The Effects of Privacy Preserving Data Publishing based on Overlapped Slicing on Feature Selection Stability and Accuracy

Mohana Chelvan P[1]
Department of Computer Science
Karpagam Academy of Higher Education (KAHE)
Coimbatore, India

Dr. Perumal K[2]
Department of Computer Applications
Madurai Kamaraj University
Madurai, India

***Abstract*—Feature selection is vital for data mining as each organization gathers a colossal measure of high dimensional microdata. Among significant standards of the algorithms for feature selection, the primary one which is currently considered as significant is feature selection stability along with accuracy. Privacy preserving data publishing methods with various delicate traits are analyzed to lessen the likelihood of adversaries to figure the touchy values. By and large, protecting the delicate values is typically accomplished by anonymizing data by utilizing generalization and suppression methods which may bring about information loss. Strategies other than generalization and suppression are investigated to diminish information loss. Privacy preserving data publishing with the overlapped slicing technique with various delicate ascribes tackles the issues in microdata with numerous touchy attributes. Feature selection stability is a vital criterion of data mining technique because of the accumulation of ever increasing dimensionality of microdata due to everyday activities on the World Wide Web. Feature selection stability is directly correlated with data utility. Feature selection stability is data centric and hence modifications of a dataset for privacy preservation affects feature selection stability along with data utility. As feature selection stability is data-driven, the impacts of privacy preserving data publishing based on overlapped slicing on feature selection stability and accuracy is investigated in this paper.**

***Keywords*—Overlapped slicing; privacy preserving data publishing; feature selection; Jaccard Index; selection stability**

I. Introduction

There will be a huge amount of high-dimensional microdata created by organizations because of regular exercises on online business, e-administration, and so on. The table that involves data with singular portrayal or single respondent but not aggregative data is called microdata [1]. In this data, each record has at least one delicate attribute and is independently elucidated in each record [2]. For most organizations, for getting an edge over the contenders, data mining which is the extraction of helpful information from the accumulated transactional data of officialdoms is typically used. Feature selection is a significant dimensionality decrease method in data mining that chooses the subset of pertinent traits. Precision, productivity, and model interpretability is improved by the application of the feature selection technique. Microdata publishing strategies including slicing, overlapped slicing, t-closeness, l-diversity, k-anonymity will bother the data to safeguard the privacy of data. This paper is connected with the effect on data utility along with feature selection stability in data mining by perturbation of dataset for the privacy preserving data publishing methods particularly slicing and overlapped slicing strategies.

II. Data Perspective Nature of Feature Selection Stability

Because of the progressions in Information Technology, the online assortment of microdata about people as high-dimensional datasets is the ordinary movement. In feature selection, just a subset of significant traits is acquired as a dimensionality decrease strategy to defeat the "curse of dimensionality". Regardless of whether there will be a little perturbation or expansion of new samples, there ought to be a selection of a comparative set of traits in feature selection. As a significant measure of algorithms for feature selection, feature selection stability is considered as the ensuing iteration of feature selection should choose a comparable set of traits. Else, it brings down their conviction of researchers, as it will make disarray in the scientist's mind about the findings of their research discoveries. Prior research contributions are toward the path that feature selection stability is generally dependent on algorithms. In any case, late explorers have demonstrated that feature selection is dataset dependent yet not algorithmic free [3-8].

Salem Alelyani and Huan Liu suggest that changes to the characteristics of the data set can affect the feature selection stability [3]. Previous researchers recommended that the feature selection stability is usually algorithmic dependent. However, Salem Alelyani and Huan Liu have tried well that the feature selection stability is usually dataset-reliant, but not entirely algorithmic independent. Salem Alelyani, Zheng Zhao, and Huan Liu suggest that there is a difficulty in measuring the stability of feature selection algorithms [4]. In his doctoral thesis, Salem Alelyani investigated the causes of instability in high-dimensional datasets using well-known feature selection algorithms and proved that feature selection stability mostly depends on data [5].

In the well-known feature selection algorithms, the stability results of different high-dimensional datasets from different domains will not be the same [6]. Some algorithms work better than the other domain datasets for a particular

domain. Noise sensitivity is a major concern for algorithm instability in the selection of features. In many fields, a large amount of high-dimensional data range from social media, e-commerce, bioinformatics, healthcare, and online education [7]. The underlying data characteristics have a significant impact on the stability of the feature selection algorithm, as the stability problem can also depend on the data and these factors include the dimensionality of the feature, the number of data instances, etc. While the supervised feature selection requires prior knowledge on the label of each data instance, a new sample that does not belong to any existing classes will be considered as an outlier and there is no need to change the selected feature set to adapt to the outliers [8]. In other words, unsupervised feature selection is more sensitive to noise and the noise affects the stability of these algorithms.

The discrepancy in the distribution of the dataset can influence the feature selection stability. Data variation will influence selection stability. Particularly for choppy datasets like privacy conserved datasets, feature selection stability is generally influenced as privacy safeguarding annoyance influences the qualities of the dataset as feature selection stability is data-driven.

III. Jaccard Index

There are different measures for feature selection stability. Spearman's and Pearson's correlations expect to gauge the consistency of ranks or weights of the two lists of traits while the Jaccard Index plans to assess the measure of crossover between two sets of trait indices. In the investigations, the Jaccard Index is utilized to gauge feature selection stability. The subsets of results that contain chosen traits' indices and the Jaccard Index assess the stability by assessing the measure of crossover between the subsets [9]. The Jaccard Index estimates the similitude between finite sample sets and is characterized as the size of the crossing point partitioned by the size of the association of the sample sets estimates. Jaccard Index for two chose subsets is shown by the accompanying equation (1). From the given various outcomes R = {R1, R2... Rl} relating to various folds of the data set D, its stability can be surveyed by the measure of crossover between the sets in R as in the equation (2).

$$S_j(R_{i,}R_j) = \frac{|R_i \cap R_j|}{|R_i \cap R_j|} \quad (1)$$

$$S_j(R) = \frac{2}{l\,(l-1)} \sum_{i=1}^{l-1} \sum_{j=i+1}^{l} S_j(R_{i,}R_j) \quad (2)$$

when the Jaccard Index value is 0, which is the result in which the feature selection results are not steady and 1 is the result where the outcomes are indistinguishable, consequently truly stable. Hence the value of Jaccard Index SJ restores is in the interval of [0, 1].

IV. Privacy Preserving Data Publishing

There will be a huge accumulation of high dimensional private data in organizations as every day we carry out our work in online technologies because of advancements in web technologies. For settling on strategic decisions, data mining is basic for business organizations. Data mining has been generally done by people who are not working in the organizations thus safeguarding the privacy of data is significant. Microdata including data about people like clinical data are published for their utility in research works yet uncovering private data could influence the notoriety of the organization and will lead the hefty monetary misfortunes. Privacy preserving data publishing is to ensure the privacy of data by some path before publishing microdata to an outsider for data mining.

V. Privacy Preserving Approaches

A few privacy preserving approaches have been intended for microdata publishing, for example, swapping, suppression, perturbation, randomization, and sampling [10]. Tuples in a similar bucket can't be recognized by their quasi identifier traits by generalization which changes the quasi identifier ascribes in each bucket into "less explicit however semantically reliable" values. Suppression substitutes the recognizing traits with values like '*'. Generalization falls flat on high dimensional data because of the scourge of dimensionality and it causes a lot of information loss because of the uniform-dispersion suspicion. For anonymizing high-dimensional data, bucketization has been chiefly utilized. By arbitrarily permuting the delicate trait values in each bucket after parcels tuples in the table into buckets, bucketization isolates the quasi identifiers with the touchy trait.

VI. Privacy Threats

Attribute disclosure, membership disclosure, and identity disclosure are the dangers of microdata publishing in data mining. Anonymising data would bring about better insurance from these dangers. Identity disclosure is worried about the revelation of identifying traits of the distinct individual. For securing attribute disclosure, coordinating different buckets was significant [11]. If the selection criteria were not a delicate trait value, at that point, it would prompt have a membership disclosure as membership information would surmise an identity of a person through different assaults [12].

VII. Microdata Publishing Techniques

A. k-anonymity

There will be a chance of aberrant identification of records from public databases utilizing a quasi-identifier ascribes, and to tackle this k-anonymity model was created. In this strategy, utilizing generalization and suppression techniques, the granularity of data representation is lessened [13]. In the k-anonymity procedure, each mix of estimations of quasi identifier traits can be vaguely coordinated to in any event k respondents. The values of the traits are discretized into stretches for quantitative ascribes or assembled into various sets of values for categorical attributes [13]. Notwithstanding, this method is lacking to forestall attribute disclosure. The background knowledge attack and the homogeneity attack are the two assaults on k-anonymity.

B. l-diversity

The l-diversity model was intended to deal with certain shortcomings in the k-anonymity model. Particularly when there is a homogeneity of delicate values inside a gathering, the k-anonymity model doesn't ensure the touchy traits relating to the quasi identifier ascribes. To forestall the

assaults on k-anonymity, for example, background knowledge attack and homogeneity attack, the idea of intra-bunch diversity of touchy values is advanced inside the anonymization plan with the bucketization method. There will be not just support of the base gathering size of k, yet besides centers around keeping up the diversity of the touchy traits, in the method of l-diversity [14]. If there are 'l well-represented' values for a delicate trait, the class is said to have l-diversity. It is the most straightforward comprehension of 'well-represented' when there will be a guarantee that there are in any event l distinct values for the touchy trait in every proportionality class [14].

C. t-closeness

All values of a given attribute along these lines independent of its dispersion in the data are in the case of the l-diversity model. In any case, for the condition for genuine data sets, the trait values might be quite slanted. Regularly, a foe may utilize background knowledge on the global appropriation to make derivations about delicate values in the data. The property utilized by the t-closeness model is that the separation between the global dispersion and the dissemination of the delicate trait inside an anonymized gathering ought not to be quite the same as by more than a threshold t and this is [15]. In comparing the numerous other privacy conserving data mining techniques for numeric ascribes, the t-closeness approach will in general be more successful. An equality class is said to have t-closeness if the separation between the spread of the trait in the entire table and the appropriation of a touchy trait in this class is close to a limit t.

D. Slicing

It doesn't matter for data, even that doesn't have a reasonable partition between quasi identifying traits and touchy traits, bucketization doesn't forestall membership disclosure. Particularly for high dimensional data, generalization loses an impressive measure of information. There will partition the data both on a level plane and vertically in the case of a slicing strategy. This method protects preferable data utility overgeneralization. Dividing ascribes into columns will secure privacy by transgressing the relationship of uncorrelated traits and safeguard data usefulness by saving the relationship amongst profoundly interrelated traits [11, 12]. Slicing is more practiced for high-dimensional data. The sliced table is shown in Table I.

E. Overlapped Slicing

An augmentation of slicing techniques is applied in overlapped slicing is. By putting the touchy trait into a quasi-identifier section after copying the touchy trait, the overlapped slicing is performed. The placing of a trait into more than one column is the thought behind overlapped slicing. In overlapped slicing, a touchy trait like sickness will be copied and placed into a quasi-identifier column [16]. Since there is more attribute relationship, this will build data utility. Consequently, Table II depicts overlapped slicing which is an expansion of Table I.

The disease has overlapped as in the first and second columns, which is shown in Table II. It will give better data utility by overlapping this trait. At that point, to keep up privacy ensure, the arbitrary permutation of tuples in each bucket from the second column is performed [17]. How arbitrary permutation is acted in overlapped slicing is portrayed below in Table III. It is not influenced by information loss as the counterfeit tuples are made by this arbitrary stage. Haphazardly permutation of values in each bucket in the second column takes place to break connecting between two columns, as shown in Table III.

Values in the primary bucket {((32, M, Teacher, Cancer)), ((30, M, Lawyer, Viral Infection)), (25, F, Clerk, Heart Disease), (31, F, Teacher, Cancer)} are haphazardly permutated as appeared in Table II, and the values {(25023, Viral Infection), (26364, Cancer), (26385, Heart Disease), (26895, Cancer)} are arbitrarily permutated. Additionally, in this way connecting between the two columns in a single bucket is camouflaged.

Quasi identifiers are not generalized or suppressed in overlapped slicing, as we find in Table III. The method of permuting arbitrarily delicate values in a bucket is shown in this technique. These will in general limit information loss if it requisite be generalized and hence this strategy doesn't create any information loss. Notwithstanding, counterfeit tuples are produced in this overlapped slicing. Touchy values permutation in a bucket results in the outcome. Since the information is complete, these phony tuples don't diminish the utility of data.

TABLE I. SLICED TABLE

(Zip Code, Disease)	**(Age, Sex, Occupation)**
(24281, Heart Disease)	(24, F, Police)
(26385, Heart Disease)	(25, F, Clerk)
(26895, Cancer)	(31, F, Teacher)
(26364, Cancer)	(39, M, Priest)
(23022, Cancer)	**(**32, M, Teacher)
(24102, Viral Infection)	(30, M, Lawyer)
(25023, Viral Infection)	(28, F, Teacher)

TABLE II. TABLE AFTER OVERLAPPED SLICING

(Zip Code, Disease)	**(Age, Sex, Occupation, Disease)**
(24281, Heart Disease)	(24, F, Police, Heart Disease)
(26385, Heart Disease)	(25, F, Clerk, Heart Disease)
(26895, Cancer)	(31, F, Teacher, Cancer)
(26364, Cancer)	(39, M, Priest, Cancer)
(23022, Cancer)	(32, M, Teacher, Cancer)
(24102, Viral Infection)	(30, M, Lawyer, Viral Infection)
(25023, Viral Infection)	(28, F, Teacher, Viral Infection)

TABLE III. OVERLAPPED SLICING TABLE WITH RANDOM PERMUTATION

(Zip Code, Disease)	**(Age, Sex, Occupation, Disease)**
(26895, Cancer)	(24, F, Police, Heart Disease)
(25023, Viral Infection)	(25, F, Clerk, Heart Disease)
(23022, Cancer)	(30, M, Lawyer, Viral Infection)
(24102, Viral Infection)	(32, M, Teacher, Cancer)
(26364, Cancer)	(28, F, Teacher, Viral Infection)
(26385, Heart Disease)	(39, M, Priest, Cancer)
(24281, Heart Disease)	(31, F, Teacher, Cancer)

VIII. EXPERIMENT

A. Methodology

The following gives the intended methodology for the privacy shielding algorithm i.e., slicing and overlapped slicing:

1) The ranking technique of information gain is utilized for picking quasi identifier traits.

2) Mean, variance and, standard deviation which are statistical properties for trial datasets are hounded.

3) CFS feature selection algorithm was utilized on trial datasets.

4) Before privacy ensuring ruffling, the accuracy for picked traits is hounded.

5) By the privacy conserving algorithm and hereby slicing and overlapped slicing, the quasi identifier ascribes, and touchy traits are disturbed.

6) Factual properties, for instance, mean, variance, and, standard deviation are hounded for privacy conserved datasets.

7) For datasets conserved for privacy, the CFS feature selection algorithm has been put on.

8) For the privacy conserved datasets, utilizing Jakkard Index JI, feature selection stability calculations are dogged.

9) After privacy preserving ruffling, the accuracy of picked attributes is hounded.

*10)*For the datasets conserved for privacy utilizing the privacy conservation algorithm i.e., slicing and overlapped slicing, accuracy, feature selection stability, and the statistical properties are dissected.

B. Information Gain IG

It is outlined as a method that provides additional Y data provided by X that displays the value by which the entropy of Y drops [18]. Entropy might be a basis for impurity in an actual training set S. This topic is assigned as IG, which is symmetrical and is determined in (3).

$$IG = H(Y) - H(Y/X) = H(X) - H(X/Y) \quad (3)$$

Knowledge picked up comparable to Y shadowed by the perception of X is equivalent to the knowledge picked up according to X, trailed by the perception of Y. This can be acclimated to help traits with high values, even though it will be not any more informative and is the deficiency of the IG rule. Taking into account the disparity amongst the entropy of the trait and along these lines, the conditional entropy gave the class label, the IG picks the opportunity inside a trait and the class label. It computes the value of a trait, considering the information gain upheld by the class as shown in (4).

$$IG\ (Class, Attribute) = H\ (Class) - H\ (Class \mid Attribute) \quad (4)$$

C. Correlation-based Feature Selection CFS

The preference is indicated for subsets of traits that are noticeably correlated inside the class with the inter-correlation among classes will be scanter [19]. By observing the level of redundancy among them, alongside the unmistakable prophetic capacity of each trait, the value of the attribute subsets is assessed by CFS. Authors have GA as a search strategy with CFS as a fitness function. CFS can be provided with elective search mechanisms and chooses the best trait subset. In (5) CFS is indicated.

$$r_{zc} = \frac{Kr_{zi}}{\sqrt{K + (k-1)r_{ii}}} \quad (5)$$

where, r_{zi} alludes to the mean connections between subset traits and the class variable, r_{ii} alludes to the mean inter-correlation between subset traits, and r_{zc} gives the relationship between different subsets of traits and the class variable, whereas k alludes to the number of subset traits [19].

D. Datasets Used

The two datasets utilized in the experimentations are the COIL 2000 - Insurance Company Benchmark dataset [20] and the KDD - Census-Income dataset [20]. The KEEL dataset repository is the source from which the datasets are accessible. Table IV shows the data sets highlights. Inside the chronicled data sets, the Coil 2000 dataset essentially has numerical values and the KDD dataset has each categorical and numerical value.

E. Experimental Results Analysis

By considering the status of traits, by utilizing the information gain concerning the class, the ranked traits are gotten. The Information Gain IG feature selection algorithm was utilized to accomplish this. The quasi identifier traits are gotten from which are maintained by the ranked ascribes and were chosen for slicing and overlapped slicing methods. For a hundred percent privacy conservation, every domain value of the picked trait has been perturbed. Even with critical background knowledge, an interloper or roguish data miner can't be sure about the exactness of a re-identification.

To pick ascribes from both original and privacy safeguarded data set, the CFS feature selection algorithm is applied. The CFS algorithm does not collaborate with any classifier inside the selection system as the algorithm is a filter-based. BestFirst is the search method utilized in the trial. By 10-fold cross-validation, overfitting is decreased. BestFirst is redesigned with backtracking abilities and it practices greedy hillclimbing to search for the region of trait subsets. BestFirst interest forward once it begins with the empty set of traits or requests toward each way anytime by requesting all single trait extensions and cancellations which are examined at a predetermined point or it can look for in switch once it begins with the total set of traits.

TABLE IV. FEATURES OF DATASETS COIL 2000 AND CENSUS

S. No.	Datasets Characteristics	Datasets	
		Coil 2000	**Census**
1	Type	Classification	Classification
2	Attribute Type	Numerical	Numerical, Categorical
3	Classes	2	3
4	Features	85	41
5	Origin	Real World	Real World
6	Mislaid Values	No	Yes
7	Instances	9782	143228

As the feature selection stability would increment be competent to up to the ideal number of relevant ascribes and after that decrease, the number of traits picked has been maintained in an idyllic assortment. For both the original and the bothered dataset numerical traits, the statistical properties, for instance, standard deviation, variance, and mean are assessed. In comparing with other privacy preserving data mining techniques, in the case of slicing and overlapped slicing methods, the statistical properties are not much affected. For the privacy safeguarded datasets, the undertakings for statistical exhibitions are engaged for validation.

For the dataset Coil 2000 and the dataset Census, feature selection stability is assessed with the stability measure Jaccard Index JI and the results have been revealed up in Fig. 1. As feature selection stability is data-centric, the discrepancy of the dataset, for instance, the privacy safeguarding perturbation is unfairly correlated with the feature selection stability along with the data utility.

The privacy preserving algorithms i.e., slicing and overlapped slicing has made stable feature selection results by considering the statistical properties of the numerical traits of the perturbed datasets which are stanch. The Census dataset has equally numerical and categorical traits, whereas the Coil 2000 dataset has all the numerical ascribes. Also, as it is found from the exploratory results, the Coil 2000 dataset is altogether more stable than the Census dataset, because it just contains numerical traits.

Because the feature selection stability results for the privacy safeguarding algorithms are effective, the accuracy of the privacy conserved datasets is nearly comparable to before modification. There will be a trade-off among accuracy, feature selection stability, and privacy safeguarding perturbation, and hence in the exploratory investigations, the accuracy results are given need alongside feature selection stability over different estimates, for example, ROC, F1-score, precision, AUC, runtime analysis. The data utility and feature selection stability are decidedly related. Thusly, data utility, feature selection stability, and privacy safeguarding have been tried for the slicing and overlapped slicing procedures with two unique datasets. At that point, the exploratory results showed that the activity of the data publishing techniques on investigational datasets prompts a stable feature selection alongside unassailable accuracy. Table V sums up the measures of the controlled experimentation on the test datasets close by the kinfolk which include feature selection stability along with accuracy.

The measures of the test results of the slicing and overlapped slicing are nearly the equivalent. By and by, the overlapped slicing method offers a marginally preferable exhibition over the slicing procedure, as appeared in the diagram beneath in Fig. 1.

TABLE V. FOR SLICING AND OVERLAPPED SLICING, ACCURACY AND FEATURE SELECTION STABILITY MENSURATION SUMMARY USING DATASETS COIL 2000 AND CENSUS

Investigational Outcomes	Slicing		Overlapped Slicing	
	Coil 2000	Census	Coil 2000	Census
Feature selection stability exploiting Jaccard Index JI	0.88	0.83	0.91	0.86
Overall accuracy before ruffling	74.93%	72.78%	74.93%	72.78%
Overall accuracy after ruffling	67.23%	66.12%	71.62%	68.73%
The Accuracy of chosen features before ruffling	80.79%	77.83%	80.79%	77.83%
The Accuracy of chosen features after ruffling	75.28%	74.12%	78.86%	76.72%

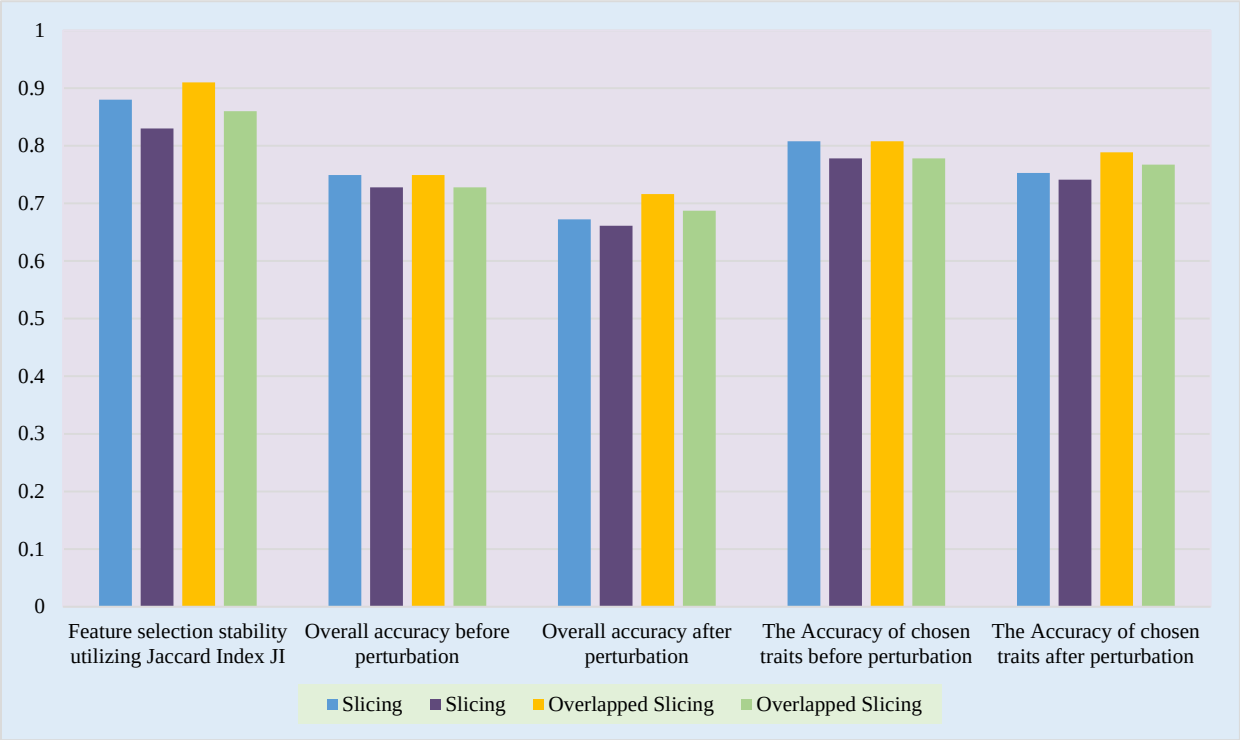

Fig. 1. For Slicing and Overlapped Slicing, Feature Selection Stability and Accuracy Mensuration using Datasets Coil 2000 and Census.

IX. CONCLUSION

This research contribution gives an overview of feature selection stability and its importance in data mining. It likewise examines different privacy preserving data publishing methods. It likewise gives a record of how the data publishing methods influence the selection stability and accuracy in privacy safeguarding data mining. Feature selection stability is now considered a significant criterion in data mining techniques especially due to the accumulation of private data with ever increasing dimensionality due to advancements in internet technologies and smartphones. The slicing and overlapped slicing are recent procedures of privacy preserving data publishing. The key findings of the experimental studies concluded the significance of the overlapped slicing technique in terms of feature selection stability along with data utility in comparison with other privacy preserving data publishing techniques. Overlapped slicing has preferred data utility over slicing strategy. The experimental analyses show that overlapped slicing is superior to the slicing technique regarding feature selection stability and accuracy.

REFERENCES

[1] Can, O. Personalised anonymity for microdata release. IET Inf. Secur. 2018, 2, 341–347.

[2] Taylor, L.; Zhou, X.H.; Rise, P. A tutorial in assessing disclosure risk in microdata. Statistics in Medicine; Wiley: Hoboken, NY, USA, 2018; pp. 1–14.

[3] Salem Alelyani, Huan Liu., The Effect of the Characteristics of the Dataset on the Selection Stability, IEEE DOI 10.1109/International Conference on Tools with Artificial Xiniun Intelligence, 2011.167, 1082-3409/11, http:// ieeexplore.ieee.org/document/ 6103458, 2011.

[4] Salem Alelyani, Zheng Zhao, Huan Liu., A Dilemma in Assessing Stability of Feature Selection Algorithms, IEEE DOI 10.1109/ International Conference on High Performance Computing and Communications, 2011.99, 978-0-7695-4538-7/11, http:// ieeexplore.ieee.org/document/6063062, 2011.

[5] Salem Alelyani, On feature selection stability: a data perspective, Doctoral Dissertation, Arizona State University, AZ, USA, ISBN: 978-1-303-02654-6, ACM Digital Library, 2013.

[6] Barbara Pes, Feature Selection for High-Dimensional Data: The Issue of Stability, Proceedings of the 26th IEEE International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE 2017), June 21–23, 2017.

[7] Jundong Li, Huan Liu, Challenges of Feature Selection for Big Data Analytics, Special Issue on Big Data, IEEE Intelligent Systems, eprint arXiv:1611.01875, 2017.

[8] Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P. Trevino, Jiliang Tang, and Huan Liu, Feature Selection: A Data Perspective. ACM Comput. Surv, 50, 6, Article 94, 45 pages. DOI: https://doi.org/10.1145/3136625, 2018.

[9] Alexandros Kalousis, Julien Prados, and Melanie Hilario. Stability of feature selection algorithms: a study on high-dimensional spaces. Knowledge and Information Systems, 12(1):95–116, May 2007.

[10] Yan Zhao, Ming Du, Jiajin Le, Yongcheng Luo, "A Survey on Privacy Preserving Approaches in Data Publishing" in the First International Workshop on Database Technology and Applications, 2009.

[11] B.Vani, D.Jayanthi, "Efficient Approach for Privacy Preserving Microdata Publishing Using Slicing" IJRCTT, 2013.

[12] Tiancheng Li, Jian Zhang, Ian Molloy, "Slicing: A New Approach for Privacy Preserving Data Publishing" IEEE Transaction on KDD, 2013.

[13] Charu C. Aggarwal, ''On k-Anonymity and the Curse of Dimensionality", Proceedings of the 31st VLDB Conference, Trondheim, Norway, pp.901-909, 2005.

[14] Ashwin Machanavajjhala, Daniel Kifer,Johannes Gehrke, Muthuramakrishnan Venkita Subramanian," ℓ-Diversity : Privacy Beyond K-Anonymity", Proc. International conference on Data Engineering.(ICDE),pp.24, 2006.

[15] Anil Prakash, Ravindar Mogili, ''Privacy Preservation Measure using t-closeness with combined l-diversity and k-anonymity", International Journal of Advanced Research in Computer Science and Electronics Engineering (IJARC SEE)Volume 1, Issue 8,pp:28-33, 2012.

[16] Widodo, Budiardjo EK, Wibowo WC. Privacy Preserving Data Publishing with Multiple Sensitive Attributes based on Overlapped Slicing. Information. 2019; 10(12):362.

[17] Widodo; Wibowo, W.C. A Distributional Model of Sensitive Values on p-Sensitive in Multiple Sensitive Attributes. In Proceedings of the International Conference on Informatics and Computational Science, UNDIP Semarang, Kota Semarang, Indonesia, 30–31 October 2018.

[18] Hall, M A., and Smith L A., Practical feature subset selection for machine learning, Proceedings of the 21st Australian Computer Science Conference, Springer.181- 191, 1998.

[19] Mark A. Hall, Correlation-based Feature Selection for Machine Learning, Dept of Computer science, University of Waikato, http://www.cs.waikato.ac.nz/ mhall / thesis.pdf, 1998.

[20] Alcalá-Fdez, A., Fernández, J., Luengo, J., Derrac, S., García, L. Sánchez, and Herrera, F., KEEL data-mining software tool: Dataset repository, integration of algorithms and experimental analysis framework, J. Multiple-Valued Logic Soft Comput. 17(2): 255–287, 2010.