$$\begin{aligned}
I(A''_1; C) &= H(C) - H(C|A''_1) = 0.98 - 0.46 = 0.52 \\
I_N(A''_1; C) &= \frac{I(A''_1; C)}{H(A''_1)} = 0.52/0.98 = 0.53 \\
|I_N(A''_1; C) - I_N(A'_1; C)| &= |0.53 - 0.56| = 0.03
\end{aligned}$$

5. Algorithm

In this section, we present an information based anonymization algorithm and its variant. The algorithm takes an input parameter of privacy requirement k and outputs an anonymized data set and the largest normalized mutual information change and Kullback-Leibler divergence caused by suppression. The variant takes the maximum allowed normalized mutual information change and Kullback-Leibler divergence, δ_G and δ_D, as input parameters, and outputs an anonymized data set and the largest k that satisfies the constraints. Both make use of the same generalization process but determine the number of suppressions based on different criteria. Both algorithms have the same time complexity.

The pseudo-code of the proposed algorithm is listed in Algorithm 1. The algorithm takes two major steps to produce anonymous tables: generalization and suppression. The generalization step generalizes attributes in the quasi-identifier to the best levels for classification. The suppression step achieves k anonymity by suppressing values in the quasi-identifier of equivalence classes whose sizes are less than k.

Lines 1-5 implement the generalization step. For each attribute, the level of the highest normalized mutual information is computed. Then all values in the attribute are generalized to that level. If there is not a level with the highest normalized mutual information for an attribute, the attribute will not been generalized. When two or more levels correspond to the highest mutual information, the attribute values will be generalized to the level closer to the bottom (opposite to the root) of the generalization taxonomy.

Lines 6-8 form equivalence classes and then create k-anonymous table. Firstly, data records are sorted by a criterion so that tuples of an equivalence class are adjacent. This is faster than forming equivalence class by computing pair wise dissimilarity between tuples. Secondly, values in equivalence classes whose size is less than k are suppressed. Line 9-11 computes the largest normalized mutual information change and Kullback-Leibler divergence caused by the suppression.

The complexity of this algorithm is estimated as the following. Let m be the size of quasi-identifier, n be the number of tuples, and l be the average height of attribute taxonomy hierarchies. In the generalization step, the normalized mutual information of every level of the attribute hierarchy of every attribute is computed. Each computation needs going through all attribute values once. The cost is $m \times h \times n$. The cost for generalization is negligible. For forming equivalence classes, all tuples are sorted by a criterion and the cost is $n \log(n)$. The time for forming equivalence classes is negligible. The total cost is $O(m \times h \times n + n \log(n))$. Note that both m (the size of quasi-identifier) and h (the average height of attribute hierarchies) are small. The complexity of the algorithm is in the order of $n \log(n)$.

We then present a variant with input parameters of distributional constraints (IACc), the maximum thresholds of normalized Kullback-Leibler divergence and normalized mutual information change δ_D and δ_G. IACc does not take k as an input but outputs the maximum k that satisfies the user's constraints. IACc calls IACk with $k = 1$, and suppresses equivalence classes from the smallest size to the largest size. The suppression stops when either normalized Kullback-Leibler divergence or normalized mutual information difference is larger than the corresponding threshold. The k is association with the largest size of equivalence classes being suppressed.

The pseudo-code of the variant is listed in Algorithm 2. Line 1 calls IACk with $k = 1$. Line 2 sorts equivalence classes by their sizes. This makes it easy in the next steps to find equivalence classes of the smallest size to suppress. Lines 5-15 implement suppression. The suppression starts from the smallest equivalence classes. The number of suppressions is constrained by the thresholds of the maximum normalized Kullback-Leibler divergence and mutual information change. The suppression stops when either of the constraints is dissatisfied.

Now we estimate the time complexity of Algorithm 2. We use notations as before. The base complexity is $O(n \log(n))$ since the input data set has been pre-processed by IACk without suppression. Let n_E be the number of equivalence classes after the generalization by IACk. The sort of equivalence classes by sizes will take $O(n_E \log(n_E))$ time. Let p be the number of the smallest size equivalence classes being suppressed. The time used for suppression is $O(m \times p)$. Since $m \ll n$ and $p \ll n$, the time for sup-

Algorithm 2 A variant of IACk for given distributional constraints (IACc)

Input: Data set D, taxonomy hierarchies $\{T_1, T_2, \ldots, T_m\}$ of quasi-identifier attributes $\{A_1, A_2, \ldots, A_m\}$, and the maximum thresholds of normalized Kullback-Leibler divergence and normalized mutual information change δ_D and δ_G.

Output: a k-anonymous view of D, and the largest k satisfying the constraints.

```
call IACk with $k = 1$
sort equivalence classes by their sizes
let $\{A'_1, A'_2, \ldots, A'_m\}$ be the quasi-identifier of the current $D'$ (after generalization)
let $e_{\min} = 0$
**while** (TRUE) **do**
    let $E_{\min} = \{E_{\min 1}, E_{\min 2}, \ldots\}$ contain all equivalence classes of the smallest size in all un-suppressed equivalence classes
    let $e'_{\min} = e_{\min}$ and $e_{\min} = |E_{\min 1}|$
    suppress values in the quasi-identifier of $E_{\min}$
    let $\{A''_1, A''_2, \ldots, A''_m\}$ be the quasi-identifier of current $D''$ (after suppression)
    **for** each attribute $i$ in the quasi-identifier **do**
        **if** $D_N(A'_i||A''_i) > \delta_D$ OR $|I_N(A''_i; C) - I_N(A'_i; C)| > \delta_G$ **then**
            restore values in $E_{\min}$, let $e_{\min} = e'_{\min}$ and then break while loop
        **end if**
    **end for**
**end while**
let $k = |e_{\min}| + 1$
return $k$ and $D''$
```

pression is negligible. The complexity of Algorithm 2 is $O(n\log(n) + n_E\log(n_E))$. Because of $n_E < n$, the complexity of Algorithm 2 is also in the order of $n\log(n)$, the same as that of Algorithm 1.

6. A proof concept experiment

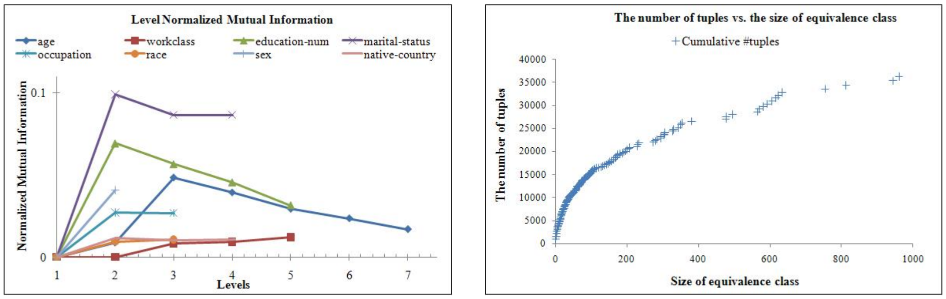
The objectives of the experiment are to demonstrate concepts discussed, and to compare the classification utility of anonymized data sets of the proposed method with a benchmark utility-aware anonymized algorithm, InfoGain Mondrian [11].

The adult data set from UCIrvine Machine Learning Repository [2] has become a benchmark data set for comparing k-anonymity methods. The data set has been used in most k-anonymity studies [4, 9, 10, 11, 24]. We eliminated records with unknown values. The resulting data set contains 45,222 tuples. We make use of 8 attributes as the quasi-identifier and one attribute as the class attribute, which are described in Table 5.

To benchmark our method with other existing methods, we compare our method with InfoGain Mondrian [11], which has been demonstrated to be better than other methods to anonymize data for various utilities including classification. Since the InfoGain Mondrian takes k as an input parameter, we use IACk for the comparison. We

	Attribute	Distinct Values	Quasi-identifier	Comments
1	Age	74	Yes	Taxonomy Tree of height 7
2	Work Class	7	No	Taxonomy Tree of height 5
3	Education	16	No	Taxonomy Tree of height 5
4	Martial Status	7	No	Taxonomy Tree of height 4
5	Occupation	14	No	Taxonomy Tree of height 3
6	Race	5	Yes	Taxonomy Tree of height 3
7	Sex	2	Yes	Taxonomy Tree of height 2
8	Native Country	41	Yes	Taxonomy Tree of height 4
9	Salary Class	2	No	Target class

Table 5: Description of Adult Data Set



(a) (b)

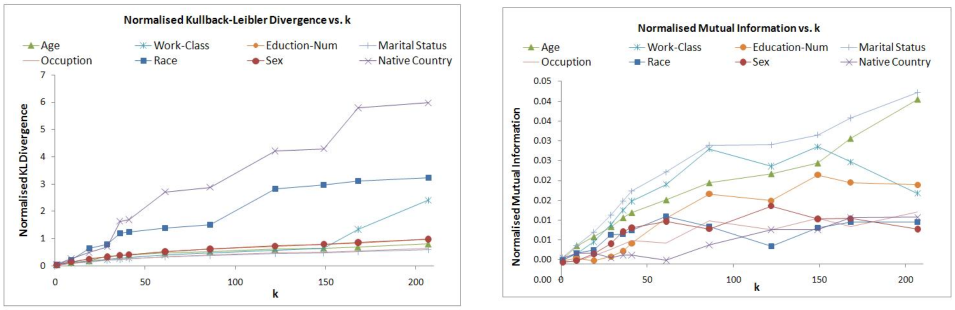
Figure 2: (a) Normalized mutual information at different levels of generalization hierarchies of eight attributes; and (b) Cumulative tuples over different sizes of equivalence classes after the generalization to the best levels for classification.

use the same attribute taxonomies for our algorithm and InfoGain Mondrian in the comparison.

Figure 2(a) shows normalized mutual information at different levels of attribute generalization hierarchies of the eight attributes. Attribute Age reaches the peak at level 3, Education at 2, Occupation at 2, Marital status at 2, Native Country at 2. These attributes are generalized to the levels corresponding to their peaks. Other three attributes do not have peaks of normalized mutual information, and hence have not been generalized. For example, Gender attribute has only two levels in its generalization hierarchy. It would cause too much information loss if values in Gender attribute were generalized.

Figure 2(b) lists the cumulative number of tuples in equivalence classes of different sizes after attributes are generalized to best levels for classification from the smallest to the largest. The generalized table is not 2-anonymous yet since around 2% tuples are unique. Some tuples are to be suppressed to achieve k-anonymity for $k \geq 2$. Normalized Kullback-Leibler divergences and normalized mutual information corresponding to various levels of suppression are listed in Figure 3.

Figure 3(a) shows normalized Kullback-Leibler divergence of different attributes



(a) (b)

Figure 3: (a) Normalized Kullback-Leibler divergences and (b) normalized mutual information corresponding to different suppression levels.

with different levels of suppression. In order to achieve k-anonymity, values in the quasi-identifier of all equivalence classes whose sizes are smaller than k are suppressed. Note that sizes of equivalence classes are not continuous integers and hence k is not the multiple of two, five or ten. The normalized Kullback-Leibler divergences of some attributes are significantly larger than others. The largest ones are from attributes Native Country and Race. Value distributions of these two attributes are skewed. For example, most records in Native Country column have the value of United-States, and remaining records have values of many other countries. The suppression of those low frequency values causes large distributional changes. However, those low frequency values are more likely in the equivalence classes of small size, and hence are more likely removed. This is a reason that normalized Kullback-Leibler divergences of the two attributes are largest. In contrast, the distribution of attribute Sex is quite even, hence its normalized Kullback-Leibler divergence is relatively small with the same number of value suppressions.

Figure 3(b) lists normalized mutual information of different attributes with different levels of suppression. The mutual information is association with the joint distribution of an attribute and the class attribute. Normalized mutual information changes are inconsistent with normalized Kullback-Leibler divergences. The largest changes are from attributes Marital Status and WorkClass. The differences between attributes are not as large as those of normalized Kullback-Leibler divergence. Since normalized Kullback-Leibler divergence and normalized mutual information depict different statistical properties of a data set, both are necessary for measuring the quality of data set after suppression.

We now assess the quality of models built on anonymized data sets by comparing the classification accuracy of various models on anonymized data sets with the accuracy of models on the original data set and the anonymized data sets by InfoGain Mondrian [11], which is a benchmark utility-aware anonymization algorithm.

Four classification model constructing methods are used in the experiment, and they are J48 decision tree (an implementation of C4.5 [17] in Weka [22]), logistic regression, LibSVM and naive Bayes. The implementation of the four classification methods are obtained from Weka APIs [22]. The classification accuracy is obtained by

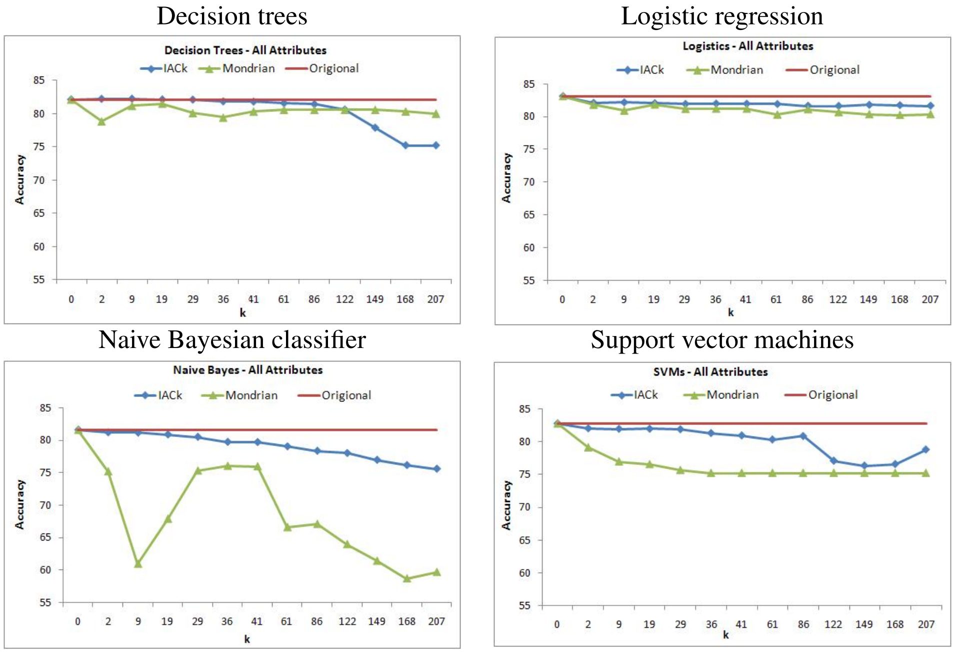


Figure 4: Accuracies: IACk versus InfoGain Mondrian bench marked by models on original data (eight attributes as the quasi-identifier). $k = 1$ is for the accuracy on the generalized data set.

10 cross-validation based on stratified sampling. A test data set is independent from its corresponding training data set. Generalization levels are determined by a training data sets solely and then mapped to test data set.

Figure 4 shows the accuracy of various models on k-anonymized data sets by IACk and InfoGain Mondrian in comparison against the accuracy of models on the original data set. Accuracy on the generalized data is very close to the accuracy on the original data in decision tree models and naive Bayes model and slightly lower in logistic regression and SVMs modules. This indicates that mutual information works well for preserving classification capability. On k-anonymized data sets, the accuracy decreases with the increase of k as expected. IACk is better than InfoGain Mondrian in the accuracy of models built on anonymized data sets with the exception of decision tree with $k > 122$. One reason for this exception is that InfoGain Mondrian makes use of a heuristic that is good for decision trees, and hence generalization has fitted decision trees. Another reason is that the current IACk can be further optimized. IACk suppresses values in all equivalence classes whose size is smaller than k. When k is large, this may cause much information loss. For example, instead of all suppression, these tuples can be further generalized by a k-anonymization algorithm, and then inconsistent values are subsequently suppressed. As a result, more values will be preserved and this will be helpful for classification.

Figure 5 shows running time of IACk and InfoGain Mondrian on the data sets with different k on the same PC computer. IACk is faster than Mandrian.

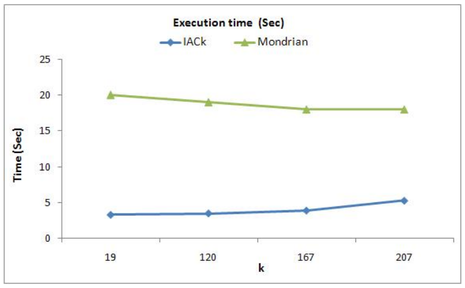


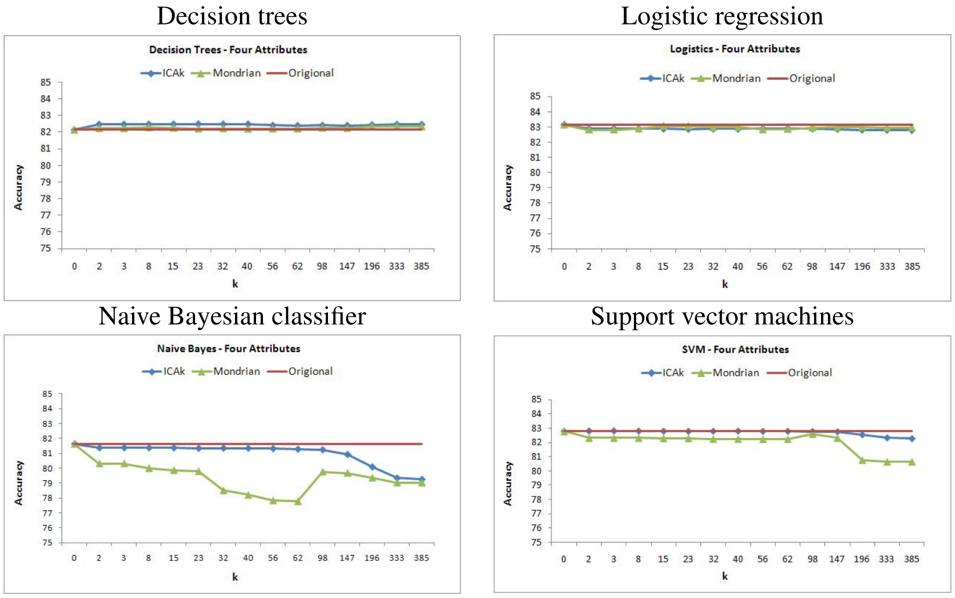
Figure 5: Running time of IACk and InfoGain Mondrian.

In the previous experiment, the set of other attributes is empty and classification models are built on the quasi-identifier only. In real world situations, the quasi-identifier is only a small portion of all attributes and there are many other attributes. Classification models are built on both quasi-identifier attributes and other attributes. To simulate real world situations, we do another experiment where four attributes, Age, Sex, Race and Native Country, are chosen as quasi-identifier attributes and remaining attributes are other attributes that do not need anonymization. Models will be built on both quasi-identifier attributes and other attributes.

Figure 6 shows the accuracy of various models on k-anonymized data sets of combined quasi-identifier attributes and other attributes by IACk and InfoGain Mondrian in comparison with the accuracy of models on the original data sets. Note that the range of y axis in Figure 6 is a third of that in Figure 4. Current accuracies are significantly higher than accuracies in the previous experiment where all eight attributes are used as quasi-identifier attributes. This is because that values in a half of attributes have not been generalized and suppressed. Accuracies of models built on data sets anonymized by IACk are very close to accuracies of models built on the orginal data set except for Naive Bayes with $k > 147$. IACk is consistently better than InfoGain Mondrian in all models. Results support our argument that it is not necessary to optimize the generalization of multiple attributes since quasi-identifier attributes are very likely to be used conjunctly with other attributes in model building.

7. Conclusion

In this paper, we have proposed two classification-aware data anonymization methods which combine global attribute generalization and local value suppression. The attribute generalization is determined by the data distribution, instead of by privacy requirement as other data anonymization methods do. Generalization levels are optimized based on the normalized mutual information for preserving classification capability. Value suppression is then determined by a privacy requirement k (IACk) or some data distributional constraints (IACc). Experiments show that the proposed method IACk anonymizes data that supports better classification models than the data anonymized by a benchmark utility-aware data anonymization method, and is faster. In the data simulating real world application scenario, the proposed method IACk is even


Decision trees

Logistic regression

Naive Bayesian classifier

Support vector machines

Figure 6: Accuracies: IACk versus InfoGain Mondrian bench marked by models on original data (four quasi-identifier attributes and four other attributes). $k = 1$ is for the accuracy on the generalized data set. Note that the range of y axis is one third of that in Figure 4

better and the generalized data supports models that are nearly as good as models from the original data.

Acknowledgement

This research has been supported by ARC Discovery DP0774450 and DP110103142.

References

[1] R. J. Bayardo and R. Agrawal. Data privacy through optimal k-anonymization. In *International Conference on Data Engineering (ICDE 05)*, pages 217–228. IEEE Computer Society, 2005.

[2] E. K. C. Blake and C. J. Merz. UCI repository of machine learning databases, http://www.ics.uci.edu/∼mlearn/MLRepository.html, 1998.

[3] A. Friedman, R. Wolff, and A. Schuster. Providing k-anonymity in data mining. *The VLDB Journal*, 14(4):789–804, July 2008.

[4] B. C. M. Fung, K. Wang, and P. S. Yu. Top-down specialization for information and privacy preservation. In *Internation Confeence on Data Engineeing (ICDE 05)*, pages 205–216, 2005.

[5] B. C. M. Fung, K. Wang, and P. S. Yu. Anonymizing classification data for privacy preservation. *IEEE Transactions on Knowledge and Data Engineering*, 19(5):711–725, 2007.

[6] V. S. Iyengar. Transforming data to satisfy privacy constraints. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 02)*, pages 279–288, 2002.

[7] S. Kisilevich, L. Rokach, Y. Elovici, and B. Shapira. Efficient multidimensional suppression for k-anonymity. *IEEE Transaction on Knowledge and Data Engineering*, 22(3):334–347, 2010.

[8] S. Kullback. *Information Theory and Statistics*. John Wiley, 1959.

[9] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. Incognito: efficient full-domain k-anonymity. In *ACM SIGMOD International Conference on Management of Data (SIGMOD 05)*, pages 49–60. ACM, 2005.

[10] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. Mondrian multidimensional k-anonymity. In *International Conference on Data Engineering (ICDE 06)*, page 25. IEEE Computer Society, 2006.

[11] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. Workload-aware anonymization techniques for large-scale datasets. *ACM Transactions on Database Systems*, 33(3):1–47, 2008.

[12] J. Li, R. C.-W. Wong, A. W.-C. Fu, and J. Pei. Anonymization by local recoding in data with attribute hierarchical taxonomies. *IEEE Transactions on Knowledge and Data Engineering*, 20(9):1181–1194, 2008.

[13] N. Li and T. Li. t-closeness: Privacy beyond k-anonymity and l-diversity. In *International Conference on Data Engineering (ICDE 07)*, pages 106–115. IEEE Computer Sciety, 2007.

[14] T. Li and N. Li. On the tradeoff between privacy and utility in data publishing. In *ACM SIGKDD international conference on Knowledge discovery and data mining (KDD 09)*, pages 517–526. ACM, 2009.

[15] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. L-diversity: Privacy beyond k-anonymity. *ACM Transactions Knowledge Discovery in Data*, 1(1), 2007.

[16] B. Malin and L. Sweeny. Inferring genotype from clinical phenotype through a knowledge based algorithm. *Proceedings of the Pacific Symposium on Biocomputing*, pages 41–52, 2002.

[17] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA, 1993.

[18] P. Sharkey, H. Tian, W. Zhang, and S. Xu. Privacy-preserving data mining through knowledge model sharing. In *ACM SIGKDD International Conference on Privacy, Security, and Trust in KDD (PinKDD 08)*, pages 97–115. Springer-Verlag, 2008.

[19] L. Sweeney. k-anonymity: a model for protecting privacy. *International journal on uncertainty, Fuzziness and knowldege based systems*, 10(5):557 – 570, 2002.

[20] H. Wang and R. Liu. Privacy-preserving publishing microdata with full functional dependencies. *Data & Knowledge Engineering*, 70(3):249 – 268, 2011.

[21] K. Wang, P. S. Yu, and S. Chakraborty. Bottom-up generalization: A data mining solution to privacy protection. In *IEEE Internation Conference on Data Mining (ICDM 04)*, pages 249–256, 2004.

[22] Weka development team. *http://www.cs.waikato.ac.nz/ml/weka/*. access on 10 Nov 2008.

[23] R. C.-W. Wong, J. Li, A. W.-C. Fu, and K. Wang. (alpha, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 06)*, pages 754–759, 2006.

[24] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, and A. W.-C. Fu. Utility-based anonymization using local recoding. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 06)*, pages 785–790. ACM Press, 2006.

[25] M. P. Zielinski and M. S. Olivier. On the use of economic price theory to find the optimum levels of privacy and information utility in non-perturbative microdata anonymisation. *Data & Knowledge Engineering*, 69(5):399 – 423, 2010.