A New Privacy Preserving Data Mining Algorithm using Slicing along with Value Swapping and Suppression Techniques

Mohana Chelvan P[1], Dr. Perumal K[2]

[1]*Dept. of Computer Science, Hindustan College of Arts and Science, Chennai, India*

[2]*Dept. of Computer Applications, Madurai Kamaraj University, Madurai, India*

Abstract. Most of the privacy preserving data mining algorithms transforms the data in order to preserve privacy which will result in the loss of accuracy. In data mining, feature selection is an important technique for managing "the Curse of Dimensionality". In recent years data become high dimensional and so feature selection is important for data mining for reducing the dimensionality of dataset as it improves accuracy, reduces computational cost and also improves model interpretability. Feature selection stability is the robustness of feature selection algorithms for selecting same or similar set records in subsequent iterations. Unstable feature selection results in confusion in researchers mind about their result findings. So, feature selection stability is recently become important and become new active topic for research. Feature selection stability is mostly depends on the characteristics of the dataset but is not completely algorithmic independent. Privacy preserving data mining transforms the dataset in order to preserve privacy which will affect selection stability as it is mostly dataset dependent. This research paper introduces a privacy preserving data mining algorithm which has good privacy preservation, improved accuracy and feature selection stability.

Keywords: Data Mining, Privacy Preservation, Feature Selection, Selection Stability, Kuncheva Index

I. INTRODUCTION

Data mining might be characterized as the examination of chronicled datasets of organizations to extract possibly valuable, already obscure, non-insignificant, verifiable and intriguing patterns or knowledge. Data mining is fundamental for organizations for getting edge over their rivals. The gathered data of people by the online frameworks are for the most part high dimensional in light of the headways in the web throughput advances which will make the data mining undertakings extremely troublesome and in like manner terms signified as "the curse of dimensionality" [1]. Feature selection is known to be a dimensionality decrease strategy in which important features framing a little subset is picked among the dataset that is unique in agreement to persuaded criteria regarding assessment that is applicable [2], [3]. Feature selection process brings about better learning presentation, for example, brings down computational cost, higher learning accuracy, better model interpretability and lessened storage room. Further, the high dimensional data that has background information or public information can recognize the record proprietors that are hidden and that thus can represent a danger for their privacy.

Feature selection stability is the insensitivity of the algorithm of feature selection for the selection of comparative or similar features that are subsets in consequent cycles of the algorithms for selection of features for the expansion or erasure of few tuples from the dataset [4]. Temperamental feature selection will bring about disarray in the specialist's psyche about their research decisions and the exploratory outcomes end up questionable [5], [6], [7]. Presently a-days, the significance of feature selection stability is acknowledged by the scientists as it diminishes their certainty on their research work. And furthermore selection stability is considered as a vital standard of feature selection algorithms as it turns into a developing point of research [6], [8]. The adjustment in the characteristics of the dataset will impact the feature selection stability. In any case, it isn't totally algorithmic independent [9], [10], [11]. The components that influence the selection stability comprise of number selected features [12], dimensionality, sample size [5] and diverse data distribution crosswise over various folds.

During the time spent data mining, the data for the most part contain delicate individual data, for example, medicinal report or compensation and other money related data which gets presented to a few gatherings including authorities, proprietors, clients and miners. These patterns contain information which is uncovered in decision trees, association rules, classification models or clusters. Private information about individuals or business is contained in the knowledge found by different data mining strategies. Privacy preserving data mining (PPDM) is worried about shielding the privacy of individual data or touchy knowledge without giving up the utility of the data. The current strategies can be for the most part ordered into two general classes [13]:

(I) Methodologies that secure the delicate data itself in the mining procedure, and

(ii) Methodologies that secure the delicate data mining results (i.e. extracted knowledge) that were delivered by the utilization of the data mining.

PPDM has a tendency to perturb the original data with the goal that the after effect of data mining task ought not to challenge privacy imperatives.

Privacy preserving data mining demonstrates the branch of mining that goes for assurance of data that is protection of information that is privacy-sensitive of persons having a place with unsanctioned and some of the time spontaneous disclosure thus guarding the tuples of dataset alongside their privacy. In data mining for safeguarding of privacy, the delicate crude data and furthermore the touchy knowledge of mining comes about are ensured somehow by the perturbation of the original dataset utilizing the created algorithm [14]. Utilizing this method, privacy of the people is protected and in the meantime helpful knowledge is separated from the dataset [15]. The real commitment of good privacy preserving systems is high data quality with privacy. Keeping in mind the end goal to shield the individual's records from being re-recognized, these systems perturb the gathered dataset by some type of change or adjustment before its release [16]. Because of these annoyances, the selection stability will be influenced as it is for the most part dataset subordinate. More changes to the dataset will bring about precarious feature selection which will prompt less data utility. It has been discovered that there has been no profitable research put significant to the point i.e., the connection between perturbation of data for data mining for conservation of privacy and feature selection stability.

II. METHODOLOGY

a. Proposed Methodology

Datasets of microdata contain lot of public information because of advancements in internet technologies which may increase the dimensionality of the datasets and is known as “the Curse of Dimensionality”. The dataset contains these types of data called as identifiers, quasi identifiers and sensitive attributes. Identifiers are the attributes that uniquely identify the tuple such as roll number of a student. Quasi identifiers are the attributes that are group of identifiers that indirectly identify the tuple as date of birth, age and sex. Sensitive attributes are the attributes that contain sensitive information like salary.

The Feature Selection Algorithm CFS has been used identify quasi identifier attributes. By applying the algorithm, ranked list of attributes obtained. From the ranked list of attributes, quasi identifier attributes are selected. Statistical properties mean, standard deviation and variance are calculated for the experimental dataset. Feature selection algorithm has been applied on the experimental dataset. Accuracy of the selected features calculated. The identified quasi identifier attributes and sensitive attributes by the privacy preserving data mining algorithm as shown in algorithm 1. After the perturbation of the experimental dataset, statistical properties of mean, standard deviation and variance are calculated. Feature selection algorithm is applied on the perturbed dataset. Accuracy of the selected features is again calculated. From the selected features, selection stability is calculated.

b. Privacy Preserving Algorithm

The proposed privacy preserving algorithm used in the experiments is shown in the Algorithm 1. The data alteration can be done in different ways including suppression, perturbation, data swapping, data shuffling, microaggregation, rounding or coarsening and noise addition. In this algorithm slicing technique is used along with value swapping and suppression techniques. Slicing technique is well suited for high-dimensional data. Slicing technique splits the table both horizontally and vertically. Highly correlated attributes are put inside the slice block and uncorrelated attributes are split up. Value swapping has been used to improve the slicing technique for negative association and background knowledge attack. For improving privacy of highly sensitive tuples, column generalization with suppression technique is used.

Input: Microdata Table T

Output: Privacy Preserved Table T*

```
1. For a given table T generates an anonymised table T*
   Privacy requirement R of l-diversity.
2. Add the Database T
3. M={T};DSB=¢;
4. B, S={T*};MI={T-T*-key}
5. While M is not empty
       Split M into buckets B
        If total no. of records are <=100
           Add fake tuples
        Else No need to add fake tuples
5. M=M- {B}
6. Sanitization of tuples by rule based id
       Return DSB
6. Check the incompatible table in each bucket $B_i$ of
   table $T^s$
7. each tuple $r_i$ =< $q_i$, $r_i$ > in $B_i$
8. Set Y = ;
9. c = count (number of rows in B)
```

```
for $i$   $c$ do
Check the tuple $s_i$'s incompatibility
if $q_i$ is not compatible with $r_i$ then
    $Y = Y$ [ {$r_i$}
end if
    end for
    if $|Y|$ = ; then
return F
    else
return Y
    end if
    for each bucket $B_i$ 2  $T^S$ ($T^s$ is sliced table) do
    if algorithm Incompatible($B_i$) then
Set $C = B_i - Y$ (C is the available tuples for value
    swapping)
for each tuple s in Y do
    take the tuple $s$ from Y and swap the sensitive
    values r with the tuple s from C, discard the tuple
    from Y and append to C.
    $Y = Y -$ {s}
    $C = C$[ {$t$}
end for
$B_i = C$
    end if
    $T^{sw} = T^{sw}$ [ {$B_i$}
    end for
    return $T^{sw}$
```

Algorithm 1. Proposed privacy preserving algorithm

III. FEATURE SELECTION ALGORITHMS

The procedure of feature selection is generally in light of the three methodologies viz. filter, wrapper and embedded. The filter approach of feature selection is by evacuating features on a few criteria or measures and in this approach, the integrity of a feature is assessed utilizing intrinsic or statistical properties of the dataset. A feature is chosen for data mining or machine learning application in the wake of assessing it as the most reasonable feature in view of these properties. In the wrapper approach the subset of features is produced and after that decency of subset is learned utilizing some classifier. The ranking of the features in the dataset is the motivation behind some classifier in this approach and a feature is chosen for the required application in view of this rank. The embedded approach tries to make utilization of the benefits of both the filter and wrapper techniques. The principle thought behind these algorithms is the lessening of scan space for a wrapper approach by the filter approach.

a. Information Gain IG

The entropy is the pollution preparing set condition S. It is portrayed as a reflecting measures more data in regards to Y introduced by X which symbolizes the real measure of the entropy of that of Y diminishes [17]. This sort of measure is called Information Gain and is given in (1).

$$IG = H(Y) - H(Y/X) = H(X) - H(X/Y) \quad (1)$$

A symmetrical measure that is inferred once the information on X on watching Y is equivalent to the information that is determined on Y on watching X is known as IG. This IG is regularly adjusted towards those features that have some extra values even in the event of not being helpful. The gain of information as to class is figured based on the value of the assessed attribute. The autonomy existing between the class label and the feature is appropriately surveyed by methods for IG on contemplating the disparity that exists among entropy of the specific feature and in addition restrictive entropy of the class label as indicated by (2).

$$IG\ (Class, Attribute) = H\ (Class) - H\ (Class \mid Attribute) \quad (2)$$

b. Correlation-based Feature Selection CFS

The particular attributes and their subset values are assessed through CFS by considering the redundancy degree among them together with the individual predictive ability of each feature. Feature subsets that are including low inter-correlation between the classes yet that are much corresponded inside the class are favoured [5]. The search systems including genetic search, best-first search, backward elimination, forward selection and bi-directional search can be joined with CFS for deciding the best feature subset which is given in (3).

$$r_{zc} = \frac{k\ r_{zi}}{\sqrt{\ } k + (k - 1)\ r_{ii}} \quad (3)$$

in which r_{zc} indicates the genuine correlation that exist in the class variable and furthermore the subset features that are summed, where k signifies the number of features of subset, r_{zi} means the average of correlations in the class variable alongside the subset features and here r_{ii} indicates the average of inter-correlation in the subset features [5].

IV. SELECTION STABILITY MEASURES

a. Kuncheva Index KI

In the vast majority of the stability measures, there will be cover between the two subsets of the features because of chance. The bigger cardinality of the chose features' lists emphatically corresponded with the chance of overlap. To beat this disadvantage, the Kuncheva Index KI is proposed in [18] which contain correction term to evade the intersection by chance. KI is the main measurement that complies with every one of the prerequisites showed up in [18] i.e., Monotonicity, Limits and Correction for chance. The correction for chance term was presented in KI thus it winds up attractive. Not at all like alternate measurements, won't the bigger estimation of cardinality influence the stability value in KI.

$$KI\,(\text{Ƒ}'_1, \text{Ƒ}'_2) = \frac{|\text{Ƒ}'_1 \cap \text{Ƒ}'_2| \,.\, m - k^2}{k\,(m-k)} \qquad (4)$$

In (4), Ƒ'1 and Ƒ'2 are subset of features chose in consequent iterations of feature selection algorithms, k is number of features in the subsets and m is the total number of features in exploratory dataset. KI's outcomes bound between the scopes of [– 1, 1], where – 1 implies k = m/2, i.e., there is no crossing point between the two subsets of features. KI progresses toward becoming 1 when the cardinality of the intersection set equivalents k, i.e., Ƒ'1 and Ƒ'2 are indistinguishable. KI turns out to be near zero for differently drawn lists of subset of features.

V. EXPERIMENTAL RESULTS

The two datasets utilized as a part of the experiments are Census-Income (KDD) dataset and Insurance Company Benchmark (COIL 2000) dataset. The datasets are acquired from the KEEL dataset store [19]. Table 1 demonstrates the qualities of the datasets. In the recorded datasets, the Census dataset has both categorical and numeric values while the Coil 2000 dataset has just numeric values.

Table 1. Characteristics of datasets Census and Coil 2000

S. No.	Datasets Characteristics	Datasets	
		Census	**Coil 2000**
1	Type	Classification	Classification
2	Origin	Real World	Real World
3	Instances	142521	9822
4	Features	41	85
5	Classes	3	2
6	Missing Values	Yes	No
7	Attribute Type	Numerical, Categorical	Numerical

The ranked attributes are acquired by assessing the noteworthiness of a attribute by estimating the information gain with respect to the class. This was finished by the feature selection algorithm Information Gain IG. In view of the got ranked attributes, the quasi identifiers are recognized and chosen for privacy preserving perturbation. The quasi identifiers and sensitive attributes are perturbed utilizing the privacy preserving algorithm which is appeared in Algorithm 2. Every single domain value of the chose trait has changed for 100% privacy conservation thus a gatecrasher or vindictive data miner even with extensive background information can't make certain about the accuracy of a re-identification.

The feature selection algorithm CFS has been utilized to choose attributes from both original and privacy preserved datasets and the search technique utilized as a part of the trial is BestFirst. CFS algorithm is filter-based, so it doesn't connect with any classifier in the determination procedure. Overfitting is lessened by utilizing 10-fold cross validation. BestFirst utilizes greedy hillclimbing for looking through the space of trait subsets and is enhanced with a backtracking facility. BestFirst may look in reverse in the wake of beginning with the full arrangement of traits or hunt forward in the wake of beginning with the unfilled arrangement of attributes or inquiry in the two bearings subsequent to beginning anytime by considering all conceivable single attribute augmentations and erasures at a predetermined point. The quantity of selected features was kept at ideal number as selection stability will enhance up to the ideal number of applicable features and afterward diminishes.

The feature selection stability estimations of the privacy preserved datasets are Census and Coil 2000 are figured utilizing the stability measure Kuncheva Index KI and the outcome is appeared in the Fig.2. On account of KI, the bigger estimation of cardinality won't influence the selection stability thus it is utilized as a part of the analyses as a stability measure. Selection stability is contrarily associated with the variety of the dataset i.e., perturbation of the training samples. The privacy preserving algorithm has created relatively stable feature selection outcomes as a result of the statistical properties for the numerical characteristics of the annoyed datasets are reliable The dataset Coil 2000 has every one of the attributes as numeric while the dataset Census has both categorical and numerical attributes. Thus from the outcomes, it has been seen that the dataset Coil 2000 is more steady than the dataset Census as it contains just numeric attributes.

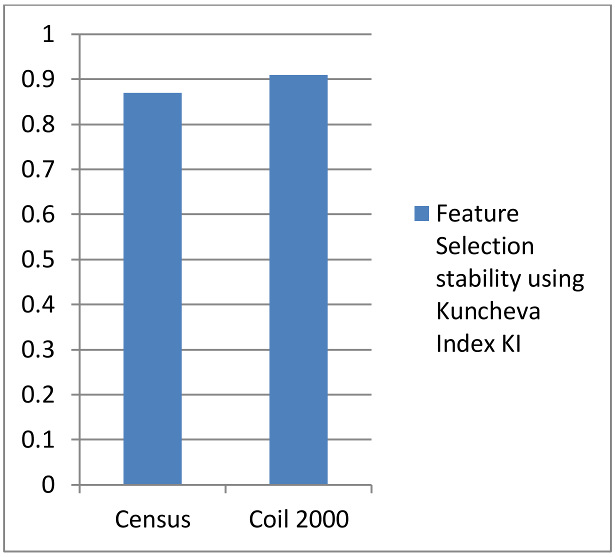

Fig. 2. Feature Selection stability using Kuncheva Index KI for the datasets Census and Coil 2000 after privacy preserving perturbation

Feature selection stability and data utility are decidedly related. As the feature selection stability comes about for the privacy preserving algorithm are great, the precision of the privacy preserved datasets are relatively same as before perturbation. The accuracy results are appeared in the Fig.3.

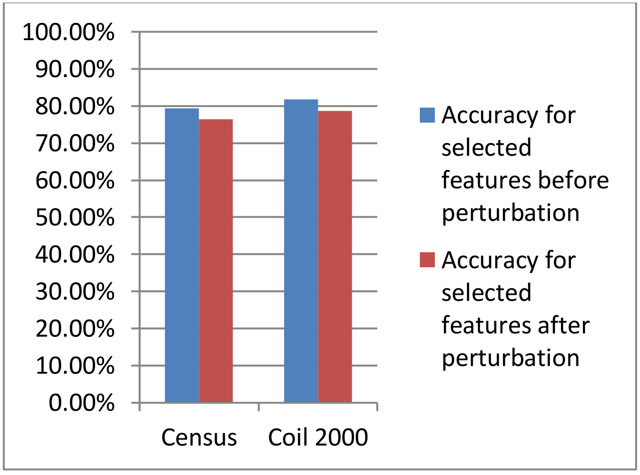

Fig. 3. Accuracy for selected features for the datasets Census and Coil 2000 before and after privacy preserving perturbation

Along these lines, the proposed privacy preserving algorithm has been tried utilizing two diverse trial datasets for its performance in privacy preservation, feature selection stability and data utility. The test comes about have demonstrated that the utilization of the algorithm on test datasets result in stable feature selection with relatively reliable accuracy. The Table 2 condenses the insights of the led probe the datasets in connection with feature selection stability and accuracy.

Table 2. Summary of feature selection stability and accuracy for datasets Census and Coil 2000

Experimental Results	**Datasets**	
	Census	**Coil 2000**
Feature Selection stability using Kuncheva Index KI	0.89	0.93
Overall accuracy before perturbation	74.41%	76.84%
Overall accuracy after perturbation	69.73%	71.62%
Accuracy of selected features before perturbation	79.72%	82.82%
Accuracy of selected features after perturbation	75.42%	77.86%

VI. CONCLUSION

The fundamental goal of privacy preserving data mining is creating algorithm to veil or offer privacy to certain sensitive information with the goal that they can't be disclosed to unapproved gatherings or interloper. Protecting the privacy-sensitive data of people and furthermore dig outing helpful information from microdata is an exceptionally complex issue. There will be tradeoffs between privacy preservation, feature election stability and accuracy. From the trial comes about, it has been reasoned that the proposed privacy preserving algorithm which is used to perturb the quasi identifier attributes and sensitive attributes of the trial datasets will save the privacy of the people. The experiments have determined that the proposed privacy preserving algorithm gave relatively stable feature selection results. In the meantime there will be least change in the accuracy because of the bother of the datasets. In this way, the proposed privacy preserving algorithm used in the tests has safeguarded the privacy of the people and in the meantime gave great feature selection stability and furthermore the diminishing in accuracy is relatively insignificant.

VII. REFERENCES

[1]. Hastie, T., Tibshirani, R., and Friedman, J.: the Elements of Statistical Learning. Springer (2001)

[2]. Guyon, I., and Elisseeff, A.: An introduction to variable and feature selection, Journal of Machine Learning Research, 3:1157–1182 (2003)

[3]. Liu, H., and H. Motoda, H.: Feature Selection for Knowledge Discovery and Data Mining. Boston: Kluwer Academic Publishers (1998)

[4]. Chad A Davis, Fabian Gerick, Volker Hintermair, Caroline C Friedel, Katrin Fundel, Robert Kfner, and Ralf Zimmer. Reliable gene signatures for microarray classification: assessment of

stability and performance. Bioinformatics, 22(19):2356–2363 (Oct 2006)

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

[6]. 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)

[7]. Zengyou He and Weichuan Yu : Stable feature selection for biomarker discovery (2010)

[8]. Kalousis, A., Prados, J., and Hilario, M.: Stability of feature selection algorithms. page 8 (Nov. 2005)

[9]. Salem Alelyani and Huan Liu. : The Effect of the Characteristics of the Dataset on the Selection Stability 1082-3409/11 IEEE DOI 10.1109 / International Conference on Tools with Artificial Intelligence. 2011.167 (2011)

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

[11]. 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)

[12]. Alexandros Kalousis, Julien Prados, and Melanie Hilario. : Stability of feature selection algorithms: a study

[13]. Aris Gkoulalas-Divanis and Vassilios S. Verikios, "An Overview of Privacy Preserving Data Mining", Published by The ACM Student Magazine, 2010.

[14]. Vassilios, S., Veryhios, Elisa Bertino, Igor Nai Fovino Loredana Parasiliti Provenza, Yucel Saygin, Yannis eodoridis. : State-of-the-art in Privacy Preserving Data Mining. SIGMOD Record, Vol. 33, No.1 (March 2004)

[15]. Xiniun, Q., Mingkui Zong.: An Overview of Privacy Preserving Data Mining, 1878-0296, doi: 10.1016/Procedia Environmental Sciences 12 (2012)

[16]. Agarwal, R and Srikant, R.: Privacy preserving data mining. In Proc. Of the ACM SIGMOD Conference of Management of Data, pages 439-450. ACM Press (May 2000)

[17]. 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)

[18]. Kuncheva, L I., A stability index for feature selection, In Proceedings of the 25th conference on Proceedings of the 25th IASTED International Multi Conference: artificial intelligence and applications, Anaheim, CA, USA,. ACTA Press, 390 – 395 (2007)

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