Journal of Ambient Intelligence and Humanized Computing
https://doi.org/10.1007/s12652-019-01489-7

ORIGINAL RESEARCH

Rounding based continuous data discretization for statistical disclosure control

Navoda Senavirathne[1] · **Vicenç Torra**[1,2]

Received: 25 January 2019 / Accepted: 9 September 2019

Abstract
"Rounding" can be understood as a way to coarsen continuous data. That is, low level and infrequent values are replaced by high-level and more frequent representative values. This concept is explored as a method for data privacy with techniques like rounding, microaggregation, and generalisation. This concept is explored as a method for data privacy in statistical disclosure control literature with perturbative techniques like rounding, microaggregation and non-perturbative methods like generalisation. Even though "rounding" is well known as a numerical data protection method, it has not been studied in depth or evaluated empirically to the best of our knowledge. This work is motivated by three objectives, (1) to study the alternative methods of obtaining the rounding values to represent a given continuous variable, (2) to empirically evaluate rounding as a data protection technique based on information loss (IL) and disclosure risk (DR), and (3) to analyse the impact of data rounding on machine learning based models. Here, in order to obtain the rounding values we consider discretization methods introduced in the unsupervised machine learning literature along with microaggregation and re-sampling based approaches. The results indicate that microaggregation based techniques are preferred over unsupervised discretization methods due to their fair trade-off between IL and DR.

Keywords Rounding for micro data · Unsupervised discretization · Micro data protection

1 Introduction

Rounding is based on the operating principle of data discretization or quantization, which maps a given continuous variable into a discrete set of values. This is achieved by replacing the values of a given set $X = \{x_1, \ldots, x_n\}$ into a much smaller set $R = \{r_1, \ldots, r_k\}$, where $k < n$ and each r_i is less specific compared to x_i values. Here, r_i is a *rounding point* and R indicates a collection of rounding points which defines a *rounding set*. Rounding values are defined so that minimum protection is guaranteed which describes the minimum criteria required for data protection. Privacy is achieved by replacing multiple x_i values by a single r_i. In rounding the biggest challenge is to generate the rounding set R, which reduces the expected distortion while avoiding the risk of "disclosure". A disclosure occurs when an adversary exploits a released dataset in order to obtain information about an individual otherwise not known to him. When each data record contains unique attribute values, it is straightforward for an adversary with relevant background information to attempt a disclosure, either by linking a given data record to a specific individual (identity disclosure) or by learning a sensitive attribute belongs to an individual (attribute disclosure).

Obtaining the rounding points can be explained with respect to scalar quantization (SQ). A given attribute/vector is partitioned into homogeneous groups, and a representative value is chosen for each partition as a rounding point. In information theory terminology a rounding point can be identified as a *code word* and the rounding set as the *code book*. In quantization, the objective is to generate a code book in a way that minimizes the distortion introduced by

This work is supported by Vetenskapsrådet project: "Disclosure risk and transparency in big data privacy" (VR 2016-03346, 2017-2020)

✉ Navoda Senavirathne
navoda.senavirathne@his.se

Vicenç Torra
vicenc.torra@mu.ie; vicenc.torra@his.se

[1] School of Informatics, University of Skövde, Skövde, Sweden

[2] Hamilton Institute, Maynooth University, Maynooth, Ireland

Published online: 25 September 2019

encoding. As explained earlier, this can be used as a data protection technique to minimize the disclosure risk.

As explained by Willenborg and De Waal (2012), the conventional way of defining the rounding points is to convert each x_i into a multiple of a given base value b as $\lfloor x_i/b \rceil * b$. For each rounding point r_i, the number of original data points (x_i) that can be mapped to it are known as the *set of attraction*. This is a half open interval indicated as $[r_i - \frac{b}{2}, r_i + \frac{b}{2})$. All the values in the original dataset within the above mentioned interval will be represented by r_i. It is necessary to decide on a suitable b value such that each set of attraction is sufficiently large enough (contains a least number of elements) in order to avoid disclosure. Further, this is explained as choosing the smallest b such that, $min\ \{F^x(0, r_1 + \frac{b}{2}), \ldots, F^x(r_n - \frac{b}{2}, x_{max}]\} \geq \alpha$ is achieved. F^x indicates the number of data points that fall within a given interval and α indicates the minimum size of the set of attraction. However, it is difficult to decide a base value b in a way that the distance between $|r_i - x_i|$ is minimized, and the required minimum protection is achieved.

In this work, we explore a few alternative approaches to generate the rounding sets. Here, we focus on the univariate case where the numerical attributes are masked one at a time. The SDC literature discusses masking methods with respect to both univariate and multivariate cases as they both have different benefits. In the univariate case, the overall distortion introduced by masking can be minimized as each variable is considered individually. However, this will intrinsically have a higher risk of disclosure compared to the multivariate case. Even though the multivariate masking can preserve the statistical relationships between the variables it would still introduce a high distortion. Moreover, grouping the variables for masking is not very straight forward.

- *Methods based on data discretization* Data points are partitioned into k non overlapping intervals, and for each interval rounding points are chosen based on a given aggregation method e.g., on mean, median, re-sampling mean or cluster centroids.

  Here, we consider methods such as equal width discretization (EWD), equal frequency discretization (EFD), re-sampling based discretization (RBD) and K-means clustering based discretization (KMD). EWD and EFD methods are general data pre-processing techniques often used in data analysis and machine learning domains. RBD is a new approach that we introduce in this paper.
- *Methods based on microaggregation* In this case, micro clusters are generated over a given set of data points ensuring a minimum of k items in each cluster. Then cluster centroids are selected as rounding points.

  Three univariate microaggregation methods are considered in this work. Maximum Distance to Average Vector (MDAV), Optimal Microaggregation (OMA) (Hansen and Mukherjee 2003), and the univariate implementation of Variable Distance to Average Vector (V-MDAV) method initially introduced for multivariate microaggregation.

Once the set of rounding values are identified, a given continuous variable can be quantized by either replacing each x_i using a deterministic approach or a stochastic approach. The methods mentioned above are explained in Sect. 3.

In the literature, rounding is explained as a statistical disclosure control method but we did not come across any experimental evaluations of the method. In this paper, we work to fill this gap. This will provide a clear understanding of rounding as a data masking tool and integrate rounding with discretization methods and microaggregation in a unified framework. We employ unsupervised and supervised discretization methods that have not been discussed before in the SDC literature in order to obtain the rounding points. Then we compare the results of different discretization methods discussed above in terms of their information loss (IL) and disclosure risk (DR).

This paper is structured as below. Related work is mentioned in Sect. 2 followed by in detail discussion of different rounding methods in Sect. 3. In Sect. 4, experimental setup and results are discussed. Impact of data rounding towards selected machine learning algorithms is discussed in Sect. 5, followed by discussion in Sect. 6 and conclusion in Sect. 7.

2 Related work

The SDC literature contains a plethora of micro-data protection methods also referred to as masking methods (Domingo-Ferrer 2008; Willenborg and De Waal 2012). Based on their operational principles, masking methods can be categorised into three categories as perturbative, non-perturbative and synthetic data generation (Torra 2017). Rounding is a perturbative data masking methods.

In this work, we are adopting data discretization methods to generate the rounding set. Data discretization is used widely in machine learning (ML) and knowledge discovery (Chmielewski and Grzymala-Busse 1996; Ibrahim and HacibeyoĞLU 2016; García et al. 2013; Dougherty et al. 1995; Ramírez-Gallego et al. 2016). Benefits of discretizing data as a pre-processing step include improvements in induction time, smaller sizes of induced trees/rules, enhanced predictive accuracy and the fact that most of the supervised learning algorithms require a discrete feature space (Pfahringer 1995; Yang and Webb 2002).

The operating principle of discretization can be used for numerical data masking in order to provide a privacy guarantee. Discretized data minimize the risk of disclosure, at the cost of information loss or decrease in analytical quality

of data. Therefore, when designing a discretization method minimizing the information loss becomes paramount. In data discretization, original attribute values are mapped into a more generic, less precise values. Microaggregation (MA) is a clustering based SDC technique that is designed for continuous data masking. It is considered as a quantization mechanism in Ramírez-Gallego et al. (2013) and Willenborg and De Waal (2012).

Lloyd (1982) and Max (1960) algorithms are the earliest and foremost attempts at creating optimal quantizers which is similar to k-means clustering. Here, the underlying principle of clustering is used for quantizer designing which is the same notion as discretization. The work presented by Rebollo-Monedero et al. (2013) introduced a modified version of Loyd-Max algorithm that shows how the concept of quantization can be used to achieve privacy by creating k-anonymous quantizers. Another algorithm for k-anonymous microaggregation is introduced by Rebollo-Monedero et al. (2011) which is based on the concept of distortion-optimized quantizers. The notion of k-anonymity is introduced by Sweeney (2002). A dataset is said to satisfy k-anonymity for $k > 1$, if each data record has at least $k - 1$ number of records sharing the same values for quasi identifiers. Data generalisation and local suppression are used to achieve k-anonymity while minimizing information loss. Having at least k records sharing the same values for quasi identifiers eliminate the risk of identity disclosure.

Data anonymization is studied as a vector quantization problem with respect to health data for minimizing individual or group re-identification from a released dataset (Miché et al. 2016). The paper proposes to use properties of vector quantization to anonymize a given dataset. Zhang (2011) discusses how fuzzy discretization can be used for protecting sensitive attribute values. Zhu et al. (2009) discuss how data discretization can be used for privacy preserving time series mining and the results indicate that data discretization causes a slight reduction of classification accuracy. However, they have not discussed the problem with respect to information loss or disclosure risk evaluation or comparatively analysing the different discretization techniques available. In our work we are, targeting to fill this gap with respect to continuous data.

3 Rounding

Rounding on micro data replaces the original continuous variable values (x_i) with selected rounding points r_i, so that the distance $|x_i - r_i|$ is minimized. All the rounding points obtained with respect to an attribute is known as a rounding set. Rounding can be either univariate or multivariate and deterministic or stochastic. In this work, we focus on deterministic-univariate rounding.

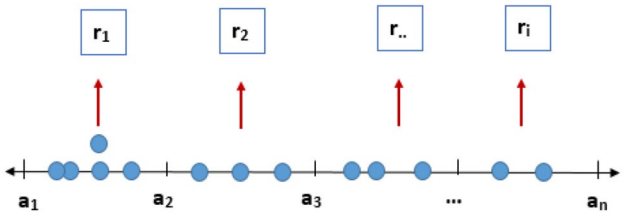


Fig. 1 Obtaining quantization regions and rounding points for a given variable x. Each quantization region is an interval and is represented by a rounding point $r_i \in \mathbb{R}$. A quantization region can be defined as $Q_i = [a_{i-1}, a_i)$

Rounding comprises three sub-processes (a) partitioning the dataset into quantization regions, (b) constructing the rounding set, and (c) encoding the original values with the nearest rounding points. Here, deriving the rounding set is the most critical step. Figure 1 depicts the process of obtaining the rounding set.

If rounding is considered as a process of quantization, there are two optimality conditions to be met (Lloyd 1982) .

- nearest neighbour condition: select the quantization point such that for all x_i arg min $d(x_i, q(x_i))$ where $q()$ is an optimal quantizer.
- centroid condition: each quantization point (rounding point) is the centroid of the quantization interval.

In the next section, we are exploring a few techniques that can be used to generate the rounding set keeping in mind the two optimality conditions of quantizer design.

3.1 Microaggregation for rounding

SDC methods are used for protecting micro-data so that the protected data can be released without a risk of disclosure. Here, we discuss microaggregation for deriving the rounding points.

This is a numerical data masking technique where the original values are divided into micro-clusters and then they are replaced by the cluster representatives. Parameter k defines the number of minimum data points required to form a micro-cluster. As the cluster centroid is used to replace the original values that fall into the particular cluster, the uniqueness of data records is concealed, thus preserving the privacy of the released data. The basic idea is to generate homogeneous clusters over the original data in a way distance between clusters are maximized so that the information loss can be minimized. Clusters are formed minimizing the sum of squared error (SSE) in groups.

Three variants of microaggregation (MA) are used to generate the rounding set as below.

- MA based on MDAV (maximum distance to average vector) algorithm with "fixed" micro-cluster sizes.
- MA based on V-MDAV (variable maximum distance to average vector) (Solanas et al. 2006) algorithm with "variable" micro-cluster sizes. This is initially introduced for multivariate microaggregation in Solanas et al. (2006). The micro-cluster sizes are between k and $2k-1$. The algorithm is explained in Algorithm 2.
- Optimal MA (OMA) (Hansen and Mukherjee 2003) based on the shortest path principle in graphs with "variable" micro-cluster sizes lies between k and $2k-1$. This derives the optimal solution for the k partition problem with minimal distortion.

Here, micro-cluster centroids are considered as rounding points (r_i) and each of these points has at least a k number of *points of attraction* from the original dataset.

There exists a wide variety of MA algorithms to generate the anonymized data other than the ones mentioned above (Chettri et al. 2012; Abidi et al. 2018).

3.2 Discretization for rounding

The process of discretization is used to map continuous data with a high cardinality into a finite set of values. Discretization of a given continuous variable includes sorting the data values, partitioning them into non overlapping intervals based on a given condition and selecting suitable values to represent the data in each interval. Discretization methods can be categorised into two, as *supervised* and *unsupervised* based on whether the class information is utilized in the data partitioning process. In this work, we are employing the unsupervised methods discussed below for obtaining the rounding set as class information is not always available for a dataset. More specifically, they are used to partition the data.

3.2.1 Equal width discretization (EWD)

The range of a sorted variable is divided into c non-overlapping partitions which are equally sized. Intervals are equidistant and c is a user defined parameter. The width of the intervals are obtained as $interval\ width = (max_x - min_x)/c$.

EWD is known to perform well on uniformly distributed data. However, with skewed distributions, this results in generating unbalanced intervals. Therefore, with EWD minimum protection cannot be ensured for rounding points obtained for each interval.

3.2.2 Equal frequency discretization (EFD)

Sorted values of a given variable are partitioned into c intervals in a way that each interval will roughly contain $\frac{n}{c}$ number of values.

The issue of unbalanced interval raised by EWD can be resolved by adopting EFD. But, EFD is known to have some other drawbacks such as duplicate data points being assigned into different intervals while very dissimilar values can be put together in order to form intervals with a given frequency (Bennasar et al. 2012). Also, when the above issue is resolved by grouping all the duplicate values within the same interval, it is not always possible to generate intervals with equal frequency (c) (Jiang et al. 2009). Therefore, the flexibility of deciding the number of intervals (anonymity constraint—c) is limited in this case.

3.2.3 K-means clustering (KMD)

Univariate k-means clustering is used on a given variable to create c clusters while minimizing the sum of distances between data points and cluster centroids. Cluster centroids are considered as rounding points, and the original data within each cluster is replaced by the centroid values.

3.2.4 Re-sampling based discretization (RBD)

Here, we explore how re-sampling can be used for discretization. First, we derive a re-sampled dataset from a given original dataset and then apply k-means clustering on the re-sampled dataset for discretization. First the sorted, original dataset ($sorted_{df}$) is partitioned into x quantiles (in this work we use 10 quantiles). Then for each quantile Q_i, m bootstrap samples ($bs_{1\ldots m}$) are extracted, where each bs_i is sized m. m is the size of the relevant quantile, that is $m = |Q_i|$. For each bootstrap sample, bs_i, its centroid (e.g., mean, median) is calculated. By repeating this process $x * m$ times, the re-sampled dataset is generated which is then discretized by using k-means clustering. Here, the discretization step is tested with other unsupervised discretization methods like EWD and EFD. Based on the results use of k-means clustering outperforms them, and thus we have used it for discretization here. The algorithm is explained in Algorithm 1.

```
Result: Discretized dataset - discretized_df
initialization;
resample_df ← list();
Partition the sorted dataset (sorted_df) into x quantiles (Q_i...x);
foreach q ∈ Q_i...x do
    m = SizeOf(q);
    #extract bootstrap samples from q;
    bs_1...m := bootstrap(q,m);
    for j in 1:m do
        #Calculate centroid values for each bs_i ;
        centroid := mean(bs_j);
        resample_df := append(resample_df,centroid);
    end
end
# k-means based discretization, c= number of clusters;
discretized_df := discretize(resample_df,c);
```

Algorithm 1: Re-sampling based discretization

3.3 Determining the quality of rounding methods

The results of each discretization method are evaluated based on three criteria, (a) information loss (b) disclosure risk and (c) the accuracy of machine learning models built on the rounded data.

3.3.1 Information loss

In simple terms, information loss measures quantify how deviated the masked data from its original version based on the statistical properties and the distance between data points. In this case, we use two methods to measure information loss (IL).

The *information loss metrics* (ILMetrics) is introduced by Domingo-Ferrer and Torra (2001). In this case, the IL is calculated by averaging the mean variance of $X - X', \bar{X} - \bar{X}', V - V', S - S'$ and the mean absolute error of $R - R'$ and finally multiplying it by 100. The symbols are explained as follows, X—the original data file and X'—the masked data file; V and R are covariance and correlation matrix of X; and S denotes the diagonal of V where $\bar{X}$ is the variable averages for X. Similarly, the other sets of symbols indicate the same properties of the masked data file.

Another IL measure is *IL1s* which computes the standardised distances between masked data and the original data scaled by the standard deviation. That is $IL1s = \frac{1}{mn} \sum_{j=1}^{m} \sum_{i=1}^{n} \frac{|x_{ij} - y_{ij}|}{\sqrt{2} S_j}$.

3.3.2 Disclosure risk

The main purpose of applying any SDC method is to minimize the risk of disclosure. Especially when personal data are handled, we want to make sure that the application of SDC methods mitigates the identity and attribute disclosures. Hence, an adversary with a substantial amount of auxiliary information would not be able to identify the data records with respect to a given data subject with certainty. In this case, two methods are employed to quantify the disclosure risk; (a) distance based record linkage (DDR) and (b) interval based disclosure (IDR).

In DDR, for each record in the masked data file, the distance to every record in the original data file is calculated. Then, the original data records that report the shortest distance to a given masked data record are considered as candidates for the linking process. A correct match is counted if the nearest record in the original data file is, in fact, the corresponding original record.

IDR defines an interval around the masked data (based on the standard deviation) and checks whether any original values fall within this interval (Templ 2017). Risk is measured as the number of times the above check is positive. However, the underlying concept of this is also the distance between masked data and original data. In order to distinguish this from DDR we refer to this as IDR in this work.

Result: Micrro-aggregated dataset - MA_{df}
initialization;
k, DataFile;
MA_{df} := DataFile ;
for $i \in 1 : ncol(DataFile)$ **do**
  centroid= ComputeCentroidOfDataSet($DataFile[i]$);
  while $ThereAreMoreThan\ [k-1]\ RecordsToAssign$ **do**
    e = SelectTheMostDistantRecordToCentroid ($DataFile[i]$,centroid);
    g_i = BuildGroupFromRecord(e,$DataFile[i]$,k);
    g_i = ExtendTheGroup(g_i,$DataFile[i]$,k);
  end
  $g_1 \dots g_s$ = AssignRemainingUnassignedRecords(DataFile[i], $g_1 \dots g_s$);
  $\text{MA}_{df}[i]$ = UpdateWithMicroaggregatedValues($g_1 \dots g_s$);
end

Algorithm 2: Univariate VMDAV based microaggregation

3.3.3 IL-DR score

As discussed above, IL is measured based on IL metrics and IL1s while disclosure risk is calculated based on IDR and DDR. These values obtained for each masked/discretized dataset is then used to derive a score as explained below (Domingo-Ferrer and Torra 2001).

$$\text{IL-DR}_{\text{score}} = (IL\ Metrics * 0.25 + IL1s * 0.25 + 0.25 * DDR + 0.25 * IDR)$$

IL-DR score gives equal weight to different IL (information loss) and DR (disclosure risk) measures we have obtained. Therefore, it is used to understand the trade-off between privacy (DR) and utility (IL). The lower the score, the better the particular discretization methods is.

3.3.4 Accuracy of ML models

This can also be considered as another way of measuring information loss. Here, we explore how rounding impacts the predictive accuracy of machine learning (ML) models. This is studied with respect to linear regression and decision trees. The prediction accuracy of the ML models trained on masked data is determined based on how accurately the models can predict the original data. Low prediction accuracy on original data indicates that the masked data used to train the models have a poor utility. For linear regression models mean squared error (MSE) and R^2 score is used to evaluate the utility of the models built on masked data along with information loss measures discussed above. With respect to decision trees, mainly classification accuracy and entropy is used for the analysis. Decision tree classifiers are built using the RPART[1] package available on R.

[1] https://cran.r-project.org/web/packages/rpart/rpart.pdf

4 Methodology and experimental results

As explained previously, we evaluate different techniques that can be used to obtain the rounding set which is based on microaggregation or unsupervised discretization techniques. These methods are then evaluated with respect to distortion (information loss) and anonymity (disclosure risk) while changing other properties such as data distribution and dataset size. Finally, the impact of data masking on ML learning based data modelling is evaluated using linear regression and decision tree algorithms.

4.1 Data

For the experiments two types of datasets are used; (a) synthetic datasets following theoretical distributions, i.e., exponential, uniform and normal distributions, and (b) Tarragona[2], Boston housing information[3], and Wine[4] classification datasets which are openly available. All synthetic datasets are 2×500 in dimension and generated using *rexp, runif* and *rnorm* functions available in R respectively to generate exponential, uniform, and normal distributions. Parameters for generating the synthetic datasets are as follows. Exponential distribution is generated with $\lambda = 0.08$, normal distribution is generated with ($\mu = 0$, $\sigma = 1$) and uniform distribution is generated with minimum and maximum values in the range of (0,1). Dimension of the other datasets are: Tarragona as 834×13, Boston housing information 506×14 and Wine classification as as 179×14.

4.2 Results

In this Section, we have reported the results from the experiments. First, we compare the information loss (IL), and disclosure risk (DR) values obtained for different synthetic datasets. Comparisons are made for microaggregation based methods and unsupervised discretization methods. MDAV, V-MDAV and OMA are compared together as they are microaggregation based discretization methods, whereas the unsupervised discretization methods such as EWD, EFD, KMD, and RBD are compared together. Here, each of the above mentioned methods is used to obtain the rounding set. This can also be explained as partitioning a dataset based on specified criteria. Then for each partition, the representative values are selected and then the original data are replaced by them. IL is the distortion caused by this encoding process while DR indicates whether we can directly identify a given masked data record (after rounding is applied) with respect to its original record with a certainty. This can also be explained as the number of successful record linkages between the original dataset and the rounded dataset.

4.3 Setting anonymity constraints

We have selected the anonymity constraint values (k and c respectively for microaggregation and unsupervised discretization) in a way that approximately the same number of data points are used to form the micro-clusters in microaggregation and the intervals/ clusters in unsupervised discretization to obtain the rounding points. The relationship between k and c can be explained as below. The k and c values are set like this in order to compare the overall result at the end.

$$k \approx \frac{\#\ of\ instances\ in\ attribute_j}{c}$$

$$c \approx \frac{\#\ of\ instances\ in\ attribute_j}{k}$$

The selected set of values for parameter k in microaggregation are { 167, 100, 50, 34, 25, 20 }. These are approximately equal to the selected parameter values c for unsupervised discretization. They are respectively { 3, 5, 10, 15, 20, 25 }. In synthetic datasets each attribute has 500 instances.

4.4 Microaggregation for rounding

In the case of microaggregation, the number of data points per micro-cluster (k) is directly proportional to IL, and it is inversely proportional to DR. The higher the number of values in micro-clusters, the quality of the selected rounding points will be low resulting in high IL. This behaviour is vice versa, when a fewer number of values are used to form the micro-clusters.

Here, three MA based methods are compared with respect to different synthetic data distributions and the results are shown in Tables 1, 2 and 3. As explained in Sect. 3.3.3 the lower the IL-DR scores the better the specific method is. With respect to exponentially distributed data VMDAV outperforms the other two microaggregation methods when evaluated based on the *IL-DR Score* (see column *IL-DR Score* on the above mentioned tables). When uniformly and normally distributed data are considered, OMA performs better than the other two MA methods. However, with respect to uniformly distributed data when the k value is low (i.e., k = 34, 25, 20) VMDAV performs better than OMA. With high k values, OMA performs better. With respect to normally distributed data when the k value is low, both OMA and MDAV perform equally. As we can see despite being the "optimal" method for microaggregation, OMA does not always produce the lowest *IL-DR Scores*. Compared to

[2] https://cran.r-project.org/web/packages/sdcMicro/index.html

[3] https://cran.r-project.org/web/packages/mlbench/index.html

[4] https://archive.ics.uci.edu/ml/datasets/wine

Table 1 Comparison of IL and DR measures obtained for the exponentially distributed synthetic dataset with varying anonymity constraints (k)

Rounding method	Anonymity constraint (k)	IL metrics	IL1s	IDR	DDR	IL-DR score
MDAV exponential	167	103.26	542.51	0.05	0.00	161.45
	100	43.06	179.47	0.24	0.03	55.70
	50	123.91	102.12	0.47	0.09	56.65
	34	52.49	98.40	0.63	0.16	37.92
	25	73.26	57.10	0.73	0.26	32.84
	20	81.86	44.48	0.82	0.31	31.87
OMA exponential	167	126.50	335.30	0.07	0.00	115.47
	100	43.06	179.47	0.24	0.03	55.70
	50	123.91	102.12	0.47	0.09	56.65
	34	75.18	74.33	0.64	0.17	37.58
	25	73.26	57.10	0.73	0.26	32.84
	20	84.40	44.36	0.81	0.29	32.47
VMDAV exponential	167	120.57	335.54	0.07	0	114.05
	100	78.66	201.09	0.15	0.01	69.98
	50	7.54	99.86	0.43	0.07	26.98
	34	10.65	79.99	0.65	0.16	22.86
	25	4.11	58.30	0.72	0.24	15.84
	20	3.30	42.45	0.75	0.30	11.70

Increasing anonymity constraint (k) value relates to low DR (high privacy) and high IL

Table 2 Comparison of IL and DR measures obtained for the uniformly distributed synthetic dataset with varying anonymity constraints (k)

Rounding method	Anonymity constraint (k)	IL metrics	IL1s	IDR	DDR	IL-DR score
MDAV uniform	167	284.75	354.33	0.04	0.00	159.78
	100	26.07	122.59	0.15	0.03	37.21
	50	9.34	62.70	0.28	0.09	18.10
	34	25.19	46.26	0.40	0.16	18.01
	25	13.15	31.96	0.57	0.28	11.49
	20	2.01	25.41	0.73	0.33	7.12
OMA uniform	167	52.66	321.22	0.04	0.00	93.48
	100	26.07	122.59	0.15	0.03	37.21
	50	9.34	62.70	0.28	0.09	18.10
	34	21.65	43.05	0.42	0.17	16.32
	25	22.49	30.98	0.59	0.27	13.58
	20	50.22	24.16	0.75	0.32	18.86
VMDAV uniform	167	138.60	321.43	0.05	0.00	115.02
	100	9.48	156.13	0.12	0.01	41.44
	50	11.02	70.86	0.25	0.08	20.55
	34	3.85	44.89	0.40	0.17	12.33
	25	3.33	32.23	0.56	0.26	9.10
	20	1.29	24.47	0.74	0.32	6.71

Increasing anonymity constraint (k) value relates to low DR (high privacy) and high IL

other methods, in most of the cases, OMA results in high disclosure risk (DR) ratios caused by the high utility of the rounded data. Eventually, this results in high *IL-DR Scores* which indicate a high privacy utility trade-off.

The same experiments are used on Tarragona dataset to measure IL and DR, as shown in Fig. 2. In this experiment, IL is also measured in terms of the sum of square error (SSE) and total sum of squares (SST). SSE indicates within the group homogeneity in micro-clusters. SST is the

Table 3 Comparison of IL and DR measures obtained for the normally distributed synthetic dataset with varying anonymity constraints (k)

Rounding method	Anonymity constraint (k)	IL metrics	IL1s	IDR	DDR	IL-DR score
MDAV normal	167	35.85	404.89	0.05	0.00	110.20
	100	44.73	169.16	0.12	0.02	53.51
	50	6.28	91.08	0.31	0.08	24.44
	34	8.44	75.21	0.41	0.15	21.05
	25	5.31	52.61	0.54	0.25	14.68
	20	3.24	42.73	0.68	0.33	11.75
OMA normal	167	35.61	377.93	0.03	0.00	103.39
	100	44.73	169.16	0.12	0.02	53.51
	50	6.28	91.08	0.31	0.08	24.44
	34	5.40	67.72	0.41	0.16	18.42
	25	5.73	51.49	0.57	0.26	14.51
	20	3.96	42.16	0.69	0.32	11.78
VMDAV normal	167	38.36	378.30	0.04	0.00	104.18
	100	30.91	201.79	0.09	0.01	58.20
	50	9.05	90.49	0.26	0.07	24.97
	34	21.13	71.15	0.39	0.16	23.21
	25	14.94	50.72	0.54	0.25	16.61
	20	6.45	43.08	0.67	0.31	12.63

Increasing anonymity constraint (k) value relates to low DR (high privacy) and high IL

summation of, between the groups' sum of squares (SSB) and SSE. Here the IL is calculated based on the following formula $IL = \frac{SSE}{SST} * 100$. The formulas are explained in detail in Chettri et al. (2012).

As shown in Fig. 2a, b both OMA and VMDAV results in low IL compared to MDAV. OMA reports the highest DR compared to the other two methods. When Consider the IL and DR trade-off VMDAV outperforms the other two methods. Most of the attributes in the Tarragona dataset are exponentially distributed. Based on the results obtained on the synthetic datasets, VMDAV is the most suitable approach for rounding when a dataset is exponentially distributed. The test results confirm this finding.

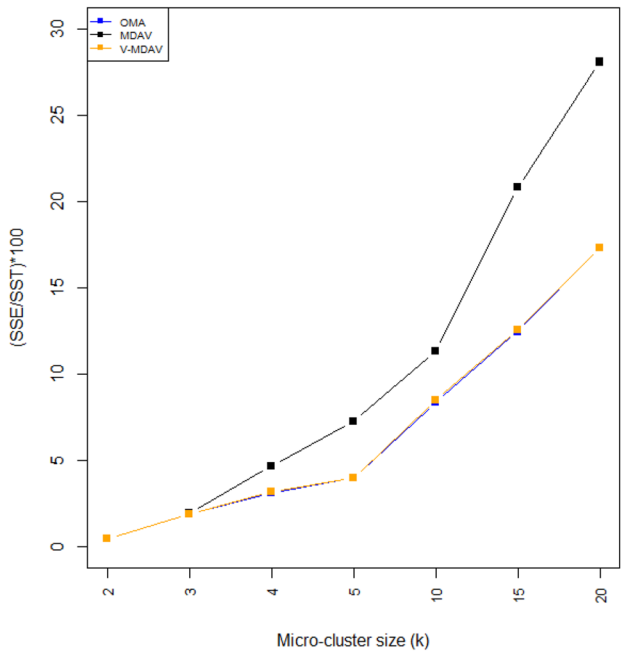
4.5 Unsupervised discretization for rounding

In the case of unsupervised discretization, the number of intervals/clusters (c) are inversely proportional to IL, whereas it is directly proportional to DR. The higher the number of intervals, the quality of the selected rounding points will also be high resulting a low IL and a DR. When the number of intervals is few, a large number of data points fall into a given interval thus introducing a high data distortion once a rounding point is selected for discretization.

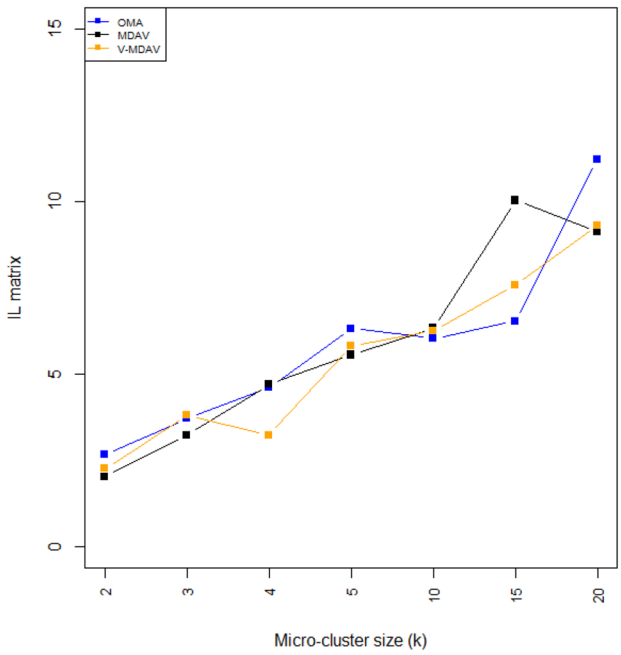
A noteworthy point is that, in the case of unsupervised discretization a high value for the anonymity constraint c indicates low privacy whereas this behaviour in microaggregation based methods are vice versa. The reason for this is the quality of the selected rounding point which discretizes the values fall into a particular interval/cluster or a micro-cluster.

Here, four unsupervised discretization methods are evaluated with respect to rounding, namely re-sampling based discretization (RBD), equal width discretization (EWD), equal frequency discretization (EFD) and k-means based discretization (KMD). Tables 4, 5 and 6 contains the results of using unsupervised discretization methods to obtain the rounding set with respect to different theoretical distributions. Generally, EWD outperforms the other methods with respect to exponential and normal distributed data. However, we analyse these results more closely. When the number of intervals are fewer (i.e., 3, 5) on exponentially and normally distributed data RBD performs better than the other methods as it reports the lowest average IL-DR score. For normally distributed data when the interval size is 3 ($c = 3$) the IL-DR scores are respectively 60.76, 84.27, 166.76 and 254.56, for RBD, EWD, KMD and EFD. Exponentially distributed data also show that RBD reports the lowest IL-DR score of 69.73 when the interval size is 3, compared to 71.31 by EFD, 104.18 by EWD and 278.75 by KMD. For both normally and exponentially distributed data, when the number of intervals (c) is high (i.e., 25) EWD results in a low IL-DR score compared to the other methods. When the data are uniformly distributed, EFD outperforms the other methods.

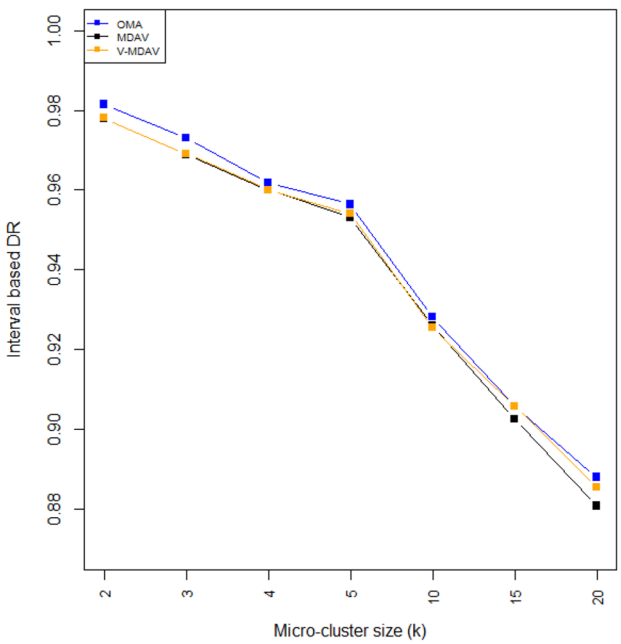
In conclusion, with respect to a fewer number of intervals RBD performs better than the other methods when the data are normally or exponentially distributed. With the aforementioned distribution types, if the data are partitioned into



(a) Tarragona dataset - IL versus annonimity constraint.



(b) Tarragona dataset - IL Metrics versus annonimity constraint



(c) Tarragona dataset - Interval based disclosure versus annonimity constraint

Fig. 2 Microaggregation based discretization on Tarragona dataset

a high number of intervals it is advisable to use EWD. However, if the data are uniformly distributed EFD is preferred irrespective of the number of intervals.

Figure 3 depicts the IL and DR analysis on Tarragona dataset when unsupervised discretization methods are used for obtaining the rounding set. A steady decrease in IL and

Table 4 Comparison of IL and DR measures obtained for the exponentially distributed synthetic dataset with varying anonymity constraints (c)

Rounding method	Anonymity constraint (c)	IL metrics	IL1s	IDR	DDR	Score
EWD exponential	3	38.36	378.30	0.04	0.00	104.18
	5	30.91	201.79	0.09	0.01	58.20
	10	9.05	90.49	0.26	0.07	24.97
	15	21.13	71.15	0.39	0.16	23.21
	20	6.45	43.08	0.67	0.31	12.63
	25	14.94	50.72	0.54	0.25	16.61
EFD exponential	3	46.69	238.38	0.12	0.04	71.31
	5	48.03	157.61	0.26	0.10	51.50
	10	110.12	93.24	0.52	0.29	51.04
	15	52.76	69.97	0.72	0.39	30.96
	20	38.58	55.39	0.79	0.43	23.80
	25	34.22	40.73	0.84	0.46	19.06
KM exponential	3	721.23	393.67	0.10	0.01	278.75
	5	246.42	281.97	0.13	0.03	132.14
	10	56.39	131.91	0.34	0.10	47.19
	15	37.69	96.27	0.60	0.20	33.69
	20	44.70	84.10	0.68	0.27	32.44
	25	33.13	70.86	0.80	0.34	26.28
RBD exponential	3	83.77	195.00	0.14	0.02	69.73
	5	52.36	131.56	0.23	0.04	46.05
	10	66.82	99.42	0.36	0.11	41.68
	15	52.78	80.98	0.56	0.19	33.63
	20	67.99	76.46	0.62	0.25	36.33
	25	73.79	71.87	0.69	0.30	36.66

Increasing anonymity constraint (c) value relates to high DR (low privacy) and low IL

gradual increment in DR can be noted when the data are split into a high number of intervals. As per the results, RBD performs poorly with regarding to IL. This is indicated by the high SSE ratio and IL metrics values shown in Fig. 3a, b. In this case, EWD outperforms the other methods in terms of IL and DR.

4.6 Comparative analysis of the rounding methods

Here, we comparatively evaluate microaggregation and un-supervised discretization methods for rounding, based on their mean *IL-DR Scores*. In this case, an average value for *IL-DR Scores* are obtained over differing k or c values. As discussed earlier, the selected k or c parameter values ensure approximately a similar number of data points are used to form each micro-cluster or interval/cluster. Therefore, this comparison can be justified. Figure 4 depicts the results of applying different rounding methods on different synthetic datasets. It can be noted that compared to normally and uniformly distributed data, exponentially distributed data incurs a high privacy-utility trade-off. When considering microaggregation based methods, OMA is more suitable for normally and uniformly distributed data. For exponentially distributed data VMDAV is preferable. Out of unsupervised discretization methods, EWD is more suitable for exponentially or normally distributed data. EFD performs better when the data are uniformly distributed. Overall OMA, VMDAV, EWD methods are more suitable for obtaining the rounding set compared to other methods under consideration.

5 Impact of rounding for modelling data

In this section, we explore the impact of rounding when data are modelled using machine learning algorithms. Two types of machine learning algorithms are used to build the models: (a) linear regression, and (b) decision trees. The quality of the models is evaluated based on the classification accuracy, R^2 values and the mean squared error (MSE) based on their relevance. We compare the impact of applying different discretization methods based on the above mentioned evaluation criteria. The results of the different discretization methods are measured with varying anonymity constraint values (k for microaggregation and c for unsupervised discretization).

Table 5 Comparison of IL and DR measures obtained for the uniformly distributed synthetic dataset with varying anonymity constraints (c)

Rounding method	Anonymity constraint (c)	IL metrics	IL1s	IDR	DDR	IL-DR score
EWD uniform	3	8.92	192.29	0.11	0.04	50.34
	5	169.69	112.67	0.19	0.10	70.66
	10	30.53	55.78	0.35	0.29	21.73
	15	33.11	35.61	0.50	0.39	17.40
	20	9.23	25.06	0.68	0.44	8.85
	25	9.55	19.14	0.85	0.46	7.50
EFD uniform	3	26.21	176.25	0.15	0.04	50.66
	5	118.17	96.02	0.25	0.10	53.64
	10	82.68	39.00	0.52	0.28	30.62
	15	55.23	25.91	0.69	0.38	20.55
	20	4.84	21.49	0.81	0.44	6.90
	25	11.19	16.44	0.93	0.46	7.25
KM uniform	3	444.61	365.00	0.09	0.02	202.43
	5	224.58	238.74	0.14	0.04	115.87
	10	127.93	124.75	0.28	0.12	63.27
	15	27.65	90.95	0.40	0.21	29.80
	20	36.80	71.42	0.56	0.28	27.27
	25	12.12	56.48	0.68	0.34	17.40
RBD uniform	3	47.60	171.48	0.13	0.02	54.81
	5	177.20	120.92	0.17	0.04	74.58
	10	15.59	73.15	0.32	0.12	22.29
	15	15.16	55.61	0.36	0.17	17.83
	20	10.53	37.41	0.51	0.28	12.18
	25	3.40	33.82	0.55	0.31	9.52

Increasing anonymity constraint (c) value relates to high DR (low privacy) and low IL

For the unsupervised discretization methods, apart from the user defined c values the number of bins are also decided based on *FreedmanDiaconis rule* which is widely used for deriving the interval width based on the following formula, $width = 2 * \frac{IQR(x)}{\sqrt[3]{n}}$. The same formula is used to decide the micro-cluster size k, for each variable in the dataset as, $k = \frac{number\ of\ data\ points}{width}$. In this case, instead of processing each variable with the same anonymity constraint value (which is either the number of intervals (c) or the micro-cluster size (k)) it is determined per variable based on the FreedmanDiaconis rule.

5.1 Evaluating model accuracy

Microaggregation and unsupervised discretization methods are used to obtain the rounding set. For linear regression R^2 and MSE values are used to evaluate the model utility. R^2 value indicates the *goodness of fit*. It explains how close the actual data points are to the fitted regression line. In other words, this is the variation of the response variable that can be explained by the model. Therefore, a higher R^2 value illustrates a model with a good fit for the underlying data. However, in this case R^2 cannot be used for comparative analysis as they are measured on different independent training data sets obtained through different rounding methods. Instead, it is used to understand the relationship between the data and the respective models. In order to compare the model utility MSE is used, and this is calculated as $MSE = \frac{1}{n}\sum_{i=1}^{n}(y - \hat{y})^2$. Here, y indicates the original values of the response variable before rounding, and $\hat{y}$ indicates the predicted values. The lower the MSE, the better the models are.

In the case of decision trees, classification accuracy is used for the comparison. Here, two types of accuracy figures are obtained. Acc_V is the validation accuracy of a given model on the rounded testing data. Acc_O is the classification accuracy of the model, with respect to original (un-rounded) data. Here, Acc_O is used for comparison.

5.2 Experimental setup

As explained earlier, in each scenario a different discretization method is used to obtain the rounding set, and then the original values are encoded using them. When applying EFD

Table 6 Comparison of IL and DR measures obtained for the normally distributed synthetic dataset with varying anonymity constraints (c)

Rounding method	Anonymity constraint (c)	IL metrics	IL1s	IDR	DDR	IL-DR score
EWD normal	3	74.72	262.23	0.09	0.04	84.27
	5	38.27	169.53	0.15	0.08	52.01
	10	12.24	79.88	0.29	0.18	23.15
	15	8.62	48.81	0.44	0.27	14.54
	20	7.76	36.00	0.53	0.34	11.16
	25	5.60	27.41	0.68	0.38	8.52
EFD normal	3	768.79	249.30	0.10	0.04	254.56
	5	50.10	151.99	0.17	0.10	50.59
	10	12.70	79.99	0.38	0.28	23.34
	15	8.71	55.19	0.57	0.38	16.21
	20	6.30	42.03	0.72	0.42	12.37
	25	6.47	34.25	0.80	0.44	10.49
KM normal	3	58.76	608.24	0.05	0.01	166.76
	5	48.48	353.31	0.10	0.03	100.48
	10	17.83	181.04	0.20	0.09	49.79
	15	5.67	122.52	0.31	0.17	32.17
	20	16.63	95.34	0.43	0.26	28.16
	25	5.10	78.30	0.58	0.32	21.07
RBD normal	3	23.62	219.33	0.07	0.02	60.76
	5	25.42	155.80	0.11	0.04	45.34
	10	17.44	87.47	0.24	0.11	26.31
	15	14.07	79.39	0.35	0.18	23.50
	20	14.51	71.02	0.38	0.22	21.53
	25	14.14	60.11	0.52	0.32	18.77

Increasing anonymity constraint (c) value relates to high DR (low privacy) and low IL

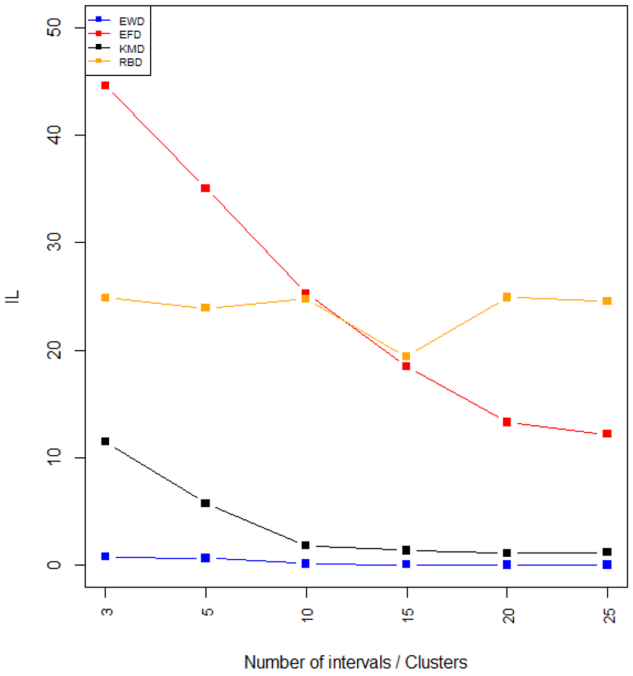
some variables cannot be partition exactly into the specified number of intervals c. This is due to the fact that unique partitions cannot be created for the variable when a particular c value is determined. In such cases, the next highest number of intervals are selected for discretization. The same approach was used when partitioning data into x number of quantiles in RBD.

Usually, SDC/masking methods are applied only to the quasi identifiers-the variables that work as indirect identifiers and can be used for re-identification or record linkage-and the sensitive (dependent) variables are left in their original form. By masking such variables their uniqueness is concealed, thus limiting the risk of disclosure. However, in these test scenarios we have considered all the continuous variables as quasi identifiers and masked them using microaggregation and discretization methods. For the datasets used to train decision trees, the sensitive variable (class variable) is left as it is since they are categorical in nature.

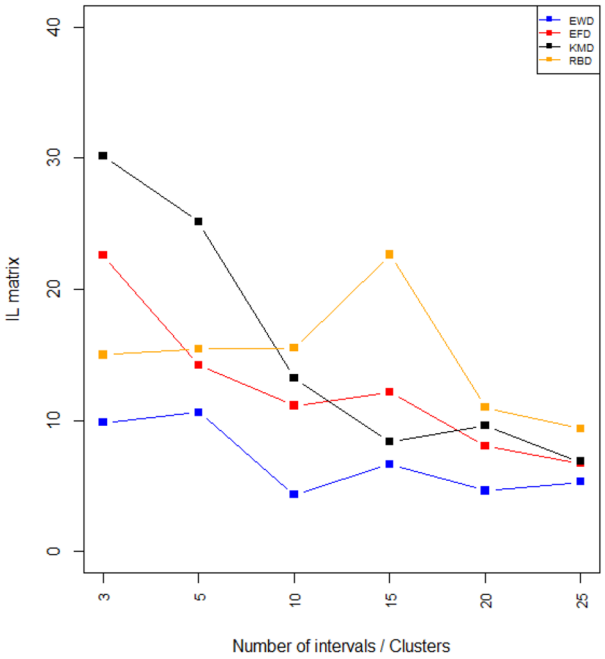
In the case of linear regression datasets, the sensitive (dependent) variable is also numerical. From the preliminary test results, it was noted that when all the continuous variables of a given dataset is masked (fully discretized), including the sensitive (dependent) variable, the trained model's utility is better than (low MSE value) leaving the sensitive (dependent) variable unmasked. For example, on Boston housing prices dataset when the LR models are built on a fully discretized dataset, the utility of the discretized models are equal or better than the original model $\frac{26}{55}$ times. As opposed to the above, when the sensitive (dependent) variable is not masked, always the original model reports a better utility value. This is resulted by the strengthened correlation among the variables when the same treatment is applied to the predictive variable. Therefore, in the LR test cases the datasets are fully masked.

5.3 Results

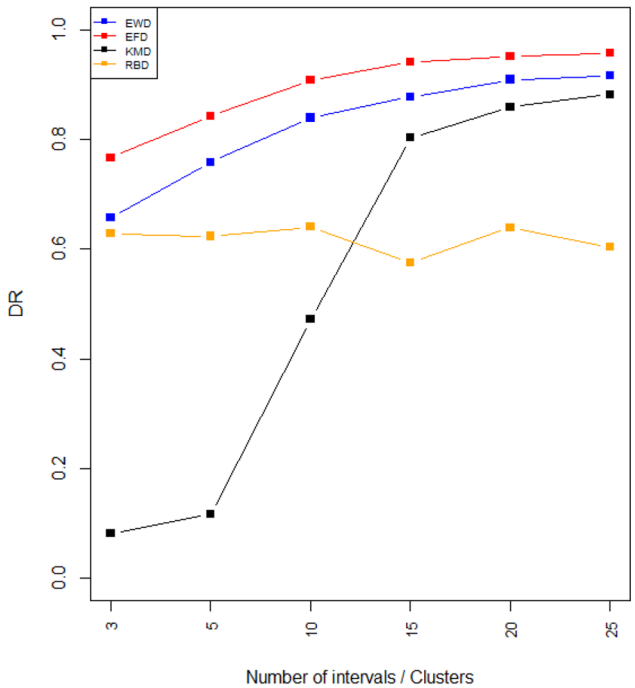
Table 7 illustrates the results of linear regression on rounded training data with varying anonymity constraint values (k or c). When the results are compared based on the MSE values, it can be seen that MDAV reports the highest number of instances where the MSE values are lower than or equal to the original model (baseline model). For a k parameter value as high as 25, all the MA based methods are reporting MSE values less than the original model, indicating that the discretized models are not only providing a privacy guarantee



(a) Tarragona dataset - IL versus annonimity constraint.



(b) Tarragona dataset - IL Metrics versus annonimity constraint



(c) Tarragona dataset - Interval based disclosure versus annonimity constraint

Fig. 3 Unsupervised discretization methods on Tarragona dataset

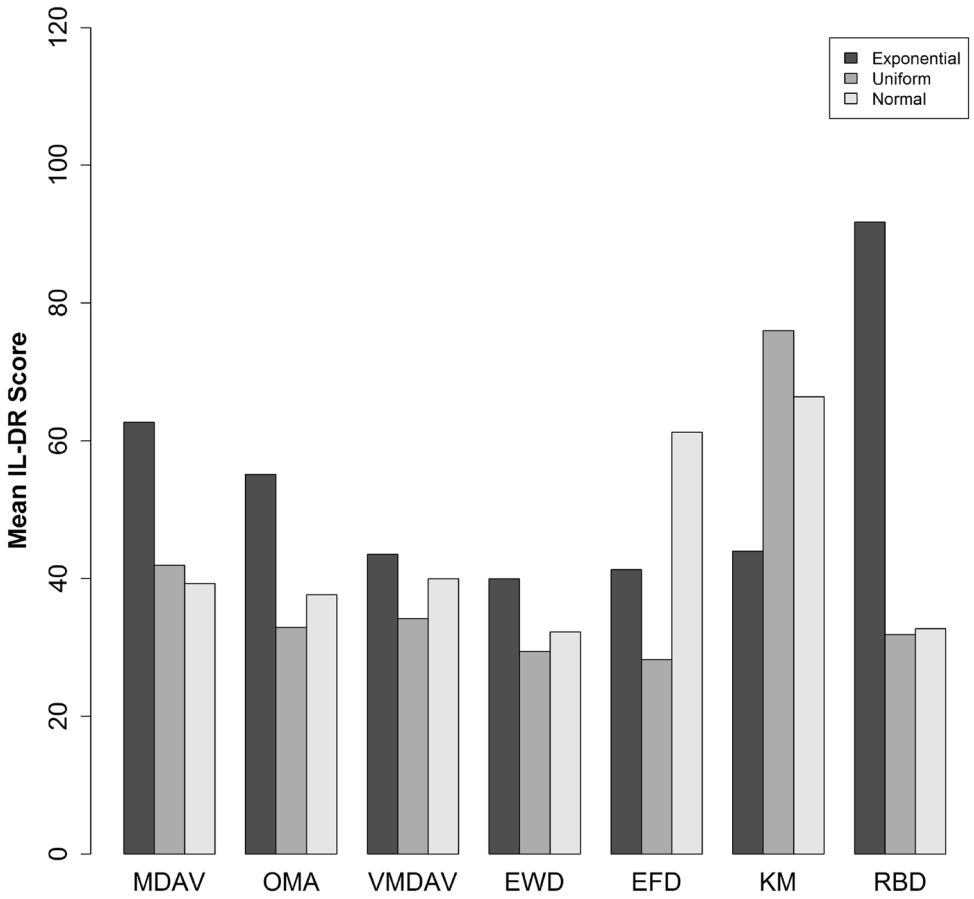


Fig. 4 Mean IL-DR score for different rounding methods

but also results in high predictive accuracy. Here, the MSE values (prediction accuracy) are obtained from the models built on rounded/discretized data which are then used to obtain the predictions on the original data. The lowest MSE values of 14.57 and 17.88 are reported by EFD when the parameter c is set to 3 and 10 respectively. Both EWD and RBD provide a higher number of instances where the predictive accuracies are better than the original case. Moreover, 3 out of 4 cases of using *Freedman–Diaconis rule* with respect to unsupervised discretization methods also result in better predictive accuracies compared to the baseline model. However, none of the KMD instances were able to achieve the baseline predictive accuracy or better. As shown by the results the IL loss values (shown by IL metrics) are not directly related to the predictive accuracy of the models. On average MA based methods work better than the unsupervised discretization methods.

Predictive accuracy of the ML models is not that susceptible to the information loss caused by discretization. The discretized data are deviated from its original form, but the statistical properties required to model the data are preserved so that they can be used to train useful ML models. In the case of LR we have obtained the mean absolute correlation that summarizes the correlation matrix of a given data file. As indicated by the results most of the anonymity constraint value and the masking methods combinations maintain their correlation values within ± 0.2 from the original value. Generally, the higher correlation results in lower MSE and vice versa.

Table 8 illustrates the results of applying rounding on decision tree classifiers. Two accuracy measures are taken for the evaluation purpose. Acc_V indicates the classification accuracy of a given ML model on its test data, and Acc_O indicates the classification accuracy of a given ML model with respect to the original data. In this case, models are built on the rounded/ discretized data and evaluate on the original data. This criterion is used as an evaluation measure to understand the utility of the models built on masked data.

The average ACC_O values reported by each method is as follows; OMA—0.9289, MDAV—0.93, VMDAV—0.936, EWD—0.95, EFD—0.946, KMD—0.918 and RBD—0.924. As shown by the results EWD, EFD and VMDAV methods perform better than the rest. Showing the same pattern as previous KMD performs poorly. On the discretized/rounded dataset we derived the entropy values based on Shannon's entropy for each variable and sum it up to obtain the total entropy for each variable. As it is shown by the results when data are discretized, the entropy is decreased. Moreover,

Table 7 Linear regression on *Boston housing prices* dataset

Rounding method	Anonymity constraint (k/c)	MSE	R^2	Mean absolute correlation	IL metrics	IL1S
Original model	–	21.89	0.74	0.46	0.00	0.00
OMA	2	21.89	0.74	0.46	2.33	12.66
	3	21.91	0.74	0.46	2.78	21.15
	5	21.90	0.74	0.46	2.36	42.25
	7	21.87	0.74	0.46	2.97	65.87
	10	22.03	0.74	0.46	2.39	96.33
	15	22.00	0.74	0.47	2.73	145.82
	20	21.65	0.74	0.47	3.62	201.15
	25	21.58	0.75	0.47	4.07	254.65
	FreedmanDiaconis rule	22.13	0.74	0.46	3.73	298.22
MDAV	2	21.89	0.74	0.46	0.82	21.97
	3	21.94	0.74	0.46	0.95	40.44
	5	21.84	0.74	0.46	1.18	63.28
	7	21.82	0.74	0.46	1.32	95.86
	10	21.79	0.74	0.46	1.61	141.55
	15	21.47	0.75	0.46	2.28	215.80
	20	21.64	0.75	0.46	2.84	263.42
	25	21.56	0.75	0.46	3.09	299.78
	FreedmanDiaconis rule	22.84	0.74	0.46	2.67	434.27
VMDAV	2	21.88	0.74	0.46	1.13	20.59
	3	21.86	0.74	0.46	1.23	37.02
	5	21.87	0.74	0.46	1.56	55.68
	7	21.99	0.74	0.46	1.11	93.31
	10	21.95	0.75	0.46	1.71	122.94
	15	22.07	0.74	0.46	1.93	176.33
	20	22.12	0.73	0.46	2.30	274.15
	25	21.73	0.75	0.47	2.86	296.07
	FreedmanDiaconis rule	22.36	0.74	0.46	2.50	346.52
EWD	3	24.13	0.69	0.43	5.66	7533.45
	5	23.58	0.74	0.46	5.60	4456.77
	10	21.79	0.74	0.46	3.78	2396.30
	15	22.07	0.74	0.46	4.65	1630.63
	20	21.62	0.74	0.45	3.09	1228.44
	30	21.74	0.74	0.46	2.38	810.90
	Freedman–Diaconis rule	21.81	0.74	0.46	4.05	1362.59
EFD	3	14.57	0.70	0.44	6.31	7268.81
	5	27.18	0.70	0.45	4.78	4517.43
	10	17.88	0.76	0.46	4.42	2829.41
	15	22.99	0.73	0.46	4.43	1987.84
	20	23.00	0.73	0.46	3.97	1860.71
	30	22.23	0.74	0.46	4.15	1364.85
	Freedman–Diaconis rule	21.61	0.74	0.45	3.94	1739.21
KMD	3	39.15	0.56	0.38	9.19	13856.35
	5	27.89	0.66	0.41	8.89	9074.86
	10	22.83	0.71	0.44	5.41	4730.21
	15	22.94	0.73	0.45	4.49	3070.48
	20	22.67	0.73	0.45	5.16	2839.30
	30	22.37	0.73	0.45	3.30	2010.76
	Freedman–Diaconis rule	22.48	0.72	0.44	5.17	3856.14

Table 7 (continued)

Rounding method	Anonymity constraint (k/c)	MSE	R^2	Mean absolute correlation	IL metrics	IL1S
RBD	3	33.49	0.60	0.44	8.26	12894.81
	5	23.40	0.70	0.47	7.26	8610.45
	10	20.41	0.73	0.48	6.15	6134.04
	15	19.47	0.75	0.49	7.41	5547.41
	20	18.96	0.75	0.48	5.70	5071.81
	30	19.49	0.75	0.49	6.82	5053.47
	Freedman–Diaconis rule	18.91	0.76	0.49	7.44	6282.77

entropy is inversely related to information loss. As anonymity constraint value (k) increases in MA methods the entropy drops gradually, and the opposite behaviour can be seen with respect to unsupervised discretization methods. However, low entropy does not impact the validation accuracy (ACC_V) of the decision tree models. The correlation between ACC_V and entropy is 0.14 which indicates a negligible positive relationship. When correlation is measured between ACC_O and entropy it shows a moderately inverse relationship of -0.54 where the accuracy increases as the entropy decrease. This behaviour can be explained as below. Low entropy levels indicate less amount of information or less uncertainty in data. This can also be attributed to the less diversity in the underlying data. When continuous data are less diverse due to discretization (as many unique data points are replaced with centroids), the ML models derived from such data are more generalized compared to the models built on original data as they are not over-fitted. Therefore, the DT models built on discretized data still shows a fairly good accuracy when they are used to predict previously unseen data despite the IL incurred in the process.

In these experiments, we have also employed two supervised discretization methods that use class information in order to discretize continuous data. The two methods are, namely discretization using minimum description length principle (MDLP) and discretization using ChiMerge (ChiM) algorithm. MDLP is an entropy based method whereas ChiM uses χ^2 statistics to determine the discretization points. Many empirical studies conducted in the literature have shown that supervised discretization methods are more effective compared to the unsupervised discretization methods in terms of maintaining a high predictive accuracy when training and validation accuracies are determined. In this case, we are also interested in measuring Acc_O which indicates the utility of a discretized model against the original data when supervised discretization methods are used. Also, we want to explore whether supervised discretization methods can be used as an alternative for generating the rounding set. When discretized using supervised methods, validation accuracy (Acc_V) is equal or greater than the original model. However, when the original data are classified using the discretized ML models (Acc_O) it can be seen that the accuracies are far lower than any other method. Nevertheless, low entropy values are noted in this case. However, the reduction of entropy has to be done carefully. Otherwise, as shown by the outcomes of the supervised discretization methods, a significant reduction of entropy can lose useful information so that the minimum amount of information required to learn the model is no longer available. Thus, the models built on such data inherits a poor predictive accuracy when tested with new data. This indicates that for rounding supervised discretization methods are not ideal.

When the overall results are considered it can be seen that *Freedman–Diaconis rule* provides a very close accuracy to the original case despite the dataset or the discretization methods adopted. Interval based DR on Boston housing dataset when the *Freedman–Diaconis rule* is applied vary from 0.40 to 0.83 (OMA-0.5889, MDAV-0.5945, VMDAV-0.5561, EWD-0.8346, EFD-0.7929, KMD-0.4996 and RBD-0.40612), which provides a good privacy-utility trade-off compared to the other methods. Therefore this method can be considered when selecting anonymity constraint values as it provides a good accuracy while minimizing the disclosure risk. Especially, in the cases where the data owners do not have a clear insight on what value should be selected as the anonymity constraint (degree of privacy).

We have also performed experiments with iris and faithful datasets, but for the sake of conciseness in the discussion it is restricted to the ones mentioned above. Results of the datasets iris and faithful were similar to the above.

6 Discussion

In this work, we have explored different discretization methods in order to mask the numerical data. Based on the results it can be concluded that microaggregation (MA) based methods incur a low IL compared to unsupervised

Table 8 Decision tree models trained on *Wine classification* dataset

Rounding method	Anonymity constraint (k/c)	Acc_V	Acc_O	Entrpoy	IL metrics	IL1S
Original model	–	0.88	0.94	82.80	0.00	0.00
OMA	2	0.88	0.93	78.42	4.96	235.16
	3	0.89	0.94	73.60	6.72	427.77
	5	0.86	0.93	65.73	6.22	756.31
	7	0.90	0.95	60.34	9.87	1054.22
	10	0.85	0.93	54.24	10.51	1456.06
	15	0.89	0.94	46.41	10.84	2194.82
	20	0.86	0.90	40.39	12.60	2913.20
	25	0.86	0.91	38.05	10.21	3420.58
	Freedman–Diaconis rule	0.89	0.93	45.07	10.52	2365.54
MDAV	2	0.88	0.93	81.12	2.33	331.50
	3	0.88	0.93	76.15	3.85	580.20
	5	0.90	0.95	68.00	6.38	959.74
	7	0.86	0.94	61.88	8.28	1260.79
	10	0.87	0.93	54.45	11.42	1854.36
	15	0.86	0.92	46.12	10.37	2718.31
	20	0.92	0.92	40.00	14.42	3593.31
	25	0.84	0.91	38.05	8.90	3543.17
	Freedman–Diaconis rule	0.90	0.94	45.33	8.32	2620.36
VMDAV	2	0.92	0.95	78.41	2.88	305.69
	3	0.86	0.94	73.15	3.91	543.03
	5	0.90	0.94	64.58	4.66	884.94
	7	0.88	0.93	58.96	6.71	1185.81
	10	0.87	0.95	52.95	8.69	1612.67
	15	0.93	0.95	45.70	10.52	2390.37
	20	0.89	0.92	40.43	10.74	3076.69
	25	0.85	0.92	35.57	11.00	3740.52
	Freedman–Diaconis rule	0.87	0.92	43.67	10.28	2609.10
EWD	3	0.92	0.96	36.90	13.23	4533.74
	5	0.89	0.94	45.64	6.97	2761.45
	10	0.93	0.96	56.12	8.40	1444.51
	15	0.90	0.96	62.04	5.63	968.42
	20	0.87	0.95	66.00	3.77	764.86
	30	0.86	0.93	70.94	4.24	452.76
	Freedman–Diaconis rule	0.89	0.95	57.17	8.09	1378.13
EFD	3	0.83	0.94	38.89	20.23	4149.45
	5	0.86	0.94	46.73	11.20	2786.93
	10	0.88	0.93	58.57	6.44	1455.19
	15	0.90	0.96	65.08	8.91	964.08
	20	0.89	0.95	68.94	8.07	722.20
	30	0.90	0.94	72.29	4.92	562.61
	Freedman–Diaconis rule	0.93	0.96	59.81	5.89	1329.15
KMD	3	0.86	0.90	21.58	17.19	7868.63
	5	0.88	0.91	30.17	12.31	4768.16
	10	0.89	0.92	42.94	8.99	2526.06
	15	0.85	0.92	49.32	8.20	1785.65
	20	0.90	0.93	54.70	5.91	1366.02
	30	0.86	0.93	61.75	5.88	961.64
	Freedman–Diaconis rule	0.87	0.92	43.35	7.67	2491.14

Table 8 (continued)

Rounding method	Anonymity constraint (k/c)	Acc_V	Acc_O	Entrpoy	IL metrics	IL1S
RBD	3	0.89	0.91	43.14	8.81	2681.74
	5	0.85	0.93	42.95	10.25	2664.94
	10	0.88	0.93	42.85	11.46	2658.41
	15	0.86	0.91	42.76	11.89	2707.15
	20	0.87	0.93	42.73	12.16	2657.79
	30	0.85	0.91	42.78	8.16	2724.58
	Freedman–Diaconis rule	0.90	0.95	42.89	10.18	2687.42
MDLP	–	0.88	0.39	18.08	33.03	10317.56
ChiM	–	0.93	0.30	31.89	12.75	6691.81

The accuracy values are obtained based 10-cross fold validation and mean values are reported

discretization methods and intuitively results in an improved DR. A single discretization technique or an anonymity constraint value cannot be determined as the best in addressing the privacy-utility trade-off, as the nature of the underlying data impacts that. However, methods like OMA, VMDAV and EFD perform well in most of the instances. For normal and exponentially distributed data with a small number of intervals (c), RBD (Re-sampling based discretization) is suitable over the other methods as it incurs a low privacy-utility trade-off. For a uniformly distributed dataset EWD or EFD can be used to obtain the rounding set with minimal privacy-utility trade-off. On average, uniformly distributed data can be discretized with minimal IL compared to other data distributions we have checked here. Normally distributed data incurs the highest IL whereas exponentially distributed data reports the highest DR despite the rounding method we use. Therefore, examining the data distribution beforehand can be helpful in deciding the rounding method and privacy parameters such as interval/cluster size, aggregation method etc. Instead of using a fixed anonymity constraint value for all the variables, we can define the size of k or c per each variable in the case of univariate discretization. We have illustrated some example cases in the experiments using *Freedman–Diaconis rule* to determine the anonymity constraint without having to specify it by the users. The experiments show that in both unsupervised discretization and MA, the outcome of using the above approach is very close to the baseline results.

As discussed earlier, releasing of rounded/discretized data helps to mitigate the disclosure risk. However, this results in an IL which directly impacts the analytical value of the underlying dataset. In this work, we explored how IL caused by rounding can influence the predictive power of the machine learning models. It seems that IL does not necessarily result in poor predictive accuracy in ML models. In many cases, rounding improves the model utility as it reduces the noise in the data so that the ML algorithm can learn without the risk of over-fitting.

For example, data owners release a perturbed version of original data to the data analysts in order to minimize the risk of disclosure. Assume the data masking method used is carefully tuned so that the analytical value of data is not completely destroyed. In this case, the models trained on such data should also have good predictive accuracy, maybe with a slight reduction compared to the original model. When we build a ML model, one of the main concerns is to avoid model overfitting. If the generated models are more generalized towards the training data a better accuracy can be seen when new data are classified using these models. In ML, this is mainly achieved through regularization. In our case, generalized data are used to train the models with the expectation that it would result in simple but accurate models. The other advantage is that these data masking techniques also guarantee a degree of privacy for the data in use. This privacy and utility trade-off in model building can be justified if we are dealing with sensitive information. Considering the above mentioned facts, it can be concluded that models built on rounded data are generalized, thus it secures a good predictive accuracy.

7 Conclusion

"Rounding" is a numerical data masking technique which has not been discussed previously in the literature with empirical results. The operating principle of rounding can be seen as discretization or quantization where the continuous values are mapped into a discrete space. In this work, we discuss rounding in a unified way with unsupervised discretization and microaggregation where these methods are used to generate the rounding sets. Also, we have introduced a re-sampling based discretization method for continuous data which works better with a small number of intervals thus minimizing the disclosure risk. Then these methods are evaluated based on their information loss and disclosure risk with respect to theoretical distributions and real world data. Finally, the rounded

data are used to train linear regression, and decision tree models and the impact of rounding towards model accuracy is discussed. Based on the results, it can be concluded that generally, microaggregation based methods are more suitable for deriving the rounding set. However, based on the data distribution in some cases unsupervised discretization methods outperforms microaggregation methods. Also, we have used *Freedman–Diaconis rule* to define the anonymity constraint value per each attribute and shown that this method can be used to minimize disclosure risk while maintaining a model utility closer to the benchmark (original) model.

This work is focused on univariate, deterministic rounding. In future work, it will be interesting to explore multivariate and stochastic rounding based on the above discussed methods. Also, a study on different aggregation methods that can be used to obtain the centroids/rounding points will be interesting in terms of managing the IL and DR.

Acknowledgements Open access funding provided by University of Skövde. This work is supported by Vetenskapsrdet project: "Disclosure risk and transparency in big data privacy (VR 2016-03346, 2017-2020).

References

Abidi B, Ben Yahia S, Perera C (2018) Hybrid microaggregation for privacy preserving data mining. J Ambient Intell Human Comput. https://doi.org/10.1007/s12652-018-1122-7

Bennasar M, Setchi R, Hicks Y (2012) Unsupervised discretization method based on adjustable intervals. In: Advances in knowledge-based and intelligent information and engineering systems - 16th annual KES conference, San Sebastian, Spain, 10–12 September 2012, pp 79–87. https://doi.org/10.3233/978-1-61499-105-2-79

Chettri SK, Paul B, Dutta AK (2012) A comparative study on microaggregation techniques for microdata protection. Int J Data Min Knowl Manag Process 2(6):27

Chmielewski MR, Grzymala-Busse JW (1996) Global discretization of continuous attributes as preprocessing for machine learning. Int J Approx Reason 15(4):319–331

Domingo-Ferrer J (2008) A survey of inference control methods for privacy-preserving data mining. In: Privacy-preserving data mining - models and algorithms, pp 53–80. https://doi.org/10.1007/978-0-387-70992-5_3

Domingo-Ferrer J, Torra V (2001) A quantitative comparison of disclosure control methods for microdata. Confid Discl Data Access Theory Pract Appl Stat Agencies 111–134

Dougherty J, Kohavi R, Sahami M (1995) Supervised and unsupervised discretization of continuous features. In: Machine learning, proceedings of the twelfth international conference on machine learning, Tahoe City, California, USA, July 9–12, pp 194–202

García S, Luengo J, Sáez JA, López V, Herrera F (2013) A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning. IEEE Trans Knowl Data Eng 25(4):734–750

Hansen SL, Mukherjee S (2003) A polynomial algorithm for optimal univariate microaggregation. IEEE Trans Knowl Data Eng 15(4):1043–1044

Ibrahim MH, Hacibeyoğlu M (2016) Comparison of the effect of unsupervised and supervised discretization methods on classification process. Int J Intell Syst Appl Eng 4(Special Issue–1):105–108

Jiang SY, Li X, Zheng Q, Wang LX (2009) Approximate equal frequency discretization method. In: WRI global congress on intelligent systems, 2009, vol 3. IEEE, pp 514–518

Lloyd S (1982) Least squares quantization in pcm. IEEE Trans Inf Theory 28(2):129–137

Max J (1960) Quantizing for minimum distortion. IRE Trans Inf Theory 6(1):7–12

Miché Y, Oliver I, Holtmanns S, Kalliola A, Akusok A, Lendasse A, Björk K (2016) Data anonymization as a vector quantization problem: Control over privacy for health data. In: Proceedings of availability, reliability, and security in information systems: IFIP WG 8.4, 8.9, TC 5th international cross-domain conference, CD-ARES 2016, and workshop on privacy aware machine learning for health data science, PAML 2016, Salzburg, Austria, August 31–September 2, pp 193–203

Pfahringer B (1995) Compression-based discretization of continuous attributes. In: Machine learning, proceedings of the twelfth international conference on machine learning, Tahoe City, California, USA, 9–12 July 1995, pp 456–463. https://doi.org/10.1016/b978-1-55860-377-6.50063-3

Ramírez-Gallego S, García S, Mouriño-Talín H, Martínez-Rego D, Bolón-Canedo V, Alonso-Betanzos A, Benítez JM, Herrera F (2016) Data discretization: taxonomy and big data challenge. Wiley Interdiscip Rev Data Min Knowl Discov 6(1):5–21

Rebollo-Monedero D, Forné J, Soriano M (2011) An algorithm for k-anonymous microaggregation and clustering inspired by the design of distortion-optimized quantizers. Data Knowl Eng 70(10):892–921

Rebollo-Monedero D, Forné J, Pallarès E, Parra-Arnau J (2013) A modification of the lloyd algorithm for k-anonymous quantization. Inf Sci 222:185–202

Solanas A, Martínez-Ballesté A, Domingo-Ferrer J (2006) VMDAV: a multivariate microaggregation with variable group size. In: 17th IASC symposium on computational statistics (COMPSTAT), pp 917–925

Sweeney L (2002) k-Anonymity: a model for protecting privacy. Int J Uncertain Fuzziness Knowl Based Syst 10(5):557–570

Templ M (2017) Statistical disclosure control for microdata: methods and applications in R. Springer, Berlin

Torra V (2017) Data privacy: foundations, new developments and the big data challenge. Springer, Berlin

Willenborg L, De Waal T (2012) Elements of statistical disclosure control, vol 155. Springer, Berlin

Yang Y, Webb GI (2002) A comparative study of discretization methods for naive-Bayes classifiers. In: Proceedings of PKAW 2002: the 2002 Pacific Rim knowledge acquisition workshop, vol 2002, pp 159–173

Zhang G (2011) Privacy data preserving method based on fuzzy discretization. In: Eighth international conference on fuzzy systems and knowledge discovery, FSKD 2011, 26–28 July 2011. Shanghai, China, pp 1201–1205

Zhu Y, Fu Y, Fu H (2009) Preserving privacy in time series data classification by discretization. In: Proceedings of machine learning and data mining in pattern recognition, 6th international conference, MLDM 2009, Leipzig, Germany, July 23–25, pp 53–67