Utility Enhancement for Privacy Preserving Health Data Publishing

Lengdong Wu, Hua He, and Osmar R. Zaïane

Department of Computing Science
University of Alberta
Edmonton, Alberta, Canada
{lengdong, hhe, zaiane@ualberta.ca}

Abstract. In the medical field, we are amassing phenomenal amounts of data. This data is imperative in discovering patterns and trends to help improve healthcare. Yet the researchers cannot rejoice as the data cannot be easily shared, because health data custodians have the understandable ethical and legal responsibility to maintain the privacy of individuals. Many techniques of anonymization have been proposed to provide means of publishing data for research purposes without jeopardizing privacy. However, as flaws are discovered in these techniques, other more stringent methods are proposed. The strictness of the techniques is putting in question the utility of the data after severe anonymization. In this paper, we investigate several rigorous anonymization techniques with classification to evaluate the utility loss, and propose a framework to enhance the utility of anonymized data.

Keywords: Data Publishing, Privacy Preservation, Anonymization, SVM

1 Introduction

Health research is central to the advancement of health care, which imperatively requires access to health data. However, health records are intrinsically identifiable, and control over the use of the data is necessary. When trying to publish health data, custodians of such data inevitable encounter hurdles relative to privacy[10]. To address such issues, a Privacy Rule was established in the United States which constitutes a comprehensive Federal protection for individuals' medical records and is known as *Health Insurance Portability and Accountability Act* (HIPAA)[12, 13]. The legislation regulates health care groups, organizations or businesses, on how to use and disclose privacy data.

In recent years, several significant privacy preserving techniques have been proposed to protect individual's privacy when sharing the information. These techniques are progressively stricter as vulnerabilities in these algorithms are discovered. However, while such strict techniques prevent identification, they can significantly hinder the utility of the data for research purposes. In this paper we investigate several rigorous anonymization techniques with novel criterions based on a classification technique to evaluate the data utility. The remainder of the paper is organized as follows: in subsequent sections, we briefly introduce k-anonymity[1, 4], l-diversity [3] and t-closeness [7] and their related limitations with tangible examples. We then present our utility evaluation methods on the anonymized data based on SVM, and the results of utility loss are analyzed. Given the sigificant utility loss, a privacy preservation with utility enhancement supervision framework is proposed. We present the implementation of the framework and algorithm with comparison experiment before concluding.

2 Privacy Preservation Technique

A typical health data table includes the basic personal information as well as their sensitive information such as diagnostic and treatment history records. All these attributes can be categorized into three classes [2]:

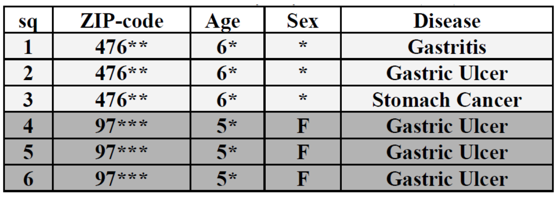
- *Identifier attributes*: a minimal collection of attributes that can explicitly identify individual records.
- *Sensitive attributes*: considered to be private.
- *Quasi-identifier (QI) attributes*: a minimal collection of attributes that can be linked with external information to re-identify individual records with high probability.

According to the HIPPA regulation, the removal of all identifier attributes is necessary. However, relinking attack [4, 5] is a notorious attack on the de-identified tables by joining two tables having common quasi-identifier attributes. For example, based on the statistics, approximately 97% of 54,805 voters in the Cambridge, U.S. can be uniquely identified on the basis of full combination of the zip-code, gender and birthday attributes; 87% can be identified with the combination of only 5-digit ZIP-code, gender and birthday; and another 69% uniquely with the ZIP-code and birthday [2]. This result reveals a serious privacy preservation problem and shows a high possibility of re-identifying the de-identified table under the re-linking attack.

2.1 *k*-Anonymity Beyond De-identification

***k*-Anonymity Principle.** The simple identifier removal process cannot guarantee the anonymity of the published data due to its potential leakage on QI attributes. The k-anonymity technique is designed to avoid re-linking attacks through generalizing the QI attribute values. For each QI attribute, a tree-structured domain generalization hierarchy is maintained, in which the node in higher levels contains more generalized information. Given this hierarchy, the specific values in the original table can be replaced by the more general values in higher level nodes of the hierarchy. Records with the same generalized value are gathered into an equivalence class[2], thus the re-linking attack cannot distinguish a certain individual from other records in the certain equivalence class. A table satisfies k-anonymity principle if at least k indistinct records exist in each equivalence class. For instance, Table 1 satisfies 3-anonymity.

Attacks on *k*-Anonymity. It is quite common for k-anonymity to generate equivalence classes with same values of sensitive attributes, especially when certain sensitive attributes have high frequent values. For example, in Table 1, an adversary can easily know that individuals in the second equivalence class suffer from Gastric Ulcer. Although the equivalence class decreases the possibility of identifying individual, the sensitive attributes can provide auxiliary clew, which can be utilized by homogeneity attacks[3]. Background attack[14] uses some background knowledge to obtain privacy information on the k-anonymity tables. Again, in Table 1, suppose an adversary knows that an individual in the first equivalence class has a certain cancer, this fact as background knowledge can assure the adversary that this individual has Stomach Cancer.



sq	ZIP-code	Age	Sex	Disease
1	476**	6*	*	Gastritis
2	476**	6*	*	Gastric Ulcer
3	476**	6*	*	Stomach Cancer
4	97***	5*	F	Gastric Ulcer
5	97***	5*	F	Gastric Ulcer
6	97***	5*	F	Gastric Ulcer

Table 1. *k*-Anonymity Health Table (k=3)

2.2 *l*-Diversity Beyond *k*-Anonymity

***l*-Diversity Principle.** To deal with the defects of *k*-anonymity, *l*-diversity requires that the sensitive attribute values in each equivalence class should be as diverse as possible, requiring at least l well-represented sensitive attribute values. The requirement of l well-represented values of sensitive attributes adds an extra protection layer over *k*-anonymity. When a table satisfies *l*-diversity, the adversary who breaches the *k*-anonymity, still needs to exclude the (l-1) possible sensitive values. The larger the parameter l, the more protection it provides.

Attacks on *l*-Diversity. However, the requirement of *l*-diversity on well-represented values cannot really ensure the real diversity for sensitive attributes. For example in Table 1, "Gastric Ulcer", "Gastritis" and "Stomach Cancer" are all stomach related, then the adversary could know that the individuals in the first equivalence class must have a problem with the stomach. Similarity attack [6] and skewness attack [7] are two typical attacks on such semantic leaks in sensitive values. The breach will be serious when the number of sensitive attribute categories is small.

2.3 *t*-Closeness Beyond *l*-Diversity

Rather than simply making sensitive attribute values numerically diverse, *t*-closeness [7] makes the sensitive values semantically diverse. The *t*-closeness requires the distribution of sensitive values in each equivalence class close to the overall distribution of the whole table.

3 Utility Loss Evaluation

3.1 Utility Loss Measures

The three important privacy preservation processes, *k*-anonymity, *l*-diversity and *t*-closeness are effective in protecting data privacy. However, there is a risk that they lower the utility of the data in the context of health research, such as building classifiers for automated diagnostic, treatment recommendation or other relevant applications requiring machine learning. Therefore, the balance between the data utility for scientific research and the privacy preservation for health data is of paramount importance; at least, reducing the loss as much as possible while keeping the same level of privacy, is imperative.

To capture data utility, some criteria measure the utility loss that is incurred by generalization based on generalization hierarchies, such as Discernability Measure (DM) [1], Utility Measure (UM) [17], Relative Error (RE) [18], Normalized Certainty Penalty (NCP) [16] etc. DM and RE is calculated based on the number of generalized group and suppressed group that overlap with the original data. NCP and UM are expressed as the weighted average of the information loss, which are penalized based on the number of ascendants in the hierarchy. Some recently proposed measures, such as multiple level mining loss [15], express utility based on how well anonymized data supports frequent itemset mining. However, all these measures are essentially evaluating the information loss of generalized items via certain penalization function based on the number of ascendants achieved in the hierarchy. A measure that can be used in the absence of hierarchies and captures the utility loss incurred by generalization is more preferred by practical application scenarios.

Machine learning applications can utilize the analysis and intelligent interpretation of large data in order to provide actionable knowledge based on the data for human decision support or automatic decision making. Support Vector Machine (SVM) is one of the effective machine learning algorithm. The standard SVM takes a set of input, each of which belonging to one of several categories; then builds a model of hyperplane separating the data space through the learning process to predict whether a new test example falls into one category or another. In this section, we are particularly interested in evaluating and discussing the utility loss induced by privacy protection via the use of Support Vector Machine (SVM), and examine the utility value through the measure of accuracy after anonymization.

3.2 Datasets and Experimental Setup

We use two census-based datasets, the Adult dataset, which is originally from the US census bureau database, and the IPUMS dataset from the historical census project at the University of Minnesota. Both datasets, available from the UCI Machine Learning repository[1], have been extensively used by recent privacy preservation studies [3, 7]. In the Adult dataset, we choose attribute set including *age*, *workclass*, *education*, *gender*, *race*, *marriage*, and *country* as QI attributes, and use the salary class as the sensitive attribute. In the IPUMS dataset, QI attribute set includes *sex*, *relationship*, *race*, *birthplace*, *children number*, *education attainment*, *weeks worked last year*, and use the *wage* class as the sensitive attribute. We remove all records with missing values. Our experiments use the following parameters: $k = 4$ for k-anonymity, $k = 4$ and $l = 2$ for l-diversity, $k = 4$ and $t = 0.2$ for t-closeness. Those settings are commonly applied in practice [3, 5, 7], which are regarded to be not too strict to make the output data completely unusable.

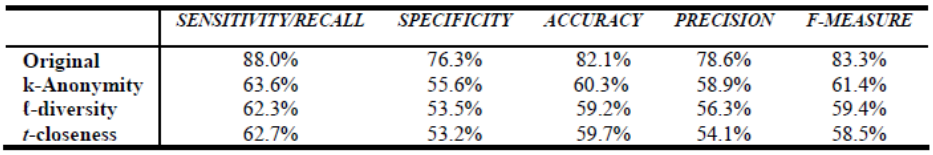
We use the LibSVM toolkit[8] to run the SVM classification algorithm, and apply the same SVM parameters for all experiments. The datasets are divided into the training and test sets randomly in three fold cross validation sets: one third of each set is used as test data while the other two thirds are used for training. In our first experiment, we use SVM on the original dataset, so that all information in the QI attributes can be fully utilized for SVM classification. We then apply SVM on the anonymized data by k-anonymity, l-diversity and t-closeness separately. By comparing the classification results, we can evaluate to what degree the anonymized data could lose utility and we examine its loss value.

3.3 Utility Loss Results

Table 2 presents the comparisons of the accuracies of correctly classified records by SVM on the Adult dataset. Significant drops, 25% in sensitivity and 21% in specificity, can be observed due to the inadvertent obfuscation of pertinent information necessary for building the classification model. In Table 2, one might expect the classification results to have lower accuracies for l-diversity and t-closeness compared to k-anonymity; however, the results are quite similar. This is due to the fact that k-anonymity already produces significant information loss, and l-diversity in each equivalent class is already established. Table 3 shows the comparison based on the IPUMS data, where there is a noticeable drop when using l-diversity and t-closeness as compared to k-anonymity. This shows an additional utility loss beyond what k-anonymity can have already done.

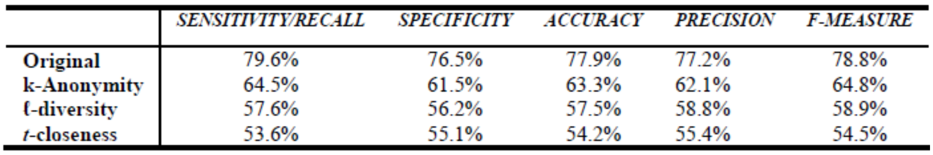
Based on the experimental results on these datasets, we can conclude that the utility value of data, after the anonymization by k-anonymity, l-diversity and t-closeness, is significantly jeopardized due to the strictness of those privacy preservation mechanisms.

[1] http://archive.ics.uci.edu/ml/machine-learning-databases/



	SENSITIVITY/RECALL	*SPECIFICITY*	*ACCURACY*	*PRECISION*	*F-MEASURE*
Original	88.0%	76.3%	82.1%	78.6%	83.3%
k-Anonymity	63.6%	55.6%	60.3%	58.9%	61.4%
ℓ-diversity	62.3%	53.5%	59.2%	56.3%	59.4%
***t*-closeness**	62.7%	53.2%	59.7%	54.1%	58.5%

Table 2. Experiment Results with Adult Dataset



	SENSITIVITY/RECALL	*SPECIFICITY*	*ACCURACY*	*PRECISION*	*F-MEASURE*
Original	79.6%	76.5%	77.9%	77.2%	78.8%
k-Anonymity	64.5%	61.5%	63.3%	62.1%	64.8%
ℓ-diversity	57.6%	56.2%	57.5%	58.8%	58.9%
***t*-closeness**	53.6%	55.1%	54.2%	55.4%	54.5%

Table 3. Experiment Results with IPUMS Data

4 Privacy Preservation with Utility Supervision

4.1 Utility Enhancement Supervision Framework

To minimize the utility loss of these privacy preserving techniques, partition-based and cluster-based anonymization algorithms have been proposed recently. The partition-based anonymization treats a record projected over QI attributes as a multi-dimensional point. A subspace that contains at least k points forms a k-anonymous group [18]. The main idea of clustering-based anonymization is to create clusters containing at least k records in each cluster separately [16]. Fung et al. [19] presented an effective top-down approach by introducing multiple virtual identifiers for utilizing information and privacy-guided specialization. However, the partitioned-based anonymization selects the attribute with the largest domain for efficiency and top-down specialization chooses the attribute with best pre-defined scoring ranking. These genetic evolution and top-down generalization algorithms do not produce any progressive attribute selection process which determines a desired balance of privacy and accuracy.

We introduce the utility enhancement supervision in the attribute selection process. The insight of our proposal is based on the acknowledgement that the anonymization process unquestionably damages the potential data correlation between the QI attributes and the sensitive attributes; and the higher generalization hierarchy is achieved the more correlation is lost. Since any prior knowledge is unknown about the class related features for QI attributes, there probably exist, among the numerous QI attributes, some that have poor correlation or no correlation with sensitive attributes. These superfluous QI attributes are definitely ideal for generalization without losing any utility. More generally, QI attributes that are less correlated with the sensitive attribute are better candidates for generalization of anonymity than others. The less the attributes with strong correlation are generalized, the more utility will be preserved for anonymization. Hence, the utility enhancement supervision is established to produce such an order of QI attribute candidates for generalization.

Figure 1 illustrates our framework of privacy preservation with utility enhancement supervision. The process is divided into four stages:

- Stage 1. Sample data extraction. De-identified dataset is submitted to D and sample data D_0 is randomly extracted from D for evaluation purpose.
- Stage 2. Anonymization candidates order. Given the randomly selected sample dataset D_0, SVM utility evaluation is applied to produce the partial order of correlation of QI attributes.
- Stage 3. Attribute generalization. Optimal attributes are chose based on the partial order to be generalized according to each own generalization hierarchy.
- Stage 4. The anonymized dataset D' is verified according to anonymity principles. If all equivalent classes satisfy all requirements of the specified principle, D' is ready for publishing.

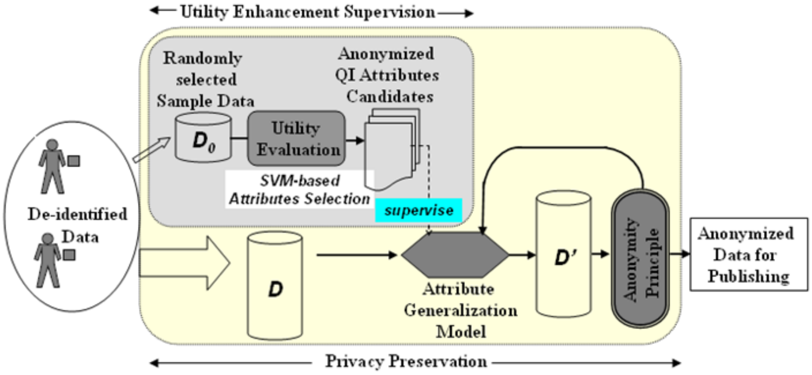

Utility Enhancement Supervision
Randomly selected Sample Data
Anonymized QI Attributes Candidates
D_0
Utility Evaluation
SVM-based Attributes Selection
supervise
De-identified Data
D
Attribute Generalization Model
D'
Anonymity Principle
Anonymized Data for Publishing
Privacy Preservation

Fig. 1. Privacy Preservation with Utility Supervision Framework

4.2 Privacy Preservation with Utility Supervision Algorithm

To produce an appropriate anonymization algorithm with utility supervision, we need to solve the following issues:

- define the standard for comparison which is essential for the partial order.
- devise an efficient algorithm to generate the partial order.
- select the optimal attribute candidates for generalization based on the utility order

For this purpose, we continue to adopt the utility evaluation based on SVM and the F-measure value for SVM-based classifier cross validation is used as the criteria for comparison. We use the notation $F(S)$ to indicate the F-measure value for cross validation with attributes set S.

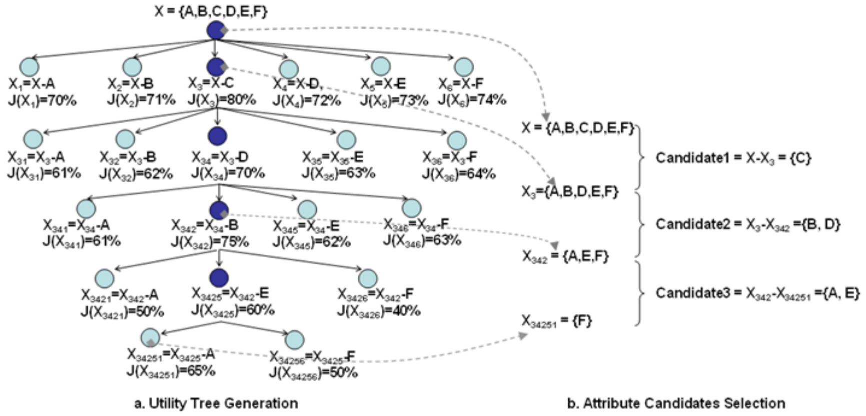
To generate the partial order of candidates, the simplest way is to compare all possible combinations. However, the number of combinations grows exponentially as the number of attributes increases, thus the brute-force solution might not always be practical. Thus we use sequential backward selection (SBS), to achieve affordable search performance. We assume the original QI attributes set is X_s. Each time one attribute ξ is removed, and SVM classifier is done based on attributes $(X - \xi)$ obtaining $F(X - \xi)$. The attribute $\hat{\xi}$ having the maximum F value implies that the left attributes $(X - \hat{\xi})$ can best preserve utility, thus $\hat{\xi}$ is removed. The removal procedure is repeated until the attribute set is empty with a utility tree established.

To extract the attribute candidates, we first find the maximum $F(X')$ value in the whole tree, then attributes existing in $(X - X')$ will be chosen for generalization. In the case that these first-batch candidate attributes are all generalized to their highest level in the generalization hierarchy and the anonymization constraints are still not satisfied, another batch of candidates need to be selected for further generalization. For this purpose, the maximum $F(X'')$ value is searched in the subtree whose root is X'. Attributes in $(X' - X'')$ will form a new group of attributes for generalization. The procedure of search, selection, generalization and check is executed repeatedly until a certain anonymization principle is achieved. Algorithm 1 demonstrates the details for the procedure.

For example, we assume there are six QI attributes $X = \{A, B, C, D, E, F\}$, as illustrated in Figure 2. After removing each attribute and executing a classifier cross validation, we find that $X_3 = X - C = \{A, B, D, E, F\}$ obtains the highest F value. Thus in the next round of tree building, we only start from X_3 rather than considering other sibling nodes. With the same manner for $X_{34} = \{A, B, E, F\}$, $X_{342} = \{A, E, F\}$, and $X_{3425} = \{A, F\}$, the utility tree can be established. To select attribute candidates, the maximum F value is achieved by $F(X_3)$ with attribute set $\{A, B, D, E, F\}$. Thus, QI attribute $\{C\}$ is first chosen to be generalized. In the subtree of X_3, X_{342} with attribute set $\{A, E, F\}$ has the highest F value. $\{B, D\}$ will be the candidates for generalization. Repeatedly, $\{A, E\}$ will be selected as next group of candidates.

4.3 Experiment Evaluation

The experiment is based on the same census-based datasets, the Adult dataset and the IPUMS dataset, and the same QI attributes set are chosen as introduced in Section 3.2. We compare the utility loss by global recoding and local recoding implementation [15, 16] and test with the common configurations for k, l and t as described in Section 3.2. The global recoding can be described by a group of functions $\phi_i : D_{X_i} \rightarrow D'$ for each attribute X_i of the Quasi-identifier. The anonymization is obtained by applying each ϕ_i to the values of X_i in each tuple of D. The global recoding generalizes the attribute for a group of records in the



a. Utility Tree Generation

b. Attribute Candidates Selection

Fig. 2. Example of Privacy Preservation with Utility Supervision Algorithm

table for efficiency. In our implementation, the data space is partitioned into a set of non-overlapping regions and the algorithm maps all records in a region to the same generalization hierarchy level. When checking whether anonymity principle is satisfied, a single SQL query can be established, for example, "SELECT *race*, *gender*, *age*, *count(*)* FROM *Table* GROUP BY *race*, *gender*, *age* HAVING *count(*)*$> k$". Alternatively, the local recoding is described by a function $\phi : D_{X_1} \times D_{X_2} \times ... \times D_{X_n} \rightarrow D'$, which recodes the domain of value vectors associated with the set of Quasi-identifier attributes. Under this definition, the anonymization is obtained by applying ϕ to the vector of Quasi-identifier values in each tuple of D. We implemented the local recoding by generalizing the Quasi-identifier attribute to a higher level only for the distinct individual record that does not achieve the anonymity constraints rather than all records. To check the satisfaction of anonymity constraints, we introduce an *equivalence class id*. The record satisfying the constraints is assigned with such a class id, indicating that the record belongs to the corresponding equivalence class after generalization. When each record in the table has a valid class id, the table is considered to be anonymized successfully.

Based on the algorithm 1, in the Adult Dataset, we obtained the attribute set partial order as: F(*age*, *workclass*, *education*, *country*)=84.4%, F(*workclass*, *education*)=78.2%. Thus, generalization is done firstly on attribute set {*race*, *marriage*, *gender*}, and after all these attributes have reached the highest level in the hierarchy, attribute set {*age*, *country*} is generalized. In the IPUMS dataset, attribute set partial order is calculated as: F(*sex*, *race*, *children number*, *education attainment*, *occupation*, *weeks worked last year*)=79.2%, F(*sex*, *children number*, *education attainment*, *occupation*, *weeks worked last year*)=73.4%, F(*education attainment*, *occupation*, *weeks worked last year*)=70.6%. Accordingly, attribute candidate sets are generalized based on the order: {*relationship*, *birthplace*}, {*race*}, {*sex*, *children number*}.

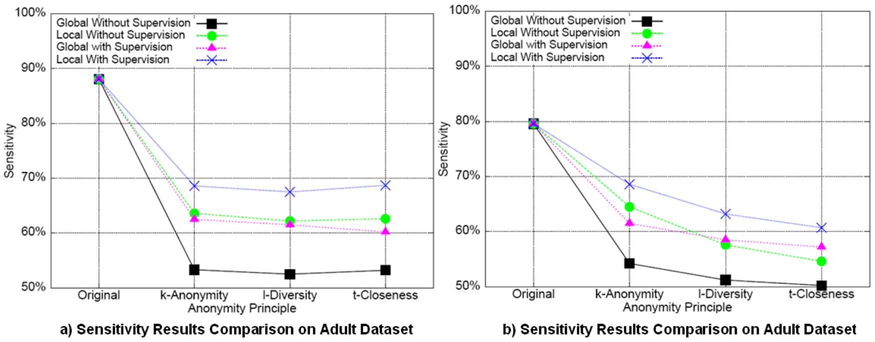


Fig. 3. Performance Comparison of Privacy Preservation with vs. without Supervision

Figure 3 shows that there is a significant increase in terms of sensitivity and specificity between anonymization with supervision and anonymization without supervision on both datasets. On the Adult dataset, we get an accuracy about 7% higher for k-anonymity and l-diversity principle, and 5% higher for t-closeness. Such significant rises are due to the deliberate retainment of pertinent attribute information necessary for building the classification utility model. Comparison on the IPUMS data shows the accuracy for the classifier can be improved even more when using l-diversity and t-closeness principle than with k-anonymity. This is because l-diversity and t-closeness, being stricter than k-anonymity, they necessarily require further generalization on additional attributes. Imposing restrictions or guidance on the attributes being generalized can reduce the risk that pertinent information contained by correlative attributes is jeopardized. Based on the experimental results on these datasets, we can conclude that our proposed privacy preservation algorithm with utility supervision can significantly increase the utility of privacy preservation mechanisms.

5 Conclusion

In this paper we examined the issue of utility of health data after the anonymization process and put forward the necessity of finding a trade-off between privacy protection and utility of data. We describe three important and recently proposed privacy preservation techniques, k-anonymity, l-diversity and t-closeness, and present the limitations of each technique. By using SVM to evaluate the utility loss, we show that the privacy preservation technique implementation we have at our disposal today can significantly jeopardize the data utility due to the obliteration of pertinent information. Protecting the privacy of patients is central. Using the wealth of health data we are collecting to improve healthcare, is also essential. To enhance the utility of the data we put forward the privacy preservation with utility enhancement supervision framework. With this framework, the anonymized data is able to preserve the data utitily as well as protect the privacy of sensitive information.

ALGORITHM 1: Anonymization with Utility Supervision

```
Input: Private de-identified Table, QI($\xi_1, ..., \xi_n$), Anonymity constraints, Domain
        generalization hierarchy $DGH_{\xi_i}, i \in [1, ..., n]$
Output: Anonymized Publishable Table containing a generalization over QI with
         respect to Anonymity principle
/* Step1. Generate utility tree of QI attributes based on SBS                */
initial selected attributes set $X_s \leftarrow$ QI($\xi_1, ..., \xi_n$) ;
initial root node $\leftarrow F(X_s)$;
repeat
    foreach $\xi_i \in X_s$ do remove one attribute from the $X_s$
        /* Use SVM-based classifier on randomly selected sample data for
           cross validation                                                   */
        $F_i \leftarrow F(X_s - \xi_i)$;
        $F(X_s)$.child node $\leftarrow F_i$;
    end
    /* Find such attribute $\xi_k$ that $F(X_s - \xi_k)$ is the maximum          */
    $F_k \leftarrow Max(F_i)$, $i \in [1, ...s]$;
    $X_s \leftarrow X_s - \xi_k$;
until $X_s = \perp$;
/* Step2. Search for candidates for generalization in utility tree            */
Initial root node $X_s \leftarrow X$ ;
repeat
    Search Tree($X_s$) $\rightarrow X'$ : $F(X')$ is maximum ;
    repeat
        /* Step3. Generalize attribute candidates                             */
        Select attribute set $< X_s - X' >$ for generalization ;
        Build hierarchy vector $< DGH_{\xi_i}$+1>, $\xi_i \in < X_s - X' >$ ;
        Replace the new hierarchy vector to $< X_s - X' >$ in equivalent class;
        /* Check data is anonymized successfully for publishing.              */
        if (Anonymity constraints are satisfied by all equivalent classes) then
            return;
        end
    until Highest level in each $DGH_{\xi_i}$;
    /* Start a new round candidates search in the Tree($X'$) for
       generalization                                                         */
    $X_s \leftarrow X'$ ;
until Anonymity constraints are achieved;
```

References

1. Bayardo, R.J. and Agrawal, R.: Data Privacy through Optimal k-Anonymization. In Proceedings ICDE'05, pp.217–228. IEEE(2005)
2. Sweeney, L.: k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness, and Knowledge-based Systems. 10(5), 557–570 (2002).
3. Machanavajjhala, A.,etc.: L-diversity: privacy beyond k-anonymity. In TKDD'07 1(1). ACM(2007)
4. Sweeney, L.: Achieving k-anonymity privacy protection using generalization and suppression. International Journal on Uncertainty, Fuzziness, and Knowledge-based Systems. 10(6):571–588 (2002).
5. LeFevre, K., DeWitt D., and Ramakrishnan, R.: Incognito: Efcient full-domain k-anonymity. In Proceedings of SIGMOD'05. pp. 49-60. ACM(2005).
6. Truta, T.M. and Vinay, B.: Privacy protection: p-sensitive k-anonymity property. In Proceedings of ICDE'06. IEEE Computer Society(2006).
7. Li, N., Li. T. and Venkatasubramanian, S.: t-closeness: privacy beyond k-anonymity and l-diversity. In 23rd International Conference on Data Engineering. pp. 106-115. IEEE Computer Society.(2007)
8. Chang, C. and Lin, C.: LIBSVM: A library for support vector machines. In ACM Transactions on Intelligent Systems and Technology. 2(3), pp. 1–27. ACM, New York, USA (2011).
9. Zhong, S., Yang, Z. and Wright, R.N.: Privacy-enhancing k-anonymization of customer data. In Proceedings of the 24th ACM SIGMOD symposium on Principles of database systems. pp. 139–147. ACM, New York, USA (2005).
10. Samarati, P.: Protecting Respondents' Identities in Microdata Release. In IEEE Trans. on Knowl. and Data Eng. 13(6), pp. 1010-1027. IEEE Educational Activities Department, Piscataway, NJ, USA (2001).
11. Kasiviswanathan, S.P., Rudelson, M., Smith, A. and Ullman, J.: The price of privately releasing contingency tables and the spectra of random matrices with correlated rows. In Proceedings of the 42nd ACM symposium on Theory of computing. pp. 775–784. Cambridge, Massachusetts, USA (2010).
12. Notice of Addresses for Submission of HIPAA Health Information Privacy Complaints Federal Register, Vol. 68, No. 54, March 20 (2003)
13. Sweeney, L.: Weaving technology and policy together to maintain confidentiality. Journal of Law, Medicine and Ethics. 25:98-110 (2000).
14. Wong, R., Li, J., Fu, A., Wang, K.:(α, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 754–759 (2006).
15. Terrovitis, M., Mamoulis, N., and Kalnis, P.: Local and global recoding methods for anonymizing set-valued data. In the International Journal on Very Large Databases, 20(1), pp. 83–106 (2011).
16. Xu, J., Wang, W., Pei, J., Wang, X., Shi, B., and Fu, A. W. C.: Utility-based anonymization using local recoding. In SIGKDD'06. pp.785–790. ACM(2006).
17. Loukides, G. and Shao, J.: Capturing data usefulness and privacy protection in k-anonymisation. In Proceedings of the 2007 ACM symposium on Applied computing. pp. 370–374). ACM (2007).
18. LeFevre, K., DeWitt, D. J., and Ramakrishnan, R.: Mondrian multidimensional k-anonymity. In ICDE'06. pp. 25–36. IEEE(2006).
19. Fung, B. C., Wang, K., and Yu, P. S.: Top-down specialization for information and privacy preservation. In ICDE'05. pp.205–216. IEEE(2005).