Measuring Information Quality for Privacy Preserving Data Mining

Sam Fletcher and Md Zahidul Islam

***Abstract*—In the strive for knowledge discovery in a world of ever-growing data collection, it is important that even if a dataset is altered to preserve people's privacy, the information in the dataset retains as much quality as possible. In this context, "quality" refers to the accuracy or usefulness of the information retrievable from a dataset. Defining and measuring the loss of information after meeting privacy requirements proves difficult however. Techniques have been developed to measure the information quality of a dataset for a variety of anonymization techniques including Generalization, Suppression, and Randomization. Some measures analyze the data, while others analyze the outputted data mining results from tasks such as Clustering and Classification. This survey discusses a collection of information measures, and issues surrounding their usage and limitations.**

***Index Terms*—Anonymization, data mining, data quality, privacy preserving data mining.**

I. Introduction

Within the Privacy Preserving Data Publishing (PPDP) community, preventing sensitive information about individuals from being inferred is a top priority. This is known as "anonymization". One of the key concepts in PPDP is the trade-off that is inherently present when "anonymizing" data: balancing the increase in security with the decrease in information quality. The majority of previous work has focused on the difficult problem of defining and measuring privacy [1], [2]. This paper explores the other side of the trade-off: information quality. A lot of the time, simplistic measures are developed to provide an estimate of the information quality, or statistical techniques are borrowed from the SDC (Statistical Disclosure Control) community. While robust, these evaluation techniques often fail to capture the nuances that can be present when evaluating specific anonymization tasks, such as generalization [1]. Information measures that target specific anonymization tasks solve this problem; however comparing the results of different measures is an ongoing problem. If two datasets[2] are anonymized with two different techniques, and each technique requires its own information measure, comparing the quality of the datasets can be problematic [1], [3], [4].

In PPDP, the information quality of an anonymized dataset is most often evaluated by measuring the similarity between the anonymized dataset and the original dataset. If the dataset could be used for a variety of reasons and there is no single purpose in mind, the dataset is evaluated in a way that applies to any scenario – we refer to this as measuring the "dataset quality" or "dataset information loss". These types of techniques are discussed in Section II.

Alternatively, if the purpose of the dataset is specific and known, the information quality can be measured in respect to that purpose. Privacy Preserving Data Mining (PPDM; a sect of PPDP) focuses on this type of data, where the quality of the dataset itself is less important than the quality of the outputted data mining[3] results produced from the dataset. Common purposes are classification[4] and clustering[5] [2]. Many patterns in the dataset can be lost after anonymization, even if the dataset itself appears to retain most of its statistical information [6]-[8]. For this reason, information measures have been designed that specifically look at the effect of anonymization on data mining results, and we discuss these in Section III. We call this type of information quality, "data mining quality" or "data mining information loss".

It should be noted that we make a distinction between "information loss" and "information quality" due to the implied comparative nature of the word "loss" – this paper focuses on measures that compare a dataset before and after modification. "Information quality" could refer to this before-and-after comparison, but also to the quality of an isolated dataset (with no comparison). "Information loss" provides more specificity. Measures of information loss are also usable in scenarios outside of privacy preservation, such as data imputation / data cleaning[6]. In these instances, information *gain* is the goal.

II. Dataset Information Loss

Dataset information loss refers to the loss of useful information in the dataset itself. Statistical tests are one way of measuring this, but additional insight can be gained through more targeted measures. Some measures target the quality of the user-defined QID groups (described below in

Manuscript received February 22, 2014; revised April 23, 2014.

S. Fletcher and M. Z. Islam are with the Center for Research in Complex Systems (CRiCS), School of Computing and Mathematics, Charles Sturt University, Bathurst NSW 2795, Australia (e-mail: {safletcher, zislam}@csu.edu.au).

[1] "Generalization" refers to making a value vaguer, such as changing all occurrences of "*apple*" and "*banana*" to "*fruit*".

[2] A "dataset" is a two dimensional table where rows represent independent records (tuples) and columns represent various attributes that describe the records and distinguish them from each other.

[3] "Data mining" refers to using automated algorithms for finding patterns in data.

[4] "Classification" refers to predicting a record's value for an attribute based on the other explanatory attributes. Decision trees and neural networks are commonly-used method for doing so [5].

[5] "Clustering" refers to grouping records in such a way that similar records are grouped together and dissimilar records are grouped in separate clusters [5].

[6] "Data imputation" and "data cleaning" refer to estimating missing values in a dataset and removing misinformation/noise [9].

DOI: 10.7763/IJCTE.2015.V7.924

Section 2.1), while others target data modifications via generalization and suppression (Section 2.2).

A. QID Group Quality

A QID is a user-defined group of quasi-identifying attributes; that is, a collection of attributes[7] that allow an attacker[8] to uniquely identify a record[9]. To do so, the attacker must be able to learn the values of the quasi-identifying attributes from outside sources, such as real life contact, social engineering[10], or from other independent datasets. Upon learning all these values for an individual, the attacker can narrow down the possible records that could represent the targeted individual and potentially learn sensitive information about them (attributes that describe sensitive information are defined by the anonymization expert[11] as "sensitive attributes"). A QID is sometimes referred to as a VID – a "Virtual Identifier" [10].

Discernibility Metric (DM) [11], [12] functions by penalizing each record for how many other records it is indiscernible from in each QID, compared to the original dataset. A "QID group" is defined as any collection of records that have the same values for all the QID attributes. If a record belongs to a QID group with *n* records, then the penalty for that record is *n* – it is indiscernible from *n* records with respect to the attributes in the QID. This naturally leads to considering the penalty per QID group, rather than per record: each QID group incurs a penalty of n^2. Interestingly, it is the conceptual opposite of k-anonymity [13], [14] – a well-known privacy technique that *requires* a user-defined minimum number of indistinguishable records per QID group.

DM is a commonly-used measure [12], [15]–[17] despite its inability to consider the data distribution[12] of the attribute values. As is often the case, the lack of a single robust information measure has led many to adopt an ensemble approach[13], with multiple measures each addressing something missed by the others [3], [16], [18]. It is worth noting that ignoring data distribution is a common shortcoming of information measures, and including statistical evaluations such as KL-divergence [1], [3], [16], chi-square distance [3], [19] and covariance comparisons [20] are possible solutions. Statistical tests are also often the solution to measuring information loss caused by randomization[14] [3], [8], [21]-[24]. In addition to those mentioned above, common tests include regression analysis, mean square error and contingency tables. For further information on these tests, we refer the reader to [25], [26].

B. Generalization and Suppression

Throughout most information measurement literature, the assumption is made that for the purposes of quality evaluation, suppression[15] can be considered as generalizations that generalize a value to its most vague state [1], [4], [13] [14], [18], [27]-[32]. We maintain this assumption, and hereafter only refer to generalization.

Minimal Distortion (MD) [14], [28], [29] (or generalization height [1], [4], [33]) is a penalty-based system where whenever a value (for one record) is generalized, the distortion count is incremented. MD harnesses the taxonomy trees of attributes (also called Domain Generalization Hierarchies [18]), in which each value of an attribute is a leaf in the tree, and the higher nodes represent collective terms for their child nodes (and are thus more vague). Fig. 1 is a simplified example of a taxonomy tree, with the number in each leaf referring to the frequency of the categorical values appearing in the dataset. MD treats each level of generalization separately – if a value is generalized to the parent of its parent node in the attribute's taxonomy tree, 2 units of distortion are added.

As an example, take Fig. 1: if all instances of value *F* were generalized to *A* (a user-defined term that collectively describes *B* and *C*), then 50 records have moved up two levels, resulting in 100 units of distortion. Additionally, however, most algorithms [10], [27], [28], [32]-[34] do not allow multiple levels of generalization for an attribute to co-exist in a dataset (e.g. a record cannot have a value of "*apple*" if another record has a value of "*fruit*" for the same attribute), as this would cause problems with data mining algorithms. Therefore if *F* is generalized to *A*, so too are *G* and *H*, bringing the total distortion up to $50+100+150+300=600$. If they were only generalized to *C*, $MD=50+100+150=300$.

Along with MD, Iyengar's **loss metric** (LM) [32] marked the first work in specifically targeting the information loss caused by generalization. LM is defined as the number of nodes a record's value has been made indistinguishable from (via generalization) compared to the total number of original leaf nodes in the taxonomy tree. This is repeated for each record for the attribute in question, and each attribute's loss is the average over all records.

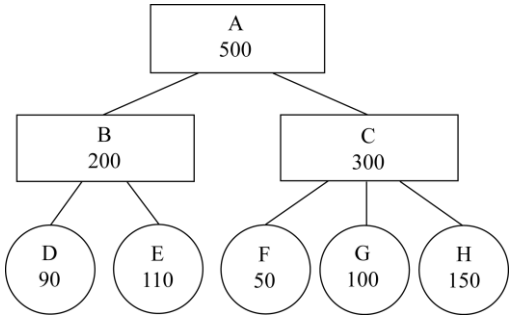

Fig. 1. A generalization taxonomy tree of an attribute.

[7] An "attribute" refers to a column in a dataset describing a particular aspect of records in the dataset.

[8] An "attacker" or "intruder" is someone who has gained access to the dataset (whether legally or not) and is using it for unintended purposes. We will assume they are attempting to target a specific individual in the dataset.

[9] A "record", or "tuple", refers to a row in a dataset, consisting of column (attribute) values that describe the record.

[10] "Social engineering" refers to deliberately misleading people into divulging certain information.

[11] An "anonymization expert" is the person charged with the responsibility of anonymizing a dataset.

[12] "Distribution" refers to how frequent each possible value for a set of data occurs. Commonly, this takes the shape of a normal distribution, otherwise known as a Gaussian distribution.

[13] An "ensemble" is any instance where multiple techniques of a certain type are used to find a more robust result.

[14] "Randomization" refers to adding random noise to numerical values (e.g. "*age=34*" becomes "*age=37*"), or changing categorical values to particular other values with a certain probability (e.g. "*England*" to "*Australia*").

[15] "Suppression" refers to a data value (or record) being completely hidden or deleted.

For example in Fig. 1, generalizing the 50 records in *F* to *C* (which collectively describes *F*, *G* and *H*) would result in the value of 50 records (for one attribute) being indistinguishable from 2 other values (nodes). With 5 leaves in the taxonomy tree, $LM = 2/5$ for those 50 records in regards to the attribute shown in Fig. 1.

For numerical attributes[16], if a value (for example, "5") is generalized (e.g. "3-7"), LM compares the size of the generalized domain[17] (e.g. $7-3=4$) to the total domain size of the attribute (e.g. if the domain is [1, 10], then the domain size is 9), giving a final result of $4/9$ in this example. The LM result for each attribute is averaged over all records. The loss of each attribute is then summed together, giving a final LM result. Unfortunately, LM does not take the distribution of the data into consideration [1].

While both MD and LM take into account how many records are affected by a value being generalized, a major downside is that they treat each generalization as equally damaging [16]. When considering categorical values and collective terms for categorical values, it is unlikely that a user (even an expert) could design each generalization in the taxonomy to have a real-world equivalence to each other. For example, something like $state \rightarrow country$ could have a far bigger impact on a dataset than $birth\_month \rightarrow birth\_year$, and even with modification it could still never truly be equivalent. Defining and measuring differences in generalizations is an open question. Perhaps weighting[18] each generalization based on changes in the information gain (using whatever measure fits the needs of the anonymization expert) of the dataset is a possibility.

ILoss [35] takes the same approach as LM when measuring the information lost to generalizations. It measures the fraction of domain values lost for an attribute by each generalization, just as LM does, and gives the same results for each generalized value as LM would. It differentiates itself by allowing each attribute to also possess a weighting, allowing for the major disadvantage of MD and LM discussed above to be partially solved. While each generalization isn't treated differently, at least each attribute is treated differently based on their user-defined importance. The ILoss values of each record (taking into account the attribute weightings) are then averaged across the whole dataset, resulting in a final ILoss result.

MD, LM and ILoss all place a reliance on the validity of the user-created taxonomy trees, and this is not considered an unreasonable assumption by most [1], [4], [14], [28]-[30], [32], [33], [36]. The assumption is not unanimous however, and solutions exist for automating the creation of taxonomy trees for numerical attributes [36]-[38] and categorical attributes [18].

Generalizations for numerical attributes can be found by finding the optimal binary split for the domain that maximizes the information gain (here, "information gain" is measured using algorithms commonly found in decision trees[19] – see Section 3.2) [37]. Splitting the domain can be repeated until the desired number of generalization levels is achieved, for example a domain [1, 10] might be split into [1, 6] and [7, 10], with [1, 6] being further split into [1, 3] and [4, 6]. Thus a value of "2" would now be "1-3" and a value of "7" would now be "7-10". Alternatively numerical attributes can be generalized using clustering techniques such as iK-Means, where adjacent values with high frequency are grouped together [40].

Categorical attributes prove much more difficult to automatically generalize due to the lack of a natural ordering. A possible solution is to dynamically combine "appropriate" categorical values together – for example, "*apple*" and "*banana*" could be generalized to a value called "*apple_or_banana*" [18]. "Appropriateness" can be defined using any similarity measure at the discretion of the anonymization expert.

Not only is user input vulnerable to human error, but even a perfectly reasonable taxonomy is a commitment that removes all other interpretations. It may seem intuitive to generalize "*apple*" and "*banana*" to "*fruit*", but what if more information (or relevancy to a specific task) could be retained by sorting by sugar content or price or color? User-defined taxonomies create or strengthen certain semantic meanings, while destroying or weakening others [18]. This can result in the anonymization expert making a drastic data mining decision, which is usually outside their job description.

Parallel to these measures of quality is an important concept that should be considered when debating the differences between generalization and randomization: "faithfulness" [1]. Faithfulness, or truthfulness, refers to how confident a data miner can be about the quality of the anonymized data at the record level. Trottini called this "perceived data utility" [41], [42] and warned of the dangers of false confidence – what if a doctor or federal security agency acts on an anonymized record that they falsely believe to be accurate? Generalization offers an advantage over randomization in this case: it can *guarantee* that each record is still as accurate as it was in the original dataset – it's simply vaguer.

III. Data Mining Information Loss

"Data mining information loss" involves comparing the data mining results of an original dataset to the results of an anonymized version. Since the output of various data mining techniques differs greatly, targeted information measures are required. Guo *et al*. described this phenomenon: "utility of any dataset, whether randomized or not, is innately dependent on the tasks that one may perform on it. Without a workload context, it is difficult to say whether a dataset is useful or not" [3]. Data mining quality and dataset quality are not mutually exclusive – often both are tested for, and the

[16] A "numerical" or "continuous" attribute refers to an attribute with natural ordering, where each value can be described relative to the values on either side of it in the ordering.

[17] The "domain" of an attribute is the set of possible values that the attribute can have.

[18] "Weighting" refers to scaling certain parts of an equation by different constants, based on their importance.

[19] A "decision tree" is a method of classifying records in a dataset based on qualities that provide the most useful information to a data miner. Records are filtered into stronger and more specific patterns the further the tree extends [39]. See Fig. 3.

empirical results support the differentiation between dataset quality and data mining quality [18], [20], [32]. The most common data mining techniques are clustering and classification, and these will be addressed below in turn.

A. Clustering

Thus far, the quality of clustering results has proven difficult to robustly capture due to the absence of a strict definition of clustering and a reference point for evaluating the results. The usefulness of a clustering result can easily vary depending on the purpose of clustering. Fung *et al*. described the problem succinctly: "the anonymity problem for cluster analysis does not have class labels to guide the generalization. It is not even clear what 'information for cluster analysis' means and how to evaluate the quality of generalized data in terms of cluster analysis" [36].

This impacts cluster analysis in two ways. Firstly, it makes ensemble approaches even more vital due to each technique measuring a different aspect of the data mining results [43]. Secondly, it makes it difficult to identify a direct loss of information when going from an original clustering to an anonymized clustering (i.e. the clustering result from the anonymized data). Therefore, "cluster information loss" is often defined as the difference in results from clustering evaluations when applied to the original clustering, and then separately applied to the anonymized clustering.

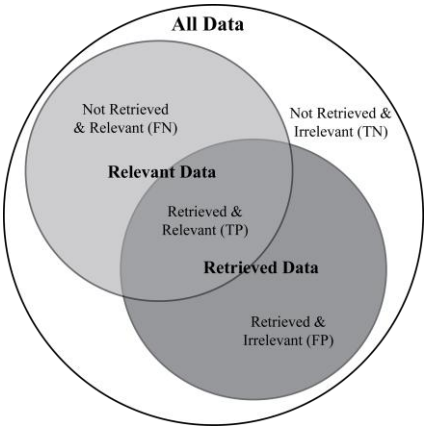

Fig. 2. A Venn diagram of possible outcomes when filtering records — True positives, false positives, true negatives, and false negatives.

Common metrics[20] used and included in ensembles are: Rand index [44], F-measure [45], Fowlkes-Mallows index [46], Davies-Bouldin index [47], and Silhouette [48]. Many of them rely on the concepts described by Fig. 2 – type I errors (false positives) and type II errors (false negatives). When executing a query[21] or filtering records based on a logic rule[22], false positives (FP) refer to records that were retrieved but shouldn't have been. False negatives (FN) refer to records that were not retrieved, but should have been. True positives (TP) and true negatives (TN) are the opposite: the records that were appropriately handled.

In order to apply these principles to a clustering scenario, a reference point is required that validates the obtained results (i.e. retrieved data) as either relevant or irrelevant. Here, "retrieved" is defined as "the records present in the cluster being assessed" and "relevant" is defined as "the records correctly belonging to that cluster". A class attribute can serve as a reference point to evaluate the clustering results. Typically, the class attribute is removed from the dataset prior to the application of a clustering algorithm. Once the clustering is complete, the class values are reassigned to the records. The most common class value in a cluster is used to define records as either relevant or irrelevant: if a record has any class value other than the majority value, it is irrelevant [39], [43]. A reference point is known as "external information", since in real life scenarios a clustering algorithm is typically applied to datasets that do not have a natural class attribute. Therefore, clustering information metrics that use external information are known as external metrics [5].

One such external metric is the **Rand index** (RI) [44]. It simply measures the fraction of correctly clustered records:

$$RI = \frac{TP + TN}{TP + FP + FN + TN}. \quad (1)$$

Unfortunately it treats false positives and false negatives as being equally undesirable, which is sometimes not the case. For example, a security agency would much rather deal with the inconvenience of a false positive than the security breach caused by a false negative.

F-measure [45] provides an easy solution to the weakness in RI, and is one of the most common clustering evaluation tool used by the PPDP and PPDM communities [36], [43]. It uses two expansions of the concepts described in Fig. 2: precision and recall. Precision measures the fraction of retrieved results that are relevant compared to irrelevant:

$$P = \frac{TP}{TP + FP}. \quad (2)$$

Recall measures the fraction of relevant results that were successfully retrieved:

$$R = \frac{TP}{TP + FN}. \quad (3)$$

Using these concepts, F-measure can be defined as:

$$F = \frac{2 \times P \times R}{P + R} \quad (4)$$

and thus it acts as a weighted average (harmonic mean) of the precision and recall. When treating false positives and false negatives differently, an expansion of the formula is used:

$$F_w = \frac{(w^2 + 1) \times P \times R}{(w^2 \times P) + R} \quad (5)$$

[20] A "metric" is a stricter variation of a measure. It acts as a distance function and satisfies four conditions: non-negativity, the identity of indiscernibles, symmetry, and the triangle inequality. Measures that satisfy these conditions can therefore be subject to more rigorous mathematical manipulation.

[21] A "query" is a list of requirements that records must meet in order to be retrieved, thus filtering out unnecessary data.

[22] A "logic rule" is a formal description of a pattern found in data. It takes the same form as a query, but is usually the result of following a chain of splitting points from a root to a leaf in a decision tree.

where $0 \le w \le 1$. When $w = 0$, $F_0 = P$, and recall has no impact on the F-measure result. F-measure is sometimes called "F_1 score", referring to the common case of $w = 1$. When P and R are expanded, the formula can be written as:

$$F_w = \frac{(w^2 + 1) \times TP}{(w^2 + 1) \times TP + w^2 \times FN + FP}. \quad (6)$$

Thus $w = 0$ equates to false negatives holding no weight.

An alternative to F-measure is the **Fowlkes-Mallows index** (FMI) [46]. While F-measure is the harmonic mean of precision and recall, FMI is the geometric mean and is defined as:

$$FMI = \sqrt{P \times R}. \quad (7)$$

There also exist a number of internal metrics that evaluate cluster quality without requiring a reference point. These metrics generally evaluate the results based on how compact each cluster is and how separated the clusters are from each other. One such metric is the **Davies-Bouldin index** (DBI) [47]. It defines a good clustering result as having low intra-cluster distances (i.e. high compactness) and high inter-cluster distances (i.e. high separation):

$$DBI = \frac{1}{n} \sum_{i=1}^{n} \max_{i \ne j} \frac{d_i + d_j}{D(c_i, c_j)} \quad (8)$$

where C_i and C_j are two different clusters out of n clusters, c_x is the centroid of cluster C_x, d_x is the average distance of records $r_i \in C_x; \forall i$ to c_x, and $D(c_x, c_y)$ is the distance between c_x and c_y.

Silhouette [48] measures how much more appropriate a record's cluster is compared to its second-most appropriate cluster:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (9)$$

$$S = \frac{\sum s(i)}{n} \quad (10)$$

where $a(i)$ is the average dissimilarity (usually measured via Euclidean distance[23]) of record r_i from each other record in the same cluster C_x, and $b(i)$ is the lowest average dissimilarity of r_i from all the records of one other cluster $C_y; \forall y \ne x$. Therefore $s(i)$ represents the "appropriateness" of record r_i's chosen cluster, and S is how appropriately all the records have been clustered. $s(i)$ ranges from -1 to 1, with 0 meaning that record *i* is on the border of two clusters, and a negative value meaning that *i* might be better off in its neighboring cluster. By comparing the result of these techniques before and after anonymization, one can make a more informed judgment on whether the clustering information has been preserved. An advantage of RI, F-measure and FMI is that they are simply reinterpretations of the same analysis, and so no extra computational time would be required if all three were to be calculated. DBI and Silhouette are clearly more computationally complex, but are arguably more explanatory and do not require reference points.

B. Classification

One of the main purposes of classification is to predict the value of a certain attribute for future records, where the values are not known. This is usually done with a decision tree, where each node in the tree filters the records it receives into two or more distinct (mutually exclusive) partitions[24] based on their value for an attribute. These partitions can then be split into more partitions until a termination requirement is met. In order to select a node's attribute, each attribute is tested to see how well it can filter records into distinct partitions, with each partition being as pure[25] as possible in respect to the attribute being predicted [49]-[51]. When dealing with decision trees, the filtering process is known as splitting, and a variety of algorithms exist for calculating the optimal "splitting point", known as splitting criteria. Examples include Information Gain [51], Gini Index [52] and Gain Ratio [53]. An example decision tree is provided in Fig. 3.

The most common technique for measuring the information quality of a classifier is **prediction accuracy** [7], [8], [10], [18], [20], [37], [38], [54], which measures the rate at which future records have their class values correctly predicted by a classifier. This is done by hiding the class value of some records not used when building the classifier, and seeing if the class values are correctly predicted.

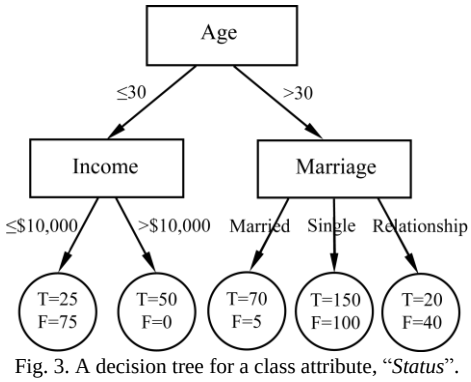

Fig. 3. A decision tree for a class attribute, "*Status*".

One way of presenting the results is to invert the accuracy so it represents prediction error, and compare this to the baseline error [10], [38], [54]. Baseline error is the prediction error of the classifier when no anonymization has occurred. Another type of error is also sometimes used: worst error [10] or removal error [54]. This error can be defined in any way which represents a relevant worst-case scenario. For example, when every attribute in the QID is suppressed (or generalized to the root of the taxonomy tree, e.g. "*anything*")

[23] "Euclidean distance" is the "ordinary" distance between two points in multi-dimensional space, as defined by the Pythagorean formula.

[24] A "partition" is a subset of records in a dataset.

[10]; or when every sensitive attribute[26] is removed from the dataset before a classification algorithm is applied [54].

These additional measures allow for further knowledge to be gained about the data mining information quality. As an example: the difference between the baseline error and worst (removal) error provides insight into the importance of the QID (sensitive) attributes in the classifier. A small difference would imply that the attributes don't influence the ability of the dataset to predict future cases. Perhaps some of the attributes can even be completely removed before publication if their utility (usefulness) doesn't warrant the privacy risk.

Unfortunately, prediction accuracy has some weaknesses. By simply measuring the percentage of records that have their class value correctly predicted, it fails to take into account any changes to the logic rules (patterns) or structure of a decision tree [7]. Sometimes an anonymized dataset can differ from the original dataset enough to result in significant structural differences between the trees obtained from the datasets, even if both trees have similar prediction accuracies [6], [7]. If the original patterns discovered through classification are weakened or destroyed by the anonymization process, but other – potentially misleading or artificial [4] – patterns are discovered, it's easily possible for the prediction accuracy to stay high, or even surpass the accuracy of the original classifier.

Some argue prediction accuracy is all that matters for a classifier, and if anonymization causes new patterns to be discovered and increases prediction accuracy then so much the better [2], [37], [38], [54]. However this represents a data mining decision made by the anonymization expert, and prevents data miners from exploring that possibility after the anonymized dataset is published. This is especially dangerous when considering that the alternate patterns may be artificial (fake), and not discovered in the original data because they do not exist [4].

Empirically this is supported by other measures sometimes disagreeing with the results of prediction accuracy [6]-[8], [18]. One of the strongest advantages of classification trees is that they provide humanly-readable patterns that can then be acted on or investigated. In other words, the logic rules themselves are valuable to a data miner, not just the predictive power of the classifier as a whole. Relying on prediction accuracy alone unnecessarily narrows the utility of the classifier. Unfortunately many do solely rely on prediction accuracy when measuring data mining information loss [10], [37], [38], [54]. More research is required to resolve these issues.

Another classification quality measure is **information-gain-to-privacy-loss ratio** (IGPL) [37], [38], [54]. This measure differs from previous measures in that it is considered a trade-off measure, or a search measure. A search measure is actively used during the anonymization process to guide it in sacrificing as little information as possible while gaining as much privacy/security as possible [2], [55].

In this instance of a search measure, IGPL is used as the filtering algorithm during decision tree construction, replacing the usual measures such as Gain Ratio. Unlike an ordinary decision tree, the authors [37] propose generalizing all attributes to their most vague state, and then using the decision tree to choose which attributes to make more specific (in other words, specialized – the opposite of generalization). The concept is surprisingly simple and effective: rather than defining a node split by the best information gain (IG), it can be defined by the best trade-off between information gain and privacy loss (IG/PL). Here, IG and PL can be defined as any information measure and any privacy measure that the anonymization expert feels appropriate. If an attribute is calculated to provide a large increase in IG and a low increase in PL, it is likely to be chosen as the filter for a node, and thus specialized. For example in Fig. 3, if "*Marriage*" was not chosen as a splitting point at any point in a tree made using IGPL, all records would simply have "*Unknown*" as their "*Marriage*" value.

A possible extension that remains unanswered is for IGPL to handle several different attributes being used as class attributes in order to maintain additional logic rules [1], [4]. We suggest an exploration into decision forests[27], using good alternate IGPL trade-offs in each tree, and perhaps weighting attributes.

IV. Conclusions

In the pursuit of quality, it is important that we can define and measure what "quality" really is. Due to anonymization requirements, measures have been developed over the years to measure changes in the quality of a dataset. It has been shown that different scenarios require different approaches to information quality evaluation, and it is important that an anonymization expert is confident in selecting the most applicable measures. Here, we have discussed a variety of methods for quality analysis, and the scenarios they are best suited for. Ensemble approaches are also important to keep in mind: in many cases, a single measure cannot robustly evaluate the utility of a dataset in all scenarios.

Another possibility is for information measures to be used as search measures. By guiding the anonymization process towards the optimal solution (as defined by the measure), it removes the need for iterative testing with estimated parameters. This can be useful for anonymization techniques or information measures with high computational cost, where heavy testing might be infeasible. Some algorithms are more computationally complex than others; but in many cases this should not be considered a deciding factor. Often the goal is to create one anonymized version of a dataset for a single release, and this is rarely time-sensitive. Here, maximizing data quality is far more important than computational cost.

Comparing the effect of different types of anonymization techniques is still an open question. Perhaps the only solution is to judge the techniques based on their principles, rather than their empirical results. For example, the faithfulness of

[25] The "purity" of a collection of records refers to the percentage of records that agree (have the same value) about the class attribute.

[26] A "sensitive attribute" is one deemed to be a risk to individuals' privacy.

[27] A "decision forest" is a collection of unique decision trees, where the predictions of each tree are combined to create a weighted average when predicting the class attribute.

generalization might outweigh the benefits of randomization, such as not relying on user-defined attribute taxonomies.

The common trade-off in PPDP and PPDM is privacy vs. quality, and measuring quality has received much less attention. Further research would prove beneficial for not only identifying effective anonymization techniques, but also for better understanding the factors that affect dataset quality and data mining quality.

REFERENCES

[1] D. Kifer and J. Gehrke, "Injecting utility into anonymized datasets," in *Proc. the 2006 ACM SIGMOD International Conference on Management of Data*, New York, New York, USA: ACM, 2006, pp. 217-228.

[2] B. Fung, K. Wang, A. W.-C. Fu, and P. S. Yu, *Introduction to Privacy-Preserving Data Publishing: Concepts and Techniques*, CRC Press, 2010.

[3] L. Guo, X. Ying, and X. Wu, "On attribute disclosure in randomization based privacy preserving data publishing," in *Proc 2010 IEEE International Conference on Data Mining Workshops (ICDMW)*, Dec. 2010, pp. 466-473.

[4] E. Bertino, D. Lin, and W. Jiang, "A survey of quantification of privacy preserving data mining algorithms," *Privacy-Preserving Data Mining*, pp. 1-20, 2008.

[5] P. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*, 1st ed. Boston, USA: Addison-Wesley, 2005.

[6] T. Lim, W. Loh, and Y. Shih, "A comparison of prediction accuracy, complexity and training time of thirty-three old and new classification algorithms," *Machine Learning*, vol. 40, no. 3, pp. 203-228, 2000.

[7] M. Z. Islam, P. Barnaghi, and L. Brankovic, "Measuring data quality: Predictive accuracy vs. similarity of decision trees," in *Proc. the 6th International Conference on Computer & Information Technology*, Dhaka, Bangladesh, 2003, vol. 2, pp. 457-462.

[8] M. Z. Islam and L. Brankovic, "Privacy preserving data mining: A noise addition framework using a novel clustering technique," *Knowledge-Based Systems*, vol. 24, no. 8, pp. 1214-1223, 2011.

[9] M. G. Rahman and M. Z. Islam, "Missing value imputation using decision trees and decision forests by splitting and merging records: Two novel techniques," *Knowledge-Based Systems*, vol. 53, pp. 51-65, Sep. 2013.

[10] K. Wang, P. Yu, and S. Chakraborty, "Bottom-up generalization: A data mining solution to privacy protection," in *Proc. Fourth IEEE International Conference on Data Mining*, 2004, pp. 249-256.

[11] A. Skowron and R. Cecylia, "The discernibility matrices and functions in information systems," in *Intelligent Decision Support*, R. Slowinski, Ed. Dordrecht: Springer Netherlands, 1992, pp. 331-362.

[12] R. Bayardo and R. Agrawal, "Data privacy through optimal kanonymization," in *Proc. 21st IEEE International Conference on Data Engineering*, 2005, pp. 217-228.

[13] P. Samarati and L. Sweeney, "Protecting privacy when disclosing information: K-anonymity and its enforcement through generalization and suppression," *SRI International, Tech. Rep.*, 1998.

[14] L. Sweeney, "K-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557-570, 2002.

[15] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Workload-aware anonymization," in *Proc. the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '06*, 2006, p. 277.

[16] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 1, no. 1, p. 3, 2007.

[17] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, and A. Fu, "Utility-based anonymization using local recoding," in *Proc. the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006, pp. 785-790.

[18] M. Nergiz and C. Clifton, "Thoughts on k-anonymization," *Data & Knowledge Engineering*, vol. 63, no. 3, pp. 622-645, 2007.

[19] H. Giggins, "VICUS - A Noise Addition Technique for Categorical Data," in *Proc. 10th Australasian Data Mining Conference (AusDM)*, Y. Zhao, J. Li, Paul Kennedy, and P. Christen, Eds. Sydney, vol. 134, 2012.

[20] C. Aggarwal and P. Yu, "A condensation approach to privacy preserving data mining," *Advances in Database Technology-EDBT* 2004, vol. 2992, pp. 183-199, 2004.

[21] M. Z. Islam, "Privacy preservation in data mining through noise addition," Ph.D. dissertation, University of Newcastle, Newcastle, 2007.

[22] K. Liu, H. Kargupta, and J. Ryan, "Random projection-based multiplicative data perturbation for privacy preserving distributed data mining," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 1, pp. 92-106, 2006.

[23] A. Evfimievski, "Randomization in privacy preserving data mining," *ACM Sigkdd Explorations Newsletter*, vol. 4, no. 2, pp. 43-48, Dec. 2002.

[24] H. Kargupta, S. Datta, and Q. Wang, "On the privacy preserving properties of random data perturbation techniques," in *Proc. Third IEEE International Conference on Data Mining*, pp. 99-106, 2003.

[25] J. Domingo-Ferrer, "Comparing SDC methods for microdata on the basis of information loss and disclosure risk," *Pre-proceedings of ETKNTTS*, 2001.

[26] J. Hair, R. Anderson, R. Tatham, and W. Black, *Multivariate Data Analysis*, 5th ed., Prentice Hall International, 1998.

[27] L. Sweeney, "Achieving k-anonymity privacy protection using generalization and suppression," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 5, pp. 571-588, 2002.

[28] L. Sweeney, "Guaranteeing anonymity when sharing medical data, the Datafly System," in *Proc. the AMIA Annual Fall Symposium*, 1997.

[29] P. Samarati, "Protecting respondents identities in microdata release," *IEEE Transactions on Knowledge and Data Engineering*, vol. 13, no. 6, pp. 1010-1027, 2001.

[30] K. Wang and B. Fung, "Anonymizing sequential releases," in *Proc. the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 2006, pp. 414-423.

[31] X. Xiao and Y. Tao, "Personalized privacy preservation," in *Proc. the 2006 ACM SIGMOD International Conference on Management of Data*, ACM, 2006, pp. 229-240.

[32] V. Iyengar, "Transforming data to satisfy privacy constraints," in *Proc. the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 2002, pp. 279-288.

[33] K. LeFevre, D. DeWitt, and R. Ramakrishnan, "Incognito: Efficient full-domain k-anonymity," in *Proc. the 2005 ACM SIGMOD International Conference on Management of Data*, ACM, 2005, pp. 49-60.

[34] S. Giessing, "Survey on methods for tabular data protection in ARGUS," *Privacy in Statistical Databases*, pp. 1-13, 2004.

[35] X. Xiao and Y. Tao, "Anatomy: Simple and effective privacy preservation," in *Proc. the 32nd International Conference on Very Large Data Bases*, 2006.

[36] B. Fung, K. Wang, L. Wang, and M. Debbabi, "A framework for privacy-preserving cluster analysis," in *Proc. 2008 IEEE International Conference on Intelligence and Security Informatics*, 2008, pp. 46-51.

[37] B. Fung, K. Wang, and P. Yu, "Top-down specialization for information and privacy preservation," in *Proc. 21st IEEE International Conference on Data Engineering*, 2005, pp. 205-216.

[38] B. Fung, K. Wang, and P. Yu, "Anonymizing classification data for privacy preservation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 5, pp. 711-725, 2007.

[39] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, 2006.

[40] R. Cordeiro de Amorim and B. Mirkin, "Minkowski metric, feature weighting and anomalous cluster initializing in K-Means clustering," *Pattern Recognition*, vol. 45, no. 3, pp. 1061-1075, 2012.

[41] M. Trottini, "A decision-theoretic approach to data disclosure problems," *Research in Official Statistics*, pp. 1-16, 2001.

[42] M. Trottini, "Decision models for data disclosure limitation," Ph.D. dissertation, Carnegie Mellon University, 2003.

[43] M. Rahman and M. Islam, "CRUDAW: A novel fuzzy technique for clustering records following user defined attribute weights," in *Proc. Tenth Australasian Data Mining Conference, AusDM12*, Sydney, Australia, 2012, vol. 134.

[44] W. Rand, "Objective Criteria for the Evaluation of Clustering Methods," *Journal of the American Statistical Association*, vol. 66, no. 336, 1971.

[45] C. van Rijsbergen, *Information Retrieval*, Butterworth, 1979.

[46] E. Fowlkes and C. Mallows, "A method for comparing two hierarchical clusterings," *Journal of American Statistical Association*, vol. 78, no. 383, pp. 553-569, 1983.

[47] D. Davies and D. Bouldin, "A cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 2, pp. 224-227, 1979.

[48] P. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53-65, 1987.
[49] M. Z. Islam, "Explore: A novel decision tree classification algorithm," *Data Security and Security Data*, vol. 6121, 2012.
[50] M. Z. Islam and H. Giggins, "Knowledge discovery through SysFor – A systematically developed forest of multiple decision trees," in *Proc. the Ninth Australasian Data Mining Conference*, Ballarat, Australia, 2011, vol. 121, pp. 195-204.
[51] J. R. Quinlan, *C4.5: Programs for Machine Learning*, 1st ed. Morgan Kaufmann, 1993.
[52] K. A. Grajski, L. Breiman, G. V. D. Prisco, and W. J. Freeman, "Classification of EEG spatial patterns with a tree-structured methodology: CART," *IEEE Transactions on Biomedical Engineering*, vol. 12, pp. 1076-1086, 1986.
[53] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81-106, 1986.
[54] K. Wang, B. Fung, and P. Yu, "Template-based privacy preservation in classification problems," in *Proc. Fifth IEEE International Conference on Data Mining*, 2005, p. 8.
[55] G. Duncan, S. Keller-McNulty, and S. Stokes, "Disclosure risk vs. data utility: The RU confidentiality map," *Chance*, 2001.

Sam Fletcher is a PhD student in the School of Computing and Mathematics, Charles Sturt University, Australia. He also received his bachelor's degree and First Class Honours in computer science at Charles Sturt University. His main research interests include data mining, privacy preservation, data quality, and information interestingness.

Md Zahidul Islam is a senior lecturer in computer science in the School of Computing and Mathematics, Charles Sturt University, Australia. He received his bachelor's degree in engineering from Rajshahi University of Engineering and Technology, Bangladesh, graduate diploma in information science from the University of New South Wales, Australia and PhD in computer science from the University of Newcastle, Australia. His main research interests include data pre-processing and cleansing, various data mining algorithms, applications of data mining techniques, privacy issues related to data mining, and privacy preserving data mining.
.