iHiMod-Perturb: Histogram Modification Based Reversible Data Perturbation Algorithm For Adaptable Privacy Preservation and Integrity

Dr Alpa K Shah, Dr Ravi M Gulati

Abstract: With the advent of users' perspective to store personal information digitized, measures to protect their private information has necessitated. Organizations also publish anonymous data to facilitate research in social science, healthcare, identification of fraudulent users, and trend analysis to foster sales and marketing techniques. Privacy Preserving Data Mining [PPDM] protects the disclosure of sensitive information present in the published datasets. Most of the existing pool of algorithms perturb the attributes forbidding verification of original data at receivers' end. This paper introduces a Reversible Data Perturbation iHiMod-Perturb approach efficient in privacy preservation and enabling recovery of original data, if needed, at the receiver end. Unlike traditional methods that adds/multiply noise or randomly project data, our method uses differences of adjoining values as basis for modification to perturb sensitive attributes. Selection of privacy factor and embedding of digital watermark ensures data integrity along with perturbation. The privacy factor enables user specific adaptable privacy preservation model in contrast to one-level model. Experiments are performed on five datasets from UCI Repository and confirms that the Classification Accuracy of the perturbed dataset is preserved well. Experiments also suggest that the Probability Information Loss [PIL] is less than 25% and Disclosure Risk [DR] is less than 8% after application of iHiMod-Perturb algorithm.

Index Terms: Histogram Modification, Privacy Preserving Data Mining, Reversible Data Perturbation, Naïve Bayes, Decision Tree, Support Vector Machines

1. INTRODUCTION

RECENT trends to collect huge pool of information has necessitated mining as an important aspect to obtain valuable information from enterprises. With more information being stored on cloud, data mining has now evolved as an important service. The main objective of cloud computing is to provide better computing facilities with the use of Internet, enabling efficient collaboration amongst different sites, which may be demographically different. Various organizations store their data on clouds, which can be collaboratively used for mining purpose. The most important issue is protecting the privacy of sensitive attributes, without affecting the knowledge. The risk of disclosing information should be minimal, and the accuracy desired by the data miners should not be affected. Data owners might not trust the users or service suppliers and hence a privacy preserving mechanism is required to augment trust of the providers. This research area, where the main focus is to preserve the privacy of contributing confidential attributes without affecting the mining results, is called Privacy Preserving Data Mining [PPDM] (Aggarwal CC et al., 2008 [2]; Agrawal R and Srikant R, 2000 [3]; Fung BCM et al., 2007 [5], Kabir SMA et al., 2007 [17]; Wu XD et al., 2006 [27]; Zhu X and Davidson I, 2007 [34]; A Shah and R Gulati, 2016 [1]; Kunta Ramu and V. Ravi, 2009 [18]). Authors [28] (Witten IH and Frank E, 2005) have suggested and showed a confirmed relationship between the information earned from knowledge and personal data, necessitating assistance in decision making by examination of true knowledge. Many algorithms for PPDM distort the original values, whereby recovery at the receiver end for verification, if at all needed, is not accurate.

Table 1 shows an example of Hospital Dataset containing Direct Attributes, Quasi-Identifier Attributes and Sensitive Attributes. Firstly, the Direct Attributes (such as Name, Addresses, or identity card numbers ID), which identify individuals are removed from published dataset. The rest of attributes are classified into two categories; (a) attributes that can indirectly identify a person's identification, in this case, Age, Cholesterol and Triglyceride, called Quasi-Identifiers (QI) (b) sensitive attributes that contain private or sensitive information, here Disease attribute. The Quasi-Identifiers are perturbed to prevent interested parties from using the attributes for record linkage to the original records. Table 2 consists of perturbed dataset, generated by using different methods of perturbation. Attribute Age is perturbed by using Global Recoding anonymization method. The values are aggregated in a variable into pre-defined classes. Attribute Cholesterol is perturbed by Local Suppression method, whereby the last value is replaced with an asterisk (*). Attribute Triglyceride is also perturbed by Local Suppression method, but by rounding off the values to tens. These anonymization methods, although are useful in hiding the information effectively, cause loss in knowledge hidden in dataset. Put it in other way, such data anonymization results in inaccurate analysis that is not desirable for statistical analysis required for mining.

- *Dr. Alpa K Shah is currently working as Assistant Professor in MCA Dept., Sarvajanik College of Engg. and Technology, Surat, Gujarat. E-mail: alpa.shah@scet.ac.in*
- *Dr. Ravi M Gulati is currently working as Associate Professor at Computer Science Dept., Veer Narmad South Gujarat University, Surat, Gujarat E-mail: rmgulati@vnsgu.ac.in*

TABLE 1
AN ILLUSTRATIVE HOSPITAL DATASET

Direct Attributes		QI Attributes			Sensitive Attribute
ID	Name	Age	Cholesterol	Triglyceride	Disease
Xc101032	Raj Shah	45	165	115	Thyroid
Ay23232	Kunal Suthar	36	255	156	Heart Disease
Bk34532	Prince Jain	42	170	148	Cancer
Sf45454	Khushi Mishra	46	182	127	Leukemia
Xy78976	Dipti Gujjar	50	169	124	Diabetes
Mw2343	Neha Shah	33	157	131	Diabetes
Li898980	Viral Modi	49	166	165	Cancer
Qx34352	Kashyap Jain	38	178	233	Heart Disease
Bz12456	Leena Mathur	41	195	142	Leukemia
Ms23888	Riya Desai	29	200	136	Thyroid
Vd34516	Shruti Kosamiya	62	210	122	Hyper Tension
Jg451282	Jinal Doctor	28	180	170	Leukemia
Uh43432	Nishit Jain	47	152	136	Paralysis
Bh89761	Raju Kulkarni	55	191	143	Thyroid
Rt32412	Minal Prajapati	43	220	111	HIV Positive

The distorted data leads to uncertainty problems that may lead to inaccurate mining decisions [15] (Hong TP et al., 2010). It is therefore essential to protect the privacy of the contributing parties and solve the restoration issue of PPDM. This class of algorithms, in Data Mining, that effectively preserve the privacy of contributing parties and are reversible is called Reversible Privacy Preserving Data Mining [RPPDM]. We have proposed a novel method to perturb the data and enable recovery of the perturbed data efficiently. Existing perturbation-based methods add/multiply noise to all the data uniformly to get perturbed data. The proposed method is novel in terms that it does not use additive/multiplicative noise to perturb data, rather decisively perturbs data.

TABLE 2
THE PERTURBED DATASET FROM TABLE 1

QI Attributes			Sensitive Attribute
Age	Cholesterol	Triglyceride	Disease
[40-49]	16*	110	Thyroid
[30-39]	25*	160	Heart Disease
[40-49]	17*	150	Cancer
[40-49]	18*	130	Leukemia
[50-59]	16*	120	Diabetes
[30-39]	15*	130	Diabetes
[40-49]	16*	160	Cancer
[30-39]	17*	230	Heart Disease
[40-49]	19*	140	Leukemia
[20-29]	20*	140	Thyroid
[60-69]	21*	120	Hyper Tension
[20-29]	18*	170	Leukemia
[40-49]	15*	140	Paralysis
[50-59]	19*	140	Thyroid
[40-49]	22*	110	HIV Positive

Reversible Data Hiding [RDH] is extensively used in Image Processing [20] (M. Wu and B. Lin, 2003) to embed a secret message in images without the host image being destroyed and imperceptible to eyes. Only a legitimate user can retrieve the secret image and restore the cover image. The reversible techniques have been proposed in various fields like video

[12] (D. Rui and J. Fridrich, 2002), visible watermarking [29] (Y. Hu and B. Jeon, 2006), audio [22] (M. Van der Veen, 2003), SVQM-based compression domain (C. C. Chang et al., 2006 [7]) and integer-to-integer wavelet domain (M. Fallahpour and M. H. Sedaaghi, 2007). Histogram modification-based technique (M. Fallahpour and M. H. Sedaaghi, 2007; S. K. Lee et al., 2006 [25]; Z. Ni et al., 2006 [32]) uses pixel difference to increase the hiding capacity. Based on the peak points and zero points, the recipient can retrieve the image. We have used the concept of RDH and Histogram based approach to design the iHiMod-Perturb method (Integrity centred Histogram Modification based Perturbation Method). The proposed method is novel in terms of being reversible and having user-specific adaptable perturbation model. Based on the privacy factor chosen by the provider, different versions of perturbed data can be produced. The paper has been organized as follows. In Section 2, we have dug into deep the research fronts of RPPDM. In Section 3, we have elucidated the proposed Histogram Modification based perturbation algorithm. Both the Perturbation and Recovery phase algorithms are described in detail. The experimental results and privacy preserving analysis is presented in Section 4. Finally, in Section 5, we have given the conclusions and future work anticipated in this direction.

2 RELATED WORK

Perturbation based PPDM algorithms generally use methods such as swapping [11] (Chun JY et al., 2013; Zhu D et al., 2009 [35]), modification [30] (Yang W and Qiao S, 2010) and deletion (Herranz J et al., 2010). The main goal is to protect the original data by minimizing the correlation between the perturbed and the original data. These methods do not allow recoverability of the original data. If the original data is lost, then users will not be able to verify the authenticity of the perturbed data (Herranz J et al., 2010 [14]; Chen TS et al. 2013). Authors [8] (Chen TS et al. 2013) had first proposed perturbation-based approach using privacy difference expansion (PDE) algorithm, capable of protecting the privacy and recovering the original data. The method pairs adjacent pixels into groups and based on parameter setting of the user, determine the difference value in each group. The method uses Principal Component Analysis to reduce the difference between successive values. Setting appropriate parameters is not easy as these parameters do not have any specific relationship with knowledge reservation. Also, there is limited length of payload that can be hidden in the data. Authors [9] (Chen-Yi Lin, 2016) have used the concept of reversible integer transformation for purpose of adjusting the ratio difference. This method can embed more watermark bits and get higher payload. Also, this method uses adjustable weight values, flexible degree of data perturbation. But the method does not specify measures to select Quasi-Identifier attributes from the dataset. For PPDM algorithms, it is essential to identify these attributes so they can be perturbed. Again, optimally selecting the group size and weight parameters are still important aspects under considerations. Authors [10] (Chen-Yi Lin et al. 2016) have described an algorithm for perturbing data that are received in streams. Continuous reversible privacy preserving (CRP) algorithm uses the concept of sliding window to protect the data and embed watermark. Size of the window determines the privacy that can be achieved via CRP algorithm. The method adds/subtracts 1 based on the differences of the group mean. If the group size is compromised all the protection will be futile. There are still other class of reversible data hiding algorithms that uses cryptographic techniques. Authors [36] (Zhuo Hao et al., 2011) have used public-private key and security parameters and have proposed remote data integrity checking protocol with data dynamics. The constructions suffer from issue of complex key management and expensive public key infrastructure. Authors [31] (Yong Yu, 2016) have used identity-based remote data integrity checking. As most of the cryptographic techniques suffer from heavy key management issues, and complex calculations, we have focused our work on perturbation-based approach. Based on the above study, our work focuses on perturbing data decisively. Rather than uniform perturbation across all the data, we have used conditions to add, subtract or letting data unaltered. Uniform perturbation is more vulnerable to attacks and it is possible to reconstruct them using Eigen Value analysis. Privacy factor measure is used to create an adaptable perturbation rather than uniform perturbation. Different values of privacy factor will create different set of perturbations. Hence this adds up a level more of privacy to the contributing parties. Watermark ensures the integrity of the data after recovery. The next section explains the proposed work in detail.

3 IHIMOD-PERTURB ALGORITHM

The main idea of RPPDM is to develop a method, which can protect the privacy of contributing attributes, and enable recovery of the mined perturbed data to the original one. The perturbed data must preserve the privacy that can be measured in terms of accuracy using algorithms like Decision Tree, Naïve Bayes and Support Vector Machine. An efficient RPPDM algorithm must preserve the knowledge reservation by having minimal information loss and be reversible. The iHiMod-Perturb method uses privacy factor to enable individually adaptable privacy protection. The privacy factor is shared between the contributors and is necessary for retrieving the original data. The perturbation algorithm is used to protect the sensitive attribute, and recovery algorithm is used to restore the privacy information. The method also embeds a watermark in the data to provide an additional level of privacy protection.

3.1 Perturbation Algorithm

Based on concept of Histogram modification (M. Fallahpour and M. H. Sedaaghi, 2007 [19]; S. K. Lee et al., 2006 [25]; Z. Ni et al. 2006 [32]) used in Image Steganography, the iHiMod-Perturb generates perturbed values from original dataset. Initially, difference in the adjoining neighbouring values is calculated for the privacy sensitive attribute under consideration. The first value remains unchanged. Peak is then calculated as an average of the differences computed excluding the first value. We do not consider the first value because this value is now an outlier. After the Peak is determined, our method uses conditional checking to add/subtract the privacy factor. The point of importance here is the way in which privacy factor is determined. The contributing parties must get a consensus to privacy factor for restoration if required. Different values of privacy factor will result into specific perturbations. This feature enables adaptable levelled perturbation model. At points where Peak is equal to difference, message bits from watermark are added/subtracted. Both sender and receiver will share privacy factor and watermark for all the attributes, and specific values

of peak for attributes under consideration. Only legitimate users having the privacy factor and peaks will be able to retrieve the original data from the perturbed data. The perturbation algorithm is listed below:

Input	Original Dataset X containing n records X= {x_i, i =1, 2, 3…n} for m attributes considered for perturbation. Watermark w (in decimal) selected by the user to be embedded of length l Privacy factor P_{factor}.
Output	Perturbed Dataset Y. $Peak_m$ for recovery of the original dataset at receiver end.
Step: 1	Set j=1, j ϵ [1, m] and iterate from i= 1, 2, …n
Step: 2	Calculate the difference between adjoining (neighbouring) values. $d_{i,j} = \begin{cases} x_i, & if\ i = 0 \\ \lvert x_{i-i} - x_i \rvert, & otherwise \end{cases}$
Step: 3	Determine Peak for the j^{th} attribute $Peak_j$ = Ceiling $(\frac{1}{n}\sum_{i=2}^{n} d_{i,j})$
Step: 4	Generate the perturbed value by embedding the Privacy Factor and Watermark Message based on current value and its comparison with Peak. If $d_{i,j} \neq Peak_j$ $y_{i,j} = \begin{cases} x_{i,j}, & if\ i = 1\ or\ d_{i,j} < Peak_j \\ x_{i,j} + P_{factor}, & if\ d_{i,j} > Peak_j\ and\ d_{i,j} \geq d_{i-1,j} \\ x_{i,j} - P_{factor}, & if\ d_{i,j} > Peak_j\ and\ d_{i,j} < d_{i-1,j} \end{cases}$ Else $y_{i,j} = \begin{cases} x_{i,j} + w_s, & if\ x_{i,j} \geq x_{i-1,j} \\ x_{i,j} - w_s, & if\ x_{i,j} < x_{i-1,j} \end{cases}$ Here $y_{i,j}$ will be the perturbed value at i^{th} row of j^{th} attribute.
Step: 5	Iterate through j until j ≤ m.

3.2 Recovery Algorithm

If the receiver wants to effectively recover the original dataset, Privacy Factor P_{factor} and $Peak_m$ for perturbed m attributes must be shared. Watermark can be extracted from the perturbed data and verified to check the integrity. The recovery algorithm is listed below:

Input	Perturbed Dataset Y containing n records Y= {y_i = 1, 2, 3,…n} for m attributes $Peak_m$ for recovery of the original dataset at receiver end. Privacy factor P_{factor}
Output	Original Dataset X containing n records and m attributes. Watermark w with length l.
Step: 1	Set j=1, j ϵ [1, m] and iterate from i =1, 2,…n.
Step: 2	The perturbed values be $y_{i,j}$. The first value remains unchanged. $y_{1,i} = x_{1,i}$
Step: 3	Extraction of Watermark Iterate through i until i ≤ n $w_{i,j} = \begin{cases} 0, & if\ \lvert y_i - x_{i-1} \rvert = Peak_j \\ 1, & if\ [\lvert y_i - x_{i-1} \rvert = Peak_j + 1 \end{cases}$ $x_{i,j} = \begin{cases} y_i + P_{factor}, & if\ \lvert y_i - x_{i-1} \rvert > Peak_j\ \&\ y_i < x_{i-1} \\ y_i - P_{factor}, & if\ \lvert y_i - x_{i-1} \rvert > Peak_j\ \&\ y_i > x_{i-1} \\ y_i, & otherwise \end{cases}$
Step: 4	Iterate through j until j ≤ m.

The recovery algorithm will retrieve the original data X, which can be verified by watermark w. Using a smaller privacy factor will yield less perturbation, thereby achieving privacy protection. The remainder of the section will now evaluate the method against its information loss and disclosure risk.

4 EXPERIMENTAL RESULTS AND PRIVACY PRESERVING ANALYSIS

Five datasets of varying sizes were chosen to test the performance of the algorithm. All the datasets were selected from UCI Learning Repository (http://archive.ics.uci.edu/ml/) and US Census Bureau (http://www.census.gov/) (Refer TABLE 3). The number of attributes, total instances and number of classes of the datasets are described in the Table 3.

TABLE 3
TEST DATASETS

Dataset	Number of Attributes	Total Instances	Number of classes
Adult	15	32561	2
Abalone	8	4177	3
Vehicle	18	846	4
Breast Cancer Wiscon	10	699	2
Heart Disease	13*	920	5

*There are total 76 attributes in original dataset, but majority published experiments use only 13 of the attributes (Barin N. Nag; Chaodong Han; Dong-qing Yao 2015 [6]) suggest that data reduction enhances information extraction process.To sort the columns by importance of their contribution for classification accuracy, we have used Decision Tree (Wang XZ, et al., 2012). Decision Tree sorts the columns by importance, resulting into Top N columns that will be used to check the effectiveness of accuracy of knowledge preservation after perturbation. WEKA 3.6.9 (https://www.cs.waikato.ac.nz/ml/weka/; Witten IH, Frank E, 2005 [28]), an open source data mining and analysis tool for knowledge analysis developed by University of Waikato, is used. Decision Tree generated from original dataset is used to identify Top3, Top5 and Top7 attributes. The Top 7 attributes identified are described in Table 4. We have considered only numeric continuous attributes for perturbation. The categorical and nominal attributes need methods like anonymization or generalization for hiding them.

TABLE 4
TOP 7 ATTRIBUTES IDENTIFIED USING DECISION TREE (J.48 ALGORITHM)

Datasets	Top 1	Top 2	Top 3	Top 4	Top 5	Top 6	Top 7
Adult	capital_gain	marital_status	education_num	capital_loss	age	hours_per_week	occupation
Abalone	shell weight	Diameter	length	shucked_weigth	whole_weight	height	viscera_weight

Vehicle	Elongatedness	max_length_aspect_ratio	Compactness	scaled_variance_minor_axis	hollows_ratio	praxis_aspect_ratio	skewness_about_major_axis
Breast Cancer Wiscon	uniformity_of_cell_size	bare_nuclei	uniformity_of_cell_shape	clump_thickness	bland_chromatin	sample_code_number	marginal_adhesion
Heart Disease	Cp	Chol	Ca	Thalach	Age	Thal	trestbps

We perturbed Top 3, Top 5 and Top 7 numeric continuous attributes using iHiMod-Perturb algorithm. The algorithm is implemented in R [https://www.r-project.org/], an open source software environment widely used for statistical computation and graphics. Three classifiers, viz. Decision Tree (DT) [Wang XZ, et al., 2012 [26]], Naïve Bayes (NB) (Zhang ML et al., 2009 [33]) and Support Vector Machine (SVM) [Amari S and Wu S, 1999[4]; Furey TS 2000 [13] ; N.R. Sakthivel, V. Sugumaran, Binoy B. Nair 2010 [23]] were then used to analyse the knowledge accuracy after perturbing Top3, Top5 and Top7 attributes. We used default arguments and 10-fold cross validation to analyse the effect of perturbing attributes on knowledge preservation. We performed experiments to test the effectiveness of the proposed work in terms of knowledge analysis in preserving the information after perturbation. Table 5 describes the results from the experimentations on the five datasets. The prediction accuracy of classification in the dataset is the knowledge analysis result derived after mining the data. The values in the Table 5 clearly indicates the similarity in accuracy significant to the original values. As can be clearly seen in Table 5, the classification accuracy of all the datasets under consideration has been retained after perturbing Top3, Top5 and Top7 attributes. This indicates that the data perturbation after iHiMod-Perturb algorithm preserves the correct knowledge.

TABLE 5
EXPERIMENTAL RESULTS FOR KNOWLEDGE ACCURACY

		DT	NB	SVM
Adult				
	Original	86.23	83.43	75.94
iHiMod-Perturb	Top 3	85.95	83.34	75.56
	Top 5	86.09	83.33	76.43
	Top 7	86.15	83.31	75.56
Abalone				
	Original	80.63	66.07	80.94
iHiMod-Perturb	Top 3	79.77	71.34	80.72
	Top 5	79.89	74.10	80.63
	Top 7	79.63	74.05	80.61
Vehicle				
	Original	72.45	44.79	30.49
iHiMod-Perturb	Top 3	73.40	45.27	30.50
	Top 5	71.04	47.28	31.21
	Top 7	71.28	46.93	31.21
Breast Cancer Wiscon				
	Original	94.99	95.99	66.38
iHiMod-Perturb	Top 3	95.46	96.34	65.01
	Top 5	95.17	96.05	65.94
	Top 7	95.47	96.05	65.94
Heart Disease				
	Original	53.26	52.5	45.00
iHiMod-Perturb	Top 3	54.54	55.55	53.87
	Top 5	52.86	55.89	53.87
	Top 7	53.87	55.22	53.87

A large perturbation may lead to information loss and lead to incorrect mined knowledge. PPDM techniques focus on trade-offs between privacy and utility. It is of utmost importance to balance between knowledge analysis, information loss and disclosure risk We have used the statistical disclosure control to obtain a probabilistic information loss measure that can be used to access the impact of the perturbation on continuous numeric data. The experimental results of Probability Information Loss (PIL) represent the information loss rate, the smaller the values, more desirable. Smaller values represent that the original and perturbed dataset have similarity. We have computed PIL suggested by Authors (Josep M. Mateo-Sanz et. al 2005 [16]) using the Mean, Variance, Covariance, Pearson's Correlation, and Quantile of original and perturbed data. The information loss is measured by these factors. Euclidean distance will measure the distance between the original and perturbed values. It is a measure based on Distance-Based Record Linkage (DR) approach described in Pagliuca and Seri 1999 [24] for micro aggregation masking that uses Euclidean distance. The method is generalized for using in any perturbation method to calculate the distance between original and perturbed attributes. Table 6 shows the calculations for PIL and DR measures computed. Experimental Results of PIL and DR suggest that the perturbation results in minimal information loss. PIL is measure of probability of information loss, smaller values desirable. All the computational results are below 25% which suggest minimal information loss. The DR values suggest that the perturbed values are not very far from the original values. The effect of perturbation on the original values permit a feasible disclosure risk. Here too, like PIL experimental results show that the values are smaller. Also, the effect of perturbing Top 3, Top 5 and Top 7 attributes do not drastically affect the

PIL and DR measures.

TABLE 6
EXPERIMENTAL RESULTS FOR INFO. LOSS

	Top 3	Top 5	Top 7
Abalone			
PIL	19.5	21.3	21.00
DR	3.57	3.67	3.68
Vehicle			
PIL	24.08	24.15	24.12
DR	1.81	2.06	2.00
Breast Cancer Wiscon			
PIL	16.33	16.31	16.25
DR	1.63	1.63	1.62
Heart Disease			
PIL	5.21	5.21	5.22
DR	1.26	1.65	1.91
Adult			
PIL	23.20	23.15	23.15
DR	7.93	11.31	18.78

5 CONCLUSIONS AND FUTURE WORK

A new RPPDM algorithm is proposed in this paper which is efficient in the privacy preservation and knowledge verification of data at the receiver's end. The proposed iHiMod-Perturb algorithm solves the problem of effectively preserving the privacy of sensitive numeric attributes and its recoverability. The individual selection of privacy attributes enables an adaptable privacy protection. Watermark allows to check the integrity of perturbed mining data. Our experiments suggest that the knowledge of mined data is effectively preserved after perturbation. Future work encompasses to study the information loss and other methods to secure knowledge analysis. We also anticipate studying various attacks to check the vulnerability of proposed method.

REFERENCES

[1] A Shah and R Gulati (2016) "Evaluating Applicability Of Perturbation Techniques For Privacy Preserving Data Mining By Descriptive Statistics" in Proceedings of 2016 Intl. Conference on Advances in Computing, Communications and Informatics (ICACCI).

[2] Aggarwal CC, Yu PS (2008) "Privacy-preserving data mining: models and algorithms" Springer, Berlin

[3] Agrawal R, Srikant R (2000) "Privacy-preserving data mining" SIGMOD Rec. 29(2):439–450. doi:10.1145/335191.335438

[4] Amari S, Wu S (1999) "Improving support vector machine classifiers by modifying kernel functions". Neural Netw 12(6):783–789. doi:10.1016/S0893-6080(99)00032-5

[5] B. C. M. Fung, K.Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments" ACM Compute. Surveys, vol. 42. no. 4, pp. 14:114:53, 2010.

[6] Barin N. Nag; Chaodong Han; Dong-qing Yao (2015), "Information enhancement in data mining: a study in data reduction", Int. J. of Data Analysis Techniques and Strategies, 2015 Vol.7, No.1, pp.3 – 20, doi: 10.1504/IJDATS.2015.067698

[7] C. C. Chang, W. L. Tai, and C. C. Lin, "A reversible data hiding scheme based on side match vector quantization" IEEE Trans. Circuits Syst. Video Technol., vol. 16, no. 10, pp. 1301–1308, Oct. 2006.

[8] Chen TS, Lee WB, Chen J, Kao YH, Hou PW (2013) "Reversible privacy preserving data mining: a combination of difference expansion and privacy preserving". J Supercomput 66(2):907–917. doi:10.1007/s11227-013-0926-7

[9] Chen-Yi Lin. "A reversible data transform algorithm using integer transform for privacy-preserving data mining" Journal of Systems and Software doi: 10.1016/j/jss.2016.02.2005

[10] Chen-Yi Lin, Yuan-Hunh Kao, Wei-Bin Lee and Ron-Chang Chen (2016) "An efficient reversible privacy-preserving data mining technology over data streams". SpringerPlus 5:1407 doi: 10.1186/s40064-016-3095-3

[11] Chun JY, Hong D, Jeong IR, Lee DH (2013) "Privacy-preserving disjunctive normal form operations on distributed sets". Inf Sci 231:113–122. doi:10.1016/j.ins.2011.07.003

[12] D. Rui and J. Fridrich, "Lossless authentication of MPEG-2 video" in Proc. IEEE Int. Conf. Image Process., vol. 2. Rochester, NY, 2002, pp. 893–896

[13] Furey TS, Cristianini N, Duffy N, Bednarski DW (2000) "Support vector machine classification and validation of cancer tissue samples using microarray expression data" Bioinformatics 16(10):906–914. doi:10.1093/bioinformatics/16.10.906

[14] Herranz J, Matwin S, Nin J, Torra V (2010) "Classifying data from protected statistical datasets". Comput Secur 29(8):874–890. doi:10.1016/j.cose.2010.05.005

[15] Hong TP, Tseng LH, Chien BC (2010) "Mining from incomplete quantitative data by fuzzy rough sets". Expert Syst Appl 37(3):2644–2653. doi:10.1016/j.eswa.2009.08.002

[16] M. Mateo-Sanz, Josep Domingo-Ferrer, Francese Sebe (2005), "Probabilistic Information Loss measures in Confidentiality Protection of Continuous Microdata", in Int. J of Data Mining and Knowledge Discovery, 11, pp 181-193. doi: 10.1007/s10618-005-0011-9

[17] Kabir SMA, Youssef AM, Elhakeem AK (2007) "On data distortion for privacy preserving data mining" In: Proceedings of the 20th Canadian conference on electrical and computer engineering, pp 308–311.doi:10.1109/CCECE.2007.83

[18] Kunta Ramu, V. Ravi (2009) "Privacy preservation in data mining using hybrid perturbation methods: an application to bankruptcy prediction in banks in Int. J. of Data Analysis Techniques and Strategies" Vol.1, No.4, pp.313 - 331. doi: 10.1504/IJDATS.2009.027509

[19] M. Fallahpour and M. H. Sedaaghi, "High capacity lossless data hiding based on histogram modification" IEICE Electron. Exp., vol. 4, no. 7, pp. 205–210, Apr. 2007.

[20] M. Wu and B. Lin, "Data hiding in image and video: Part I Fundamental issues and solutions" IEEE Trans. Image Process., vol. 12, no. 6, pp.685–695, Jun. 2003.
[21] M. Wu, H. Yu, and B. Liu, "Data hiding in image and video: Part II designs and applications" IEEE Trans. Image Process., vol. 12, no. 6, pp. 696–705, Jun. 2003.
[22] M. Van der Veen, F. Bruekers, A. Van Leest, and S. Cavin, "High capacity reversible watermarking for audio" in Proc. SPIE Security Watermarking Multimedia Contents V, Santa Clara, CA, Jan. 2003, vol.5020, pp. 1–11.
[23] N.R. Sakthivel, V. Sugumaran, Binoy B. Nair (2010) ," Application of Support Vector Machine (SVM) and Proximal Support Vector Machine (PSVM) for fault classification of monoblock centrifugal pump" in Int. J. of Data Analysis Techniques and Strategies, Vol.2, No.1, pp.38 – 61
[24] Pagliuca, D., and G. Seri (1999) "Some Results of Individual Ranking Method on the System of Enterprise Accounts Annual Survey", Esprit SDC Project, Deliverable MI-3/D2.
[25] S. K. Lee, Y. H. Suh, and Y. S. Ho, "Reversible image authentication based on watermarking" in Proc. IEEE Int. Conf. Multimedia Expo, Toronto, ON, Canada, Jul. 2006, pp. 1321–1324
[26] Wang XZ Dong LC, Yan JH (2012) "Maximum ambiguity-based sample selection in fuzzy decision tree induction" IEEE Trans Knowl Data Eng 24(8):1491–1505. doi:10.1109/TKDE.2011.67
[27] Wu XD, Yue DM, Liu FL, Wang YF, Chu CH (2006) "Privacy preserving data mining algorithms by data distortion" In: Proceedings of the international conference on Management science and engineering, pp 223–228
[28] Witten IH, Frank E (2005) "Data mining: practical machine learning tools and techniques" 2nd edn. Morgan Kaufmann, San Francisco
[29] Y. Hu and B. Jeon, "Reversible visible watermarking and lossless recovery of original images" IEEE Trans. Circuits Syst. Video Technol., vol. 16, no. 11, pp. 1423–1429, Nov. 2006.
[30] Yang W, Qiao S (2010) "A novel anonymization algorithm: privacy protection and knowledge preservation" Expert Syst Appl 37(1):756–766. doi:10.1016/j.eswa.2009.05.097
[31] Yong Yu, Man Ho Au, Member, Giuseppe Ateniese, Xinyi Huang, Willy Susilo, Yuanshun Dai, and Geyong Min. "Identity-Based Remote Data Integrity Checking with Perfect Data Privacy Preserving for Cloud Storage". pg 767-778
[32] Z. Ni, Y. Q. Shi, N. Ansari, and W. Su, "Reversible data hiding" IEEE Trans. Circuits Syst. Video Technol., vol. 16, no. 3, pp. 354–362, Mar. 2006.
[33] Zhang ML, Peña JM, Robles V (2009) "Feature selection for multi-label naive Bayes classification". Inf Sci 179(19):3218–3229. doi:10.1016/j.ins.2009.06.010
[34] Zhu X, Davidson I (2007) "Knowledge discovery and data mining: challenges and realities". Information science reference. Hershey, New York
[35] Zhu D, Li XB, Wu S (2009) "Identity disclosure protection: a data reconstruction approach for privacy-preserving data mining". Decision Support Syst 48(1):133–140. doi:10.1016/j.dss.2009.07.003
[36] Zhuo Hao, Sheng Zhong, Member and Nenghai Yu. "A Privacy-Preserving Remote Data Integrity Checking Protocol with Data Dynamics and Public Verifiability" IEEE Transactions on Knowledge and Data Engineering, Vol. 23, No. 9, September 2011 Pg 1432-1437