k-anonymised reducts

Lior Rokach*

Department of Information Systems Engineering,
Ben-Gurion University of the Negev,
P.O. Box 653, Beer-Sheva, 84105, Israel
E-mail: liorrk@bgu.ac.il
*Corresponding author

Alon Schclar

School of Computer Science,
Academic College of Tel-Aviv Yafo,
P.O. Box 8401, Tel-Aviv, 61083, Israel
and
Deutsche Telekom Research Laboratories,
Ben-Gurion University of the Negev,
Beer-Sheva, 84105, Israel
E-mail: alonschc@mta.ac.il

Abstract: Privacy-preserving data mining aims to prevent the exposure of sensitive information as a result of mining algorithms. This is commonly achieved by data anonymisation. One way to anonymise data is by adherence to the *k*-anonymity concept which requires that the probability to identify an individual by linking databases does not exceed 1/*k*. In this paper, we propose an algorithm which utilises rough set theory to achieve *k*-anonymity. The basic idea is to partition the original dataset into several disjoint *reducts* such that each one of them adheres to *k*-anonymity. We show that it is easier to make each reduct comply with *k*-anonymity if it does not contain all quasi-identifier attributes. Moreover, our procedure ensures that even if the attacker attempts to rejoin the reducts, the *k*-anonymity is still preserved. Unlike other algorithms that achieve *k*-anonymity, the proposed method requires no prior knowledge of the domain hierarchy taxonomy.

Keywords: *k*-anonimity; rough set theory; reducts.

Biographical notes: Lior Rokach is a Senior Lecturer at the Department of Information Systems Engineering and the Software Engineering programme at Ben-Gurion University. He received his BSc, MSc and PhD in Industrial Engineering from Tel Aviv University. His main areas of interest are data mining, pattern recognition and information retrieval. He is the author of over 80 refereed papers in leading journals and conference proceedings. He is also the author of six books and over a dozen book chapters. He regularly participates in programme committees of conferences on data mining and recommender systems.

Alon Schclar is a Lecturer at the School of Computer Science at the Academic College of Tel-Aviv Yafo. He received his BSc, MSc and PhD in Computer Science from Tel Aviv University. His areas of interest include supervised and unsupervised machine learning, data mining and signal and image processing. He is the author of over 15 refereed papers in leading journals and conference proceedings.

1 Introduction

Many organisations maintain databases with personal information about individuals which are of interest to them. For example, banks and insurance companies collect data about their customers, hospitals store medical information about their patients and online stores maintain information about their shoppers. Mining these sources of information can prove to be highly beneficial for both commercial and research purposes. Thus, many organisations share their information and, in some cases, make it available to the public. Unfortunately, these sources of information are vulnerable to attacks that try to reveal private and sensitive information by linking sensitive values to the individual they belong to. Accordingly, many countries have privacy regulations to prevent private and sensitive information from being freely available. For example, the UK Data Protection Act (DPA) prohibits the use of data in case an original customer, account, secure entity or overall data trends can be identified from it. In order to still be able to share useful information while obeying privacy regulations, *privacy preserving data publishing* methods have been proposed. These methods apply an *anonymisation* process whose purpose is to reduce the probability that a sensitive value be connected to the individual it belongs to by combining any number of published databases.

Tables 1 and 2 demonstrate how sensitive information such as income can be revealed by linking two tables. Table 1 contains voting registration data which is publicly available for purchase while Table 2 contains anonymised income data. An attacker can deduce from these tables that Rachel earns 100 K by linking the tables according to the date of birth, zip code and gender columns.

Table 1 Commercially available voting registration data

Name	*Date of birth*	*Zip code*	*Gender*
Rachel	3-Jan-67	79,777	Female
Jack	25-Oct-55	12,123	Male
Bob	23-May-74	79,777	Male
Vic	19-Sep-53	12,990	Female
Vera	11-Jan-71	90,221	Female

Table 2 Anonymised income information

Gender	*Date of birth*	*Zip code*	*Income*
Male	15-Jun-80	59,555	50 K
Male	17-Nov-77	79,777	75 K
Female	3-Jan-67	79,777	100 K
Male	7-Feb-72	79,888	67 K
Female	4-Aug-78	89,555	92 K

Formally, given a database table, we distinguish between the following column types:

- *Identifiers (ID)* – attributes that uniquely identify a person, e.g., social security number. The column *name* in Table 1 is an identifier.
- *Quasi-identifiers (QID)* – attributes that do not individually identify a person, however their combination may do so by means of linking attacks. In Tables 1 and 2, the columns date of birth, zip code and gender constitute a quasi-identifier.
- *Sensitive* – attributes that contain private information such as income (Table 2), health condition, etc.
- *Non-sensitive* – attributes that do not fall into the above categories.

Several methods have been proposed for incorporating privacy-preserving requirements in various data mining tasks. In Shah and Zhong (2007), two semi-honest sharing models were extended to the malicious model for classification tasks. In Zhong (2007), algorithms for distributed mining of frequent itemsets were presented. Algorithms for mining sequential pattern on network traffic data were presented in Kim et al. (2008). In this paper, we focus on classification tasks.

In order to evaluate the anonymisation level of a published table, a formal model is required. These models are usually formed as a constraint on the distribution of values in the anonymised table. One of the most common anonymisation models is *k*-anonymity (Sweeney, 2002b). A dataset complies with the *k*-anonymity constraint if for each individual the data stored in the published dataset cannot be distinguished from at least $k - 1$ individuals whose data also appears in the dataset. More generally, each original individual record can only be reconstructed based on the published data with a probability that does not exceed $1/k$, given knowledge that is based on information available from external sources, e.g., published databases.

Several approaches were proposed to achieve *k*-anonymity: generalisation, suppression, permutation and perturbation. Generalisation is the most common technique used to process a given dataset so it complies with *k*-anonymity (Sweeney, 2002a). This method generalises attribute values and substitutes them with semantically consistent but less precise content. For example, the precise date of birth can be replaced by the year of birth which occurs in more records. Another example is to publish only part of the zip code, e.g., the first three digits. Generalisation makes the identification of a specific individual more difficult. One major drawback of existing generalisation techniques is that domain hierarchy trees (also known as *taxonomy* trees) are required for every quasi-identifier attribute of the datasets for which *k*-anonymity is to be achieved. These trees need to be generated manually before applying the generalisation process. Given a taxonomy tree, generalisation is achieved by replacing a value with values that are located higher in the tree. For example, in a taxonomy tree for professions, 'medical profession' may be the ancestor of 'doctor', 'nurse' and 'paramedic' and thus may be used to generalise them.

Suppression (Sweeney, 2002a) can be considered as a special case of generalisation in which a value is suppressed by being generalised to the most general value in the domain (the root of the taxonomy tree). Suppression can be applied to an entire row – in which case the row is eliminated from the table. Alternatively, suppression can be applied to an entire table column or to specific columns of individual rows (cells).

Permutation, which was proposed in Zhang et al. (2007), achieves anonymisation by permuting the content of each sensitive column. This destroys the connection between the identifying and sensitive attributes while preserving the privacy and aggregation properties of the table.

Perturbation replaces the original sensitive values with synthetic values whose statistical properties are similar to those of the original values. However, one can merely publish the statistical properties of the original data instead of the anonymised table (Domingo-Ferrer, 2008) since only these properties remain useful.

Although *k*-anonymity is the focus of this paper, other constraints have been proposed. The L-diversity model (Machanavajjhala et al., 2006) requires that each quasi-identifier group has at least *L* different values in the sensitive attribute. This takes into account cases where a sensitive value is the same for a set of records that belong to the same quasi-identifier group. Thus, although complying with the *k*-anonymity constraint, the sensitive value for these individuals can be easily deduced.

In the K-uncertainty model (Yao et al., 2005), anonymity is preserved if each identifier value is associated with at least *k* distinct values of a sensitive attribute. This model is commonly used in cases of multiple releases publishing (the content of the same tables is published at different time points), where the association between identifier and sensitive values is checked via the intersection of several releases. Finally, the (X, Y)-privacy model (Wang et al., 2006) generalises the above anonymity models by assuming there are two disjoint groups of attributes – X and Y – that describe individuals and sensitive properties, respectively. In order to satisfy (X, Y)-privacy, the published tables need to satisfy both: (X, Y)-anonymity and (X, Y)-linkability. The former requires that each value of X needs to be linked to at least *k* distinct values of Y while the latter requires that the inference probability to deduce a specific value of Y given a specific value of X, must to be less than a threshold *p*.

Granular computing is a computing paradigm of information processing that is gaining a growing amount of research in recent years. It deals with the processing of complex information entities, called 'information granules', which are commonly an integral part of data abstraction and knowledge inference processes. The granular computing paradigm has been applied to numerous tasks. Zhan (2010) addresses the application of granular computing for privacy-preserving data mining. Specifically, privacy-preserving association-rule mining, privacy-preserving *k*-nearest neighbour classification and privacy-preserving support vector machine classification were used to illustrate the paradigm of granular computing.

Most recently Zhou et al. (2009) presented two privacy preserving attribute reduction algorithms based on rough set theory (Pawlak, 1991): one uses vertically partitioned data and the other uses horizontally partitioned data.

In this paper, we use rough set theory (Pawlak, 1991) in order to achieve *k*-anonymity. The basic idea is to partition the original dataset into several disjoint reducts such that each one of them adheres to the *k*-anonymity constraint. We hypothesise that it is easier to make each reduct comply with *k*-anonymity if it does not contain all quasi identifier attributes. Moreover, our procedure ensures that even if the attacker attempts to rejoin the reducts, the *k*-anonymity is still preserved.

2 Preliminaries

We begin by presenting common definitions from rough set theory. Specifically, we are interested in the notion of reduct.

Definition 1: information table: An information table (also known as an *attribute-value system*) is a pair $T = (U, A)$ where U is the non-empty universe of primitive objects and A is a non-empty set of attributes. Let V_{a_i} be the domain values of attribute a_i. The attribute value of an object x is given by the function: a: $U \rightarrow V_a$ where $a(x)$ denotes the value of attribute a in object x.

Decision tables are a special type of information tables that are commonly found in the area of machine learning. In such tables, one or more attributes are labels that are the outcome of a classification process. These attributes are referred to as *decision* attributes and are denoted by D. The remaining attributes $C = A - D$ are referred to as the *condition* attributes since they form the basis for the decision rules that produce the values of the attributes in C. The attributes in D may be either binary or multi-valued.

Definition 2: The indiscernibility relation IND: Let $R \subset A$ be a subset of attributes. U can be partitioned into disjoint subsets of objects where the objects in each subset have the same attribute values and thus are indiscernible or indistinguishable. Formally, R induces an equivalence relation upon U which is referred to as the indiscernibility relation $IND(R)$ and is defined as:

$$IND(R) = \{(x, y) \in U \times U; \forall a \in R, a(x) = a(y)\}$$

The family of all equivalence classes of $IND(R)$ is denoted by $U / IND(R)$. Each element in $U / IND(R)$ is a set of indiscernible objects with respect to R. The equivalence classes $U / IND(C)$ and $U / IND(D)$ are called the condition and decision classes, respectively.

Let $X \subseteq U$ be a subset of objects and $R \subset A$ an attribute subset. Subsets which are of particular interest are ones that share the same value for a specific decision attribute and thus they may describe a certain class of objects in a learning problem. These subsets are also referred to as *concepts*.

We wish to provide an *exact* or *crisp* representation of X using the attributes in R. This is equivalent to expressing X as a union of the equivalence classes induced by R. This exact expression is impossible when X includes objects that have indistinguishable counterparts in their corresponding equivalence classes that are not in X. In such cases, rough set theory comes into play by providing means to approximate X. Specifically, X can be approximated by an R-lower approximation and an R-upper approximation which are defined as:

$$R_*(X) = \bigcup \{E \in U / IND(R) : E \subseteq X\} \text{ and}$$
$$R^*(X) = \bigcup \{E \in U / IND(R) : E \cap X \neq \phi\},$$

respectively. The R-lower approximation is composed of objects along with their indiscernible counterparts and thus these objects can be classified as part of X with certainty. On the other hand, the objects in the R-upper approximation can only be classified as possible members of X. Given this observation, we proceed to the next definitions:

Definition 3: POS: The positive region of decision classes $U / IND(D)$ with respect to the object subset X and condition attributes C is denoted by:

$$POS_C(D) = \bigcup R_*(X).$$

Definition 4: reduct: A reduct is the minimal set of attributes that preserves the positive region. Formally, a subset $R \subset C$ is said to be a D-reduct of C if $POS_C(D) = POS_R(D)$ and there is no $R' \subset R$ such that $POS_C(D) = POS_{R'}(D)$. The D-reduct of C is denoted by $RED_D(C)$.

The *CORE* is defined as the set of attributes that appear in all reducts, i.e., $CORE_D(C) = \cap RED_D(C)$.

The following definitions are required for the k-anonymity protocol and are adopted from Sweeney (2002b).

Definition 5: quasi-identifier: Given a universe U and a set of attributes A, $Q \subseteq A$ is said to be a quasi-identifier set, if $\exists e \in E$; such that $f_2(\pi_Q f_1(e)) = e$ where π is the projection operator and:

$$f_1 : U \to S,\ f_2 : S \to U'$$

such that $U \subseteq U'$, where E is the set of all individuals and S is the Cartesian product of all attributes domains which is given by:

$$S = \underset{\forall a_i \in A}{\times} V_a.$$

The above formulation defines a quasi-identifier as a set of attributes whose associated values may be linked in order to reveal the object that is described by the data.

We suggest an alternative definition of k-anonymity using the terminology of rough set theory:

Definition 6: k-anonymity: The k-anonymity level of an information table is the size of the smallest equivalence class in the Q-indiscernibility relation where Q is a quasi-identifier set.

Table 3 presents a portion of the *adult* dataset which can be obtained from the UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/support/Adult; Merz and Murphy, 1998). This dataset contains census data and has become a commonly used benchmark for k-anonymity. The *adult* dataset has six continuous attributes and eight categorical attributes. The class attribute which is used to label the instances is income level. This attribute has two possible values, <= 50 K or > 50 K. We use this dataset in order to illustrate Definitions 5 and 6.

The set of attributes Q = (*age*, *workclass*, *fnlwgt*, *edu*, *edu-nun*, *marital-status*, *occupation*, *relationship*, *race*, *sex*, *native-country*) in the *adult* dataset constitutes a quasi-identifier since the values of these attributes can be linked to identify an individual. As in previous studies, we assume that there is only one set of quasi-identifiers and that it is provided by the user. If we project Table 3 onto the attributes of Q, we get Table 4.

Note that the projection result does not comply with *two*-anonymity (k = 2). Inspection of records 12–14 shows that they comply with three-anonymity since they have the same values for the quasi-identifiers ($k = 3 > 2$). However, the remaining records are unique, and thus do not comply with two-anonymity ($k = 2 > 1$).

Table 3 Illustration of *adult* dataset

Age	*Workclass*	*Fnlwgt*	*Edu*	*Edu-num*	*Marital-status*	*Occupation*	*Relationship*	*Race*	*Sex*	*Capital-gain*	*Capital-loss*	*Hours-per-week*	*Native-country*	*Salary* > 50 *k*
39	Private	77,516	BA	13	Married	Executive	Not-in-family	White	M	2,174	0	40	USA	<= 50 K
39	Private	83,311	BA	13	Married	Executive	Husband	White	M	0	0	13	USA	<= 50 K
38	Private	215,646	BA	9	Divorced	Executive	Not-in-family	White	M	0	0	40	USA	<= 50 K
53	Private	234,721	BA	7	Married	Executive	Husband	Black	M	0	0	40	USA	<= 50 K
28	Private	338,409	BA	13	Married	Executive	Wife	Black	M	0	0	40	Cuba	<= 50 K
37	Private	284,582	BA	14	Married	Executive	Wife	White	M	0	0	40	Cuba	<= 50 K
49	Private	160,187	BA	5	Married	Executive	Not-in-family	Black	M	0	0	16	Cuba	<= 50 K
52	State-gov	209,642	BA	9	Married	Executive	Husband	White	M	0	0	45	Cuba	> 50 K
31	State-gov	45,781	BA	14	Married	Executive	Not-in-family	White	M	14,084	0	50	Cuba	> 50 K
42	State-gov	159,449	MA	13	Married	Executive	Husband	White	M	5,178	0	40	Cuba	> 50 K
37	State-gov	280,464	MA	10	Married	Executive	Husband	Black	F	0	0	80	Cuba	> 50 K
30	State-gov	141,297	MA	13	Married	Sales	Husband	Asian	F	0	0	40	Cuba	> 50 K
30	State-gov	141,297	MA	13	Married	Sales	Husband	Asian	F	0	2	60	Cuba	> 50 K
30	State-gov	141,297	MA	13	Married	Sales	Husband	Asian	F	0	1	80	Cuba	<= 50 K
34	State-gov	245,487	7th–8th	4	Married	Sales	Husband	Indian	F	0	0	45	Mexico	<= 50 K

Table 4 Projection of the *adult* dataset onto the set of attributes

Age	*Workclass*	*Fnlwgt*	*Edu*	*Edu-num*	*Marital-status*	*Occupation*	*Relationship*	*Race*	*Sex*	*Native-country*
39	Private	77,516	BA	13	Married	Executive	Not-in-family	White	M	USA
39	Private	83,311	BA	13	Married	Executive	Husband	White	M	USA
38	Private	215,646	BA	9	Divorced	Executive	Not-in-family	White	M	USA
53	Private	234,721	BA	7	Married	Executive	Husband	Black	M	USA
28	Private	338,409	BA	13	Married	Executive	Wife	Black	M	Cuba
37	Private	284,582	BA	14	Married	Executive	Wife	White	M	Cuba
49	Private	160,187	BA	5	Married	Executive	Not-in-family	Black	M	Cuba
52	State-gov	209,642	BA	9	Married	Executive	Husband	White	M	Cuba
31	State-gov	45,781	BA	14	Married	Executive	Not-in-family	White	M	Cuba
42	State-gov	159,449	MA	13	Married	Executive	Husband	White	M	Cuba
37	State-gov	280,464	MA	10	Married	Executive	Husband	Black	F	Cuba
30	State-gov	141,297	MA	13	Married	Sales	Husband	Asian	F	Cuba
30	State-gov	141,297	MA	13	Married	Sales	Husband	Asian	F	Cuba
30	State-gov	141,297	MA	13	Married	Sales	Husband	Asian	F	Cuba
34	State-gov	245,487	7th–8th	4	Married	Sales	Husband	Indian	F	Mexico

Note: $Q = \{age, workclass, fnlwgt, edu, edu\text{-}nun, marital\text{-}status, occupation, relationship, race, sex, native\text{-}country\}$.

Definition 7: STR: Given an attribute subset R, $STR(R)$ is the size of the smallest equivalence class in the R-indiscernibility relation.

Based on the above definitions we can derive the following lemmas. The proofs of the lemmas are straightforward and are thus left to the reader.

Lemma 1: Given a table T, a set of quasi identifiers Q, and a reduct R, the anonymity level of the projection of T upon the reduct R, which is denoted by $\pi_R T$, is equal to:

1 $|T|$ if $R \cap Q = \varnothing$

2 $STR(R)$ if $R \subseteq Q$

3 $\leq STR(R)$ otherwise (in most cases it is strictly smaller).

If we ignore the simplest case when R and Q are mutually exclusive, our goal is to find reducts with the highest STR values.

Lemma 2: The upper bound for the anonymity level of any reduct of a table T is $STR(CORE \cap Q)$.

3 Methods

In this section we describe two novel methods for construction of tables that comply with k-anonymity. Recall that the goal of this study is to find *anonymised reducts*, such that the predictive performance of a classifier trained on the anonymous dataset will be as close as possible to the performance of a classifier trained on the original dataset. Our first method utilises a greedy forward selection approach. Like in feature selection methods, the CORE can be used as the starting point since all the features in it cannot be removed. We successively add attributes until we reach a valid reduct. Figure 1 presents the proposed algorithm.

Figure 1 Forward selection heuristics for *k*-anonymity

Heuristics 1: Single *k*-Anonymized reduct

Input:

A – a set of attributes

Output:

R – an anonymized reduct

$R \leftarrow CORE(A)$

while (R is not a *reduct*)

Add a to R which maximizes the ratio:

$(Card(POS_R(D)) - Card(POS_{R+a}(D))) / (STR(R) - STR(R + a))$

end while

return R

Given a number of reducts, they can be combined into an ensemble classifier. This approach was proposed in Øhrn and Komorowski (1997). It is useful to generate many reducts using fast approximation heuristics and then construct a classifier by selecting a

subset of them (Wroblewski, 2001; Rokach and Maimon, 2005). In this paper, we utilise the ensemble methodology by combining a set of anonymised reducts, in order to obtain a better grasp of the original information table without violating privacy. The anonymity level of an ensemble of reducts can be bounded by the following lemma.

Lemma 3: The upper bound of the anonymity level of an ensemble of reducts $R_1,\ldots,R_m$ is bounded by $\min_{i=1,\ldots,l}(STR(R_i \cap Q))$.

The proof is straightforward because the conditions of Lemma 3 are a specific case of Proposition 1 presented in Matatov et al. (2010). It should be noted that even if the reducts are mutually exclusive, the anonymity level can still be smaller than the proposed bound because potential attackers can use the decision attributes in order to identify some of the individuals. Figure 2 presents a simple heuristic method for creating an ensemble of reducts by randomly selecting the input attributes. Note that it is required to calculate the combined anonymity level of the entire ensemble at the last step of the algorithm [see Matatov et al. (2010) for additional details].

Figure 2 Forward selection heuristics for *k*-anonymity

Heuristics 2: Random Ensemble *k*-Anonymized reducts

Input:

A – an attribute set

m – ensemble size

l – number of attributes

Output:

Ensemble, *Anonymity Level*

for *i* = 1 to *m*

*A** ← Random subset of *l* attributes of *A*

R ← Call Heuristics 1 (*A**)

Add *R* to *Ensemble*

end for

Calculate the anonymity level of *Ensemble*

return *Ensemble*

4 Experimental evaluation

In order to evaluate the proposed methods they were utilised for classification. Two benchmark datasets were used. Our experiments had the following goals:

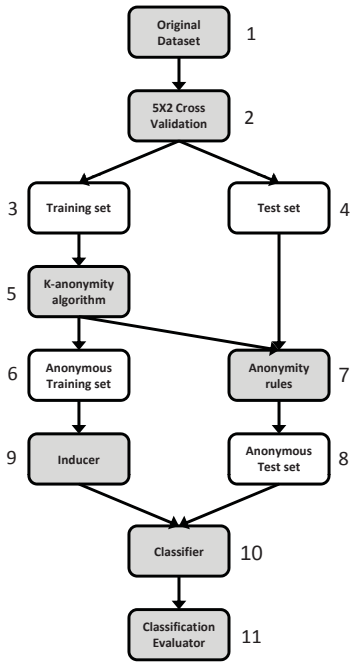
a compare between the classification accuracy obtained for the *k*-anonymised datasets with the accuracy obtained for the original datasets (without applying *k*-anonymity)

b compare the proposed methods with existing *k*-anonymity methods in terms of classification accuracy.

4.1 Experimental process

Figure 3 outlines the experimental process that was conducted for each benchmark dataset where un-shaded boxes represent datasets. The main purpose of this process is to estimate the generalised accuracy of the classifier, i.e., the probability that an instance was classified correctly. First, the dataset (box 1) was divided into a training set (box 3) and a test set (box 4) using five iterations of two-fold cross validation (box 2 – known as the 5 × 2 CV procedure) as proposed by Alpaydin (1999). In each iteration, the dataset was randomly partitioned into two equal-sized subsets S1 and S2 and the algorithm was evaluated twice: the first evaluation used S1 as the training set and S2 as the test set while the second iteration switched the roles of S1 and S2. We applied the *k*-anonymity method (box 5) to the training set and obtained an anonymous training set (box 6). Additionally, we obtained a set of anonymisation rules (box 7) that is used to transform the test set into an anonymous test set (box 8). An inducer is trained (box 9) on the anonymous training set to generate a classifier (box 10). Finally, the classifier is applied to the anonymous test set to estimate the performance of the algorithm (box 11). The same cross-validation folds are used for all the algorithms that were compared. Since the average accuracy is a random variable, the confidence interval was estimated by using the normal approximation of the binomial distribution. Moreover, we used the combined 5 × 2 CV F-test to accept or reject the hypothesis that the two methods have the same error rate with a 0.95 confidence level.

Figure 3 The experimental process



4.2 Datasets

The evaluation of most privacy preserving classification algorithms is solely based on the *adult* dataset which has become a commonly used benchmark for *k*-anonymity (Fung et al., 2005; Wang et al., 2004; Friedman et al., 2008). Fung et al. (2007) also used the *German credit* dataset. We evaluated the proposed algorithms using these two datasets which can be obtained from the UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/support/Adult; Merz and Murphy, 1998) – an extensive collection of datasets that are widely used by the machine learning community for the evaluation of learning algorithms.

We compared the results to the ones obtained by two of the current state-of-the-art algorithms DMPD (Matatov et al., 2010) and KACTUS (Kisilevich et al., 2010).

For the induction phase (box 9 in Figure 3), we used C4.5 (Quinlan, 1993) as the base induction algorithm since it is considered to be a state-of-the-art decision tree classification algorithm and thus it has been widely used for the evaluation of many other algorithms. All experiments were performed in the WEKA environment (Witten and Frank, 2005) where the C4.5 experiments were conducted using *J48* – the Java version of C4.5.

4.3 Results

In this section, we analyse the effect of the anonymity level – k – on the accuracy. Figure 4 shows the accuracy levels obtained by the different algorithms when applied to the *adult* dataset. As expected, the results indicate that there is a trade-off between the accuracy performance and the anonymity level. Namely, increasing the anonymity level decreases the accuracy. The results show that the proposed heuristics (H1 and H2) are superior for low values of k. When k is greater than 800, the results of the proposed algorithms are comparable to those obtained by the KACTUS and DMPD algorithms. Moreover, Heuristics 2 outperforms Heuristics 1 in most cases. However, Heuristics 2 also requires additional computational cost.

Figure 4 Comparing accuracy with the DMPD (Matatov et al., 2010) and KACTUS (Kisilevich et al., 2010) algorithms in the *adult* dataset (see online version for colours)

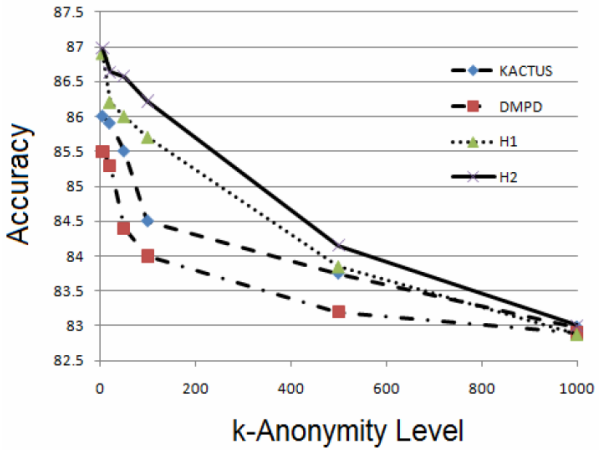


Table 5 presents the accuracy obtained by the different algorithms when applied to the *German credit* dataset. The asterisk '*' indicates that the degree of accuracy of Heuristics 2 was significantly different from that of the compared algorithm with a confidence level of 95%. Heuristics 2 significantly outperforms Heuristics 1 and KACTUS in four out of the seven cases and it significantly outperforms DMPD in only one case. Nevertheless, there is no case in which DMPD significantly outperforms Heuristics 2.

Table 5 Comparing accuracy with the DMPD (Matatov et al., 2010) and KACTUS (Kisilevich et al., 2010) algorithms in the *German credit* dataset

Method	*Anonymity level*			
	k = 2	*k = 5*	*k = 10*	
KACTUS	71.67 ± 1.28	*69.99 ± 1.69	*70.18 ± 1.48	
DMPD	72.79 ± 3.48	72.79 ± 3.48	72.53 ± 3.20	
H1	71.03 ± 2.43	*70.40 ± 1.42	30.17 ± 2.17	
H2	32.69 ± 3.3	32.69 ± 3.3	32.44 ± 3.3	
	k = 15	*k = 20*	*k = 30*	*k = 50*
KACTUS	*69.68 ± 2.56	*69.57 ± 1.42	*69.94 ± 1.24	70.52 ± 1.64
DMPD	71.70 ± 1.79	*70.72 ± 6.36	70.38 ± 2.70	71.68 ± 1.98
H1	*70.02 ± 1.17	*69.83 ± 1.65	*69.38 ± 1.9	70.01 ± 2.2
H2	71.44 ± 2.00	72.17 ± 2.70	71.38 ± 3.20	71.05 ± 2.9

4.4 Discussion

The advantages of the proposed method, as observed from the experimental study, can be summarised as following:

1 The proposed method is capable of applying *k*-anonymity to a given table with a minimal effect on classification accuracy.

2 When compared to the state-of-the-art *k*-anonymity methods, *k*-anonymised reducts can be used to induce classifiers which are of an equivalent or slightly higher degree of accuracy.

3 The proposed method, unlike other methods, does not require any prior knowledge. In many existing algorithms such as TDS (Fung et al., 2005), kADET (Friedman et al., 2008), TDR (Fung et al., 2007), GA-based anonymiser (Iyengar, 2002) and Incognito (Lefevre et al., 2005), the user is required to provide a taxonomy tree for categorical attributes. This makes them difficult to use. Additionally, it can become a source for disagreements among experts as described in Fung et al. (2005) and Friedman et al. (2008).

5 Conclusions

In this paper, we presented a new method for preserving the privacy in datasets using rough set theory. The proposed method requires no prior knowledge of the domain

hierarchy taxonomy and can be used with any inducer. The new method achieves higher predictive performance, compared to existing state-of-the-art methods.

The promising results in the previous section motivate the further investigation of the proposed methods. Specifically, other inducers such as support vector machines (SVM) and neural networks should be examined. Additionally, other anonymity measures should be used to evaluate the performance of the proposed methods, for example, *l*-diversity, which responds to different known attacks, such as homogeneous and background attacks. Finally, the proposed method should be extended to handle other data mining tasks such as clustering and association rules extraction.

References

Alpaydin, E. (1999) 'Combined 5 × 2 CV F test for comparing supervised classification learning classifiers', *Neural Computation*, Vol. 11, No. 8, pp.1975–1982.

Domingo-Ferrer, J. (2008) 'A survey of inference control methods for privacy-preserving data mining', in Aggarwal, C.C. and Yu, P.S. (Eds.): *Privacy-Preserving Data Mining*, *Advances in Database Systems*, Vol. 35, pp.53–80, Springer, New York, NY, USA.

Friedman, A., Schuster, A. and Wolff, R. (2008) 'Providing k-anonymity in data mining', *The VLDB Journal*, Vol. 14, No. 4, pp.789–804.

Fung, B.C.M., Wang, K. and Yu, P.S. (2005) 'Top-down specialization for information and privacy preservation', *Proc. of the 21st IEEE International Conference on Data Engineering (ICDE)*, Tokyo, Japan, pp.205–216.

Fung, B.C.M., Wang, K. and Yu, P.S. (2007) 'Anonymizing classification data for privacy preservation', *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, Vol. 19, No. 5, pp.711–725.

UCI Machine Learning Repository, http://archive.ics.uci.edu/ml/support/Adult (accessed on June 2010).

Iyengar, V.S. (2002) 'Transforming data to satisfy privacy constraints', *Proc. of the 8th ACM SIGKDD*, Edmonton, AB, Canada, pp.279–288.

Kim, S.W., Park, S., Won, J.I. and Kim, A.W. (2008) 'Privacy preserving data mining of sequential patterns for network traffic data', *Information Sciences*, Vol. 178, No. 3, pp.694–713.

Kisilevich, S., Rokach, L., Elovici, Y. and Shapira, B. (2010) 'Efficient multidimensional suppression for k-anonymity', *IEEE Trans. Knowl. Data Eng.*, Vol. 22, No. 3, pp.334–347.

Lefevre, K., Dewitt, D.J. and Ramakrishnan, R. (2005) 'Incognito: efficient full-domain k-anonymity', *Proc. of ACM SIGMOD*, Baltimore, ML, pp.49–60.

Machanavajjhala, A., Gehrke, J., Kiferl, D. and Venkitasubramaniam, M. (2006) 'L-diversity: privacy beyond k-anonymity', *Proc. of the 22nd IEEE International Conference on Data Engineering (ICDE)*, Atlanta, GA.

Matatov, N., Rokach, L. and Maimon, O. (2010) 'Privacy-preserving data mining: a feature set partitioning approach', *Information Sciences*, Vol. 180, No. 14, pp.2696–2720.

Merz, C.J. and Murphy, P.M. (1998) 'UCI repository of machine learning databases', Department of Information and Computer Science, University of California, Irvine, CA.

Øhrn, A. and Komorowski, J. (1997) 'Rosetta – a rough set toolkit for analysis of data', *Proc. of Third International Join Conference on Information Sciences (JCIS97)*, Durham, NC, USA, 1–5 March, pp.403–407.

Pawlak, Z. (1991) *Rough Sets: Theoretical Aspects of Reasoning about Data*, ISBN 0-7923-1472-7, Kluwer Academic Publishing, Dordrecht.

Quinlan, J.R. (1993) *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, CA, USA.

Rokach, L. and Maimon, O. (2005) 'Feature set decomposition for decision trees', *Intelligent Data Analysis*, Vol. 9, No. 2, pp.131–158.

Shah, D. and Zhong, S. (2007) 'Two methods for privacy preserving data mining with malicious participants', *Information Sciences*, Vol. 177, No. 23, pp.5468–5483.

Sweeney, L. (2002a) 'Achieving k-anonymity privacy protection using generalization and suppression', *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, Vol. 10, No. 5, pp.571–588.

Sweeney, L. (2002b) 'k-anonymity: a model for projecting privacy', *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, Vol. 10, No. 5, pp.557–570.

Wang, K., Fung, B. and Fu, C.M. (2006) 'Anonymizing sequential releases', *Proc. of the 12th ACM SIGKDD*, Philadelphia, PA.

Wang, K., Yu, P.S. and Chakraborty, S. (2004) 'Bottom-up generalization: a data mining solution to privacy protection', *Proceedings of the 4th IEEE International Conference on Data Mining*, November.

Witten, I.H. and Frank, E. (2005) *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed., Morgan Kaufmann, San Francisco.

Wroblewski, J. (2001) 'Ensembles of classifiers based on approximate reducts', *Fundamenta Informaticae*, Vol. 47, Nos. 3–4, pp.351–360.

Yao, C., Wang, X.S. and Jajodia, S. (2005) 'Checking for k-anonymity violation by views', *Proc. of the 31st Very Large Data Bases (VLDB)*, Trondheim, Norway, pp.910–921.

Zhan, J. (2010) 'Granular computing in privacy-preserving data mining', *International Journal of Granular Computing, Rough Sets and Intelligent Systems*, Vol. 1, No. 3, pp.272–288.

Zhang, Q., Koudas, N., Srivastava, D. and Yu, T. (2007) 'Aggregate query answering on anonymized tables', *Proc. of the 23rd IEEE International Conference on Data Engineering (ICDE)*.

Zhong, S. (2007) 'Privacy-preserving algorithms for distributed mining of frequent itemsets', *Information Sciences*, Vol. 177, No. 2, pp.490–503.

Zhou, Z., Huang, L. and Yun, Y. (2009) 'Privacy preserving attribute reduction based on rough set', *Second International Workshop on Knowledge Discovery and Data Mining, WKDD2009*, pp.202–206.