IPASJ International Journal of Information Technology (IIJIT)
Web Site: http://www.ipasj.org/IIJIT/IIJIT.htm
A Publisher for Research Motivation ..... Email:editoriijit@ipasj.org
Volume 6, Issue 5, May 2018 ISSN 2321-5976

A New Privacy Preserving Data Mining Algorithm with better Feature Selection Stability and Accuracy

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

Data mining digs up formerly not known and precious type of patterns and information acquired from hefty storage of data that is archived. In the latest couple of decades, those progressions over internet technologies realize colossal augmentation in the dimensionality of the dataset stressed for data mining. Feature selection is a vital dimensionality reduction method as it advances accuracy, efficiency and model interpretability of data mining algorithms. Feature selection stability may be viewed to be the sturdiness of the algorithm for feature selection which facilitates picking alike or the same subset of features for minute perturbations in the dataset. The indispensable inspiration driving data mining that is utilised for the privacy preservation is the adaptation of original datasets by means of a method to safeguard privacy of the persons and work out consequent data mining algorithm to acquire information from it. This has led to pull together users' individual data and fed into data mining systems which should guarantee that there is no trouncing of privacy. This perturbation of the dataset will influence the feature selection stability. There will be a relationship sandwiched between privacy preserving data mining and feature selection stability. This paper scrutinizes on this issue after giving an introduction to privacy preserving data mining methods and also it brings is a new-fangled privacy preserving algorithm which has less impact on feature selection stability in addition to accuracy.

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

1. INTRODUCTION

1.1. Feature Selection Stability

Data mining might be characterized as the investigation of stored datasets of trading companies to extract conceivably helpful, beforehand obscure, non-inconsequential, certain and fascinating patterns or knowledge. Data mining is basic for trading companies for getting edge over their rivals. The gathered information of people by the online frameworks are for the most part high dimensional as a result of the headways in the web throughput innovations which will make the data mining assignments exceptionally troublesome and in like manner terms indicated as "the curse of dimensionality" [1]. Feature selection is known to be a dimensionality diminishment method in which relevant features shaping a small subset is picked among the dataset that is unique in agreement to persuaded criteria regarding assessment that is significant [2], [3]. Feature selection processes brings about better learning presentation, for example, bring down computational cost, higher learning exactness, better model interpretability and decreased storage space. Additionally, the high dimensional data that has background data or open data can distinguish the record proprietors that are fundamental and that thusly can represent a risk for their protection.

Steadiness of feature selection is the lack of care of the algorithm of feature selection for the choice of comparative or similar features that are subsets in ensuing iterations of the algorithms for selection of features for the expansion or erasure of few tuples from the dataset [4]. Insecure feature selection will bring about perplexity in the scientist's brain about their examination decisions and the test outcomes about wind up questionable [5], [6], [7]. Presently, the

significance of feature selection stability is acknowledged by the analysts as it diminishes their certainty on their examination work. And furthermore selection stability is considered as an essential basis of feature selection algorithms as it turns into a budding point of research [6], [8]. The adjustment in the attributes of the dataset will impact the feature selection stability. But it is not completely algorithmic independent [9], [10], [11]. The variables that influence the selection stability comprise of number selected features [12], dimensionality, sample size [5] and diverse data appropriation crosswise over various folds.

1.2. Privacy Preserving Data Mining

During the time spent data mining, the data for the most part contain delicate individual data, for example, therapeutic report or pay and other monetary data which gets presented to a few gatherings including authorities, proprietors, clients and mineworkers. These examples contain data which is uncovered in decision trees, association rules, classification models or clusters. Private data about individuals or business are contained in the information found by different information mining strategies. Privacy preserving data mining (PPDM) is worried about guarding the privacy of individual data or delicate knowledge without yielding the utility of the information. The current strategies can be for the most part arranged into two general classifications [13] as (i) Methodologies that secure the touchy information itself in the mining procedure, and (ii) Methodologies that secure the touchy data mining outcomes (i.e. extracted knowledge) that were delivered by the use of the data mining. PPDM has a tendency to bother the original data so the after-effects of data mining undertaking ought not to resist privacy requirements.

Data mining for safeguarding of privacy demonstrates the branch of mining that goes for assurance of data that is privacy-sensitive of people having a place with unsanctioned and infrequently spontaneous disclosure thus guarding the tuples of dataset alongside their privacy. In data mining for conservation of privacy, the touchy crude data and furthermore the delicate knowledge of mining outcomes are ensured somehow by the perturbation of the original dataset utilizing the created algorithm [14]. Utilizing this system, privacy of the people is safeguarded and in the meantime valuable information is extorted from the dataset [15]. The real commitment of good privacy preserving strategies is high data quality with privacy. To shield the individual's records from being re-identified, these methods bother the gathered dataset by some type of change or alteration before its release [16]. Because of these perturbations, the selection stability will be influenced as it is generally dataset reliant. More changes to the dataset will bring about unsteady feature selection which will prompt less data utility. It has been discovered that there has been no significant research put pertinent to the subject i.e., the connection between perturbation of data for data mining for protection of privacy and feature selection stability.

2. DIMENSIONS OF PPDM TECHNIQUES

Following is the rundown of five measurements that are used for PPDM Techniques [17]:

- Data distribution
- Data modification
- Data mining algorithms
- Data or rule hiding
- Privacy preservation

2.1. Data Distribution

This measurement is connected with the dissemination of data. There are a few methods which are created for centralized data, while others allude to a distributed data situation. Distributed data situations can be isolated as horizontal data partition and vertical data partition. Horizontal distribution alludes to these situations where distinctive arrangements of records exist in different places, while vertical data distribution alludes where every one of the qualities for various attributes dwell in different places.

2.2. Data Modification

Data modification is utilized with the point of changing the unique values of a database that needs to be permitted to people in general and along these lines to ensure high privacy protection. Techniques for data modification include:

- Perturbation: This can supplant attribute value by esteem or including noise.

- Blocking: which is the substitution of a current attribute value with a "*"
- Swapping: This alludes to interchanging values of individual tuple.
- Sampling: This alludes to losing data for just example of a populace
- Encryption: numerous Cryptographic strategies are utilized for encryption.

2.3. Data Mining Algorithms

The privacy preservation system is intended for data mining algorithm as given underneath:

- Classification data mining algorithms
- Association Rule mining algorithms
- Clustering algorithms

2.4. Data or Rule Hiding

This measurement alludes to whether raw data or assembled data ought to be covered up. Data hiding means securing delicate data values of a few people. Furthermore, Rule hiding means Protecting Confidential Knowledge in data, for instance, association rule. The trouble for covering up totalled data as tenets is extremely bothersome, and for this reason, heuristics have been created.

2.5. Privacy Preserving Techniques

So as to guarantee secrecy crosswise over mining stages, there are diverse privacy preserving techniques that can be utilized in light of mining algorithm. There are various methods like the centralization of condition, utilization of cryptographic algorithms, anonymization of k-anonymity which guarantees better privacy with a lesser computation in independent situations.

- **Heuristic-based techniques:** It is a versatile alteration that adjusts just selected values that limit the viability misfortune instead of every single available value.

- **Cryptography-based techniques:** This strategy incorporates secure multiparty computation where a computation is secure toward the finishing of the computation; nobody can know anything aside from its own information and the outcomes. Cryptography-based algorithms are considered for defensive privacy in a dispersed circumstance by utilizing encryption methods.

- **Reconstruction-based techniques:** where the original conveyance of the data is reassembled from the randomized data.

3. CLASSIFICATION OF PPDM TECHNIQUES

The data mining strategies for privacy preservation might be isolated into two major categories known as the systems of masking and the procedures for synthetic generation of data. In the systems for masking, they protect secrecy of respondents by which the original dataset is adjusted such that it delivers new-fangled datasets that are reasonable for investigation of statistical data. The methods for masking might be isolated into two kinds i.e., non-perturbative and perturbative. In non-perturbative masking strategy, the original dataset stays in place, yet certain sorts of data are hidden and now and then certain points of details are severed. Cases of the non-perturbative systems can be Sampling, Global recoding, Local suppression, Top coding, Bottom coding and Generalization. On account of the strategy of perturbative masking, the original data are adjusted Cases for the perturbative masking procedures are MASSC, PRAM, Resampling, Swapping, Random noise, Rank swapping, Micro-aggregation, Lossy compression, and Rounding. In the methods of generation of synthetic data, the tuples in an arrangement of data is supplanted with the tuples of another set yet the original data's main statistical properties are however saved. That microdata that is discharged is properly blended with that of the original or is set aside completely synthetic. In view of these measurements, distinctive PPDM systems might be characterized into following five classifications [17, 18, 19, 20, 21].

- Anonymization based PPDM
- Perturbation based PPDM
- Randomized Response based PPDM

- Condensation approach based PPDM
- Cryptography based PPDM

We talk about these in detail in the accompanying subsections.

3.1. Anonymization based PPDM

The fundamental type of the data in a table comprises of following four kinds of characteristics:

- Explicit Identifiers is an arrangement of attributes containing data that distinguishes a record proprietor expressly.
- Quasi Identifiers is an arrangement of attributes that could conceivably recognize a record proprietor when joined with openly accessible data.
- Sensitive Attributes is an arrangement of attributes that contains delicate individual particular information.
- Non-Sensitive Attributes is an arrangement of attributes that makes no issue if uncovered even to conniving gatherings.

Anonymization alludes to an approach where identity or/and sensitive data about record proprietors are to be covered up. It even expects that sensitive data ought to be held for investigation. Clearly explicit identifiers ought to be banished yet at the same time there is a threat of privacy interruption when quasi identifiers are connected to openly accessible data. Such assaults are called as linking attacks. Sweeney [22] proposed k-anonymity model utilizing generalization and suppression to accomplish k-anonymity i.e. any person is discernable from at any rate k-1 different ones regarding quasi-identifier attribute in the anonymized dataset. Supplanting a value with less particular yet semantically predictable value is called as generalization and suppression includes obstructing the values. Releasing such data for mining lessens the danger of identification when joined with publically accessible data. Be that as it may, in the meantime, precision of the applications on the changed data is decreased. Various algorithms have been planned to put into practice k-anonymity utilizing generalization and suppression lately. Despite the fact that the anonymization strategy guarantees that the changed data is valid however endures substantial information misfortune. Also it isn't safe to homogeneity attack and background knowledge attack for all intents and purposes [18]. Be that as it may, as a research bearing, k-anonymity in blend with other privacy preserving techniques should be examined for identifying and notwithstanding blocking k-anonymity infringement. Identity or sensitive data about record proprietors are to be covered up.

3.2. Perturbation Based PPDM

Perturbation has been being utilized as a part of statistical disclosure control as it has an inborn property of straightforwardness, proficiency and capacity to save statistical information. In perturbation the original values are transformed with a number of synthetic data values so the statistical information figured from the perturbed data does not vary from the statistical information figured from the original data to a bigger degree. The perturbed data records don't consent to exact record holders, so the assailant can't play out the astute linkages or recuperate sensitive knowledge from the accessible data. Perturbation should be possible by utilizing synthetic data generation or additive noise or data swapping. In the perturbation approach, any distribution based data mining algorithm works under a certain supposition to treat each measurement freely. Applicable information for data mining algorithms, for example, classification stays covered up in inter-attribute relationships. This is on the grounds that the perturbation approach treats distinctive attributes independently. Consequently the distribution based data mining algorithms have a characteristic weakness of loss of concealed information accessible in multidimensional records. Another branch of privacy preserving data mining that deals with the weaknesses of perturbation approach is cryptographic procedures. In this method diverse attributes are protected separately. The inconveniences incorporate that the original data values can't be recovered and furthermore loss of information.

3.3. Randomized Response Based PPDM

In Randomized response, the data is wound such that the focal place can't state with chances superior to a pre-characterized threshold, regardless of whether the information from a client contains remedy information or mistaken information. The information got by each single client is twisted and if the quantity of clients is substantial, the total information of these clients can be evaluated with great amount of exactness. This is exceptionally profitable for decision-tree classification. It depends on joined values of a dataset, to some degree individual data items. The data gathering process in randomization strategy is done utilizing two stages [18]. Amid initial step, the data suppliers

randomize their data and exchange the randomized data to the data recipient. In second step, the data collector modifies the original distribution of the data by utilizing a distribution reconstruction algorithm. Randomization technique is generally extremely basic and does not necessitate knowledge of the dispersion of different records in the data. Henceforth, the randomization technique can be executed at data gathering time. It doesn't necessitate a trusted server to enclose the whole original records keeping in mind the end goal to play out the anonymization procedure [23]. The shortcoming of a randomization response based PPDM system is that it indulgences every one of the records even with regardless of their neighbourhood density. These demonstrate to an issue where the outlier records turn out to be more subject to opponent's assaults when contrasted with records in more compacted regions in the data [24]. One key to this is to be pointlessly adding noise to every one of the records in the data. However, it lessens the benefit of the data for mining intentions as the recreated distribution may not yield results about congruity of the reason for data mining. It is moderately straightforward helpful for concealing data about people. Better proficiency contrast with cryptography based PPDM procedure [25]. The burdens incorporate loss of person's information and this strategy isn't for multiple attribute databases because of linking attack.

3.4. Condensation approach based PPDM

Condensation approach develops constrained clusters in dataset and after that creates pseudo data from the statistics of these clusters [26]. It is called as condensation on account of its approach of utilizing dense statistics of the clusters to create pseudo data. It makes sets of disparate size from the data, with the end goal that it is certain that each record lies in a set whose size is at any rate alike to its anonymity level. Propelled, pseudo data are produced from each set in order to make a synthetic data set with an indistinguishable aggregate distribution from the unique data. This approach can be adequately utilized for the classification issue. The utilization of pseudo-data gives an extra layer of security, as it winds up hard performing antagonistic assaults on synthetic data. In addition, the aggregate behaviour of the data is saved, building it valuable for an assortment of data mining issues [23]. This strategy assists in better privacy preservation when contrasted with different methods as it utilizes pseudo data instead of changed data. Also, it works even without overhauling data mining algorithms because the pseudo data has an indistinguishable format from that of the original data. It is extremely compelling if there should arise an occurrence of data stream issues where the data is exceedingly dynamic. In the meantime, data mining comes about get influenced as colossal measure of data is discharged in light of the pressure of a bigger number of records into a solitary statistical group entity [18]. Utilize pseudo data as opposed to changed data. This technique is genuine if there should arise an occurrence of stream data. The drawbacks incorporate huge measure of information lost and it contains an indistinguishable format from the original data.

3.5. Cryptography Based PPDM

Consider a situation where different organizations desire to lead a joint research for some shared advantages without uncovering superfluous information. In this situation, research about in light of different parameters is to be directed and in the meantime privacy of the people is to be ensured. Such situations are demoted as distributed computing scenarios [27]. The parties engaged with mining of such assignments can be common un-trusted parties, contenders; hence ensuring privacy turns into a noteworthy concern. Cryptographic systems are in a perfect world implied for such situations where numerous gatherings team up to process results or offer non sensitive mining results and accordingly keeping away from exposure of sensitive information. Cryptographic procedures offer a very much characterized show for privacy that incorporates techniques for demonstrating and measuring it. Huge set of cryptographic algorithms develops to actualize privacy preserving data mining algorithms are accessible in this domain. The data might be dispersed among various teammates vertically or on a level plane. Every one of these strategies is relatively in light of an exceptional encryption convention known as Secure Multiparty Computation (SMC) innovation. SMC utilized as a part of distributed privacy preserving data mining comprises of a set of secure sub protocols that are utilized as a part of on a level plane and vertically apportioned data: scalar product, secure size of intersection, secure set union and secure sum. Albeit cryptographic systems guarantee that the changed data is correct and secure yet this approach neglects to convey when in excess of a couple of gatherings are included. Besides, the data mining results may rupture the privacy of person records. There exist a decent number of arrangements in the event of semi-honest models. Be that as it may, if there should arise of an occurrence of noxious models, fewer investigations have been made. Changed data are correct and secured and has better privacy contrast with randomized approach [25].

4. METHODOLOGY

4.1. Proposed Methodology

The greater part of the techniques for data mining for privacy preservation are influenced by the inescapable revile known as the curse of dimensionality of datasets in happening of public information. The attributes in microdata are for the most part gathered as identifiers, quasi identifiers, and confidential attributes or sensitive attributes [28]. Identifiers are those viewpoints that only perceive a microdata respondent. For instance, government security number and representative number are the attributes that will particularly recognize the tuple with which they are related. The ones known as quasi identifiers are the gathering of viewpoints that can be connected alongside outside data with the end goal of re-recognizing those respondents that are passed on in the information. For example, attributes like birth date, pin code and sex are named as quasi identifiers and can be associated appropriately to outside data on civics and can be utilized to recognize the personality of those respondents that match or even to the vulnerability of a specific gathering of respondents that can be cut down. The sensitive or confidential attributes are those attributes where microdata enclose private information. For example, attributes like compensation or consequence of sickness test can be deemed as private.

The proposed technique is appeared in Algorithm 1. The feature selection algorithm Information Gain IG was utilized to recognize the quasi identifier attributes of the datasets taken for the specific experiment. The IG gives a list of ranks of attributes in view of their implication. From the ranked list of attributes, quasi identifier attributes are chosen. The limit between sensitive attributes and quasi identifiers may wind up obscured because of the curse of dimensionality especially when the foes may have significant background information. The identified attributes known as quasi identifiers and the attributes that are delicate are bothered by the algorithm for data mining for the privacy preservation and are additionally appeared in Algorithm 2. The algorithm for feature selection which is Correlation-based Feature Selection CFS is being connected on the datasets both before and after the perturbation by the privacy preserving algorithm. From the picked feature subsets, feature selection stability is computed by influencing utilization of the stability measure called Kuncheva Index KI. The privacy preserved datasets are appropriately tried for recognizing their data mining utility.

Step 1: Quasi identifiers are selected using Information Gain ranking method.

Step 2: Statistical properties, i.e., mean, standard deviation and variance are calculated for experimental datasets.

Step 3: Feature selection algorithm CFS has applied on the experimental datasets.

Step 4: Accuracy for selected features is calculated before privacy preserving perturbation.

Step 5: Quasi identifier attributes and sensitive attributes are perturbed using privacy preserving algorithm.

Step 6: Statistical properties, i.e., mean, standard deviation and variance are calculated for privacy preserved datasets.

Step 7: Feature selection algorithm CFS has applied on the privacy preserved datasets.

Step 8: Feature selection stability values of the privacy preserved datasets are calculated using Kuncheva Index KI.

Step 9: Accuracy for selected features is calculated after privacy preserving perturbation.

Step 10: Statistical properties, feature selection stability and accuracy are analysed for the privacy preserved datasets and hence for the privacy preserving algorithm.

Algorithm 1. Proposed methodology

4.2. Privacy Preserving Algorithm

The suggested privacy preserving algorithm utilized as a part of the analyses is appeared in the Algorithm 2. The data modification should be possible in various ways together with data swapping, suppression, noise addition, microaggregation, rounding or coarsening, perturbation and data shuffling. In this algorithm, estimated noise was included for numerical trait domain values while generalization method was connected for categorical attribute domain values relating to the identified attributes of quasi identifiers and in addition attributes that are delicate. The algorithm so recommended for the data mining for privacy preservation ensures the data that is delicate with great feature selection stability and precision which will make a model for better data utility.

Input : Experimental dataset

Output : Perturbed dataset for attributes of quasi identifiers as well as sensitive

1. Let the dataset D includes T tuples as D = {T1, T2…Tn}. In every tuple of T contains an attribute set T = {A1, A2,…Ap} in which Ai ϵ T and Ti ϵ D
2. Select attributes that are ranked high based on the method of Information Gain IG ranking technique as the attribute for Quasi Identifier A_Q
3. For Each Sensitive Attribute or Quasi Identifier Attribute A_Q of Tuple T where A_Q ϵ T
   Repeat Step 4 to Step 20
4. If selected attribute A_Q =Numeric
   Then Go to Step 5
   Else Go to Step 17
5. The domain values of the attributes are divided into ranges
6. Assign N = Original value of attribute A_Q
7. Assign L = Lowest value of the range to which A_Q belongs
8. Compute R = L / N
9. If R <= 0.5
   Then Go to 10
   Else Go to 11
10. Compute R1 = 1 – R
11. Compute R1 = R
12. Assign X = Number of Digits of N
13. Compute R2 = R1 * (10 ^ X)
14. If N < mean
    Then Go to Step 15
    Else Go to Step 16
15. Numeric attribute perturbation: N1 = N + Noise of value R2
    Go to Step 20
16. Numeric attribute perturbation: N1 = N – Noise of value R2
    Go to Step 20
17. If selected attribute A_Q = Categorical
    Then Go to Step 18
    Else Go to Step 20
18. An attribute perturbation in case of categorical : N1 = perturb of the attribute's original value by a process of generalization of the value of the attribute and allocation of a number
19. The number assigned for N1 is mapped to a range.
20. End

Algorithm 2. Proposed privacy preserving algorithm

5. FEATURE SELECTION ALGORITHMS

The procedure of feature selection is for the most part in view of the three methodologies viz. filter, wrapper and embedded. The filter approach of feature selection is by expelling features on a few criteria or measures and in this approach, the decency 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 appropriate feature in view of these properties. In the wrapper approach the subset of features is created and afterward integrity of subset is found out utilizing some classifier. The ranking of the features in the dataset is the reason for several classifiers in this approach and a feature is chosen for the obligatory application in light of this rank. The embedded approach endeavours to make utilization of the upsides of both the filter and wrapper strategies. The principle contemplation following these algorithms is the lessening of search space for a wrapper approach by the filter approach.

5.1. 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 [29]. This sort of measure is described as 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 if there should be an occurrence of not being valuable. The gain of information with respect to class is figured based on the value of the assessed attribute. The autonomy existing between the class label and the feature is properly evaluated by methods for IG on mulling over the difference that exists among entropy of the specific feature and conditional entropy of the class label as indicated by (2).

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

5.2. Correlation-based Feature Selection CFS

The particular attributes and their subset values are assessed through CFS by considering the redundancy degree between them together with the individual predictive ability of each feature. Feature subsets that are together with low inter-correlation among the classes yet that are exceedingly correlated inside the class are chosen [5]. The search methodologies together with backward elimination, bi-directional search, forward selection, genetic search and best-first 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 means the quantity of features of subset, r_{zi} signifies the average of correlations in the class variable alongside the subset features and here r_{ii} signifies the average of inter-correlation in the subset features [5].

6. SELECTION STABILITY MEASURES

The three primary classes of the stability measures in light of the portrayal of the yield of the selection technique are stability by index, stability by rank and stability by weight [30]. Let A and B be subsets of features and A, B ⊂ X, of the comparable dimension or cardinality, k and let r = |A ∩ B| be the cardinality of the convergence of the two subsets. The vital properties of the various stability estimations [8] are as per the following:

- **Monotonicity.** For various features, n and fixed subset size, k, and the more noteworthy the crossing point among the subsets, the consistency index value is bigger.
- **Limits.** The index should dependably hop by ceaseless that does not rely upon n or k. As far as possible value should be come to once the two subsets are same, i.e., when r= k.
- **Correction for chance.** The index must have an enduring value dependably in which the independently drawn feature subsets will have comparable cardinality, k.

Notwithstanding these prerequisites, there are some essential properties which will have affected on the selection stability result that must be mulled over [9] [10].

- The dimensionality of the dataset.
- The number of selected features.
- The sample size.
- The data variance.
- The symmetry of the measurement.

6.1. Kuncheva Index KI

In the greater part of the stability measures, there will be overlap between the two subsets of the features because of chance. The bigger cardinality of the selected features' lists emphatically related with the chance of overlap to beat this disadvantage, the Kuncheva Index KI is proposed in [31] which contain correction term to stay away from the intersection by chance. KI is the main measurement that complies with every one of the necessities showed up in [31] i.e., Monotonicity, Limits and Correction for chance. The correction for chance term was presented in KI thus it ends up attractive. Not at all like alternate measurements, won't the bigger value of cardinality influence the stability value in KI.

$$\left| F'_1 \cap F'_2 \right| . \ m - k^2$$

$$KI(\mathcal{F}'_1, \mathcal{F}'_2) = \frac{}{k(m-k)} \quad (4)$$

In (4), $\mathcal{F}'_1$ and $\mathcal{F}'_2$ are subset of features chose in consequent cycles of feature selection algorithms, k is number of features in the subsets and m is the aggregate number of features in test 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 moves toward becoming 1 when the cardinality of the crossing point set equivalents k, i.e., $\mathcal{F}'1$ and $\mathcal{F}'_2$ are indistinguishable. KI turns out to be near zero for uniquely drawn lists of subset of features.

7. 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 gotten from the KEEL dataset repository [32]. Table 1 demonstrates the attributes 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	138591	9911
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 criticalness of an attribute by estimating the information gain with worship 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 distinguished 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 last domain value of the chose attribute has altered for 100% privacy preservation thus a gatecrasher or noxious data miner even with significant background information can't make certain about the accuracy of a re-distinguishing proof.

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 analysis is BestFirst. CFS algorithm is filter-based, so it doesn't connect with any classifier in the selection procedure. Overfitting is decreased by utilizing 10-fold cross validation. BestFirst utilizes greedy hillclimbing for looking through the space of attribute subsets and is enhanced with a backtracking facility. BestFirst may seek in reverse in the wake of beginning with the full set of attributes or hunt forward in the wake of beginning with the empty set of attributes or pursuit in the two headings in the wake of

beginning anytime by considering all conceivable single attribute increments and cancellations at a predefined point. The quantity of chose features was kept at ideal number as selection stability will enhance up to the ideal number of pertinent features and after that abatements.

The statistical properties, i.e., mean, variance and standard deviation for the numerical attributes of original dataset and adjusted dataset have been ascertained. The analyses for statistical exhibitions have been directed on the privacy preserved datasets for confirmation. The Fig.1 demonstrates that privacy preserving perturbation has shaped nearly the indistinguishable statistical outcomes after the change too.

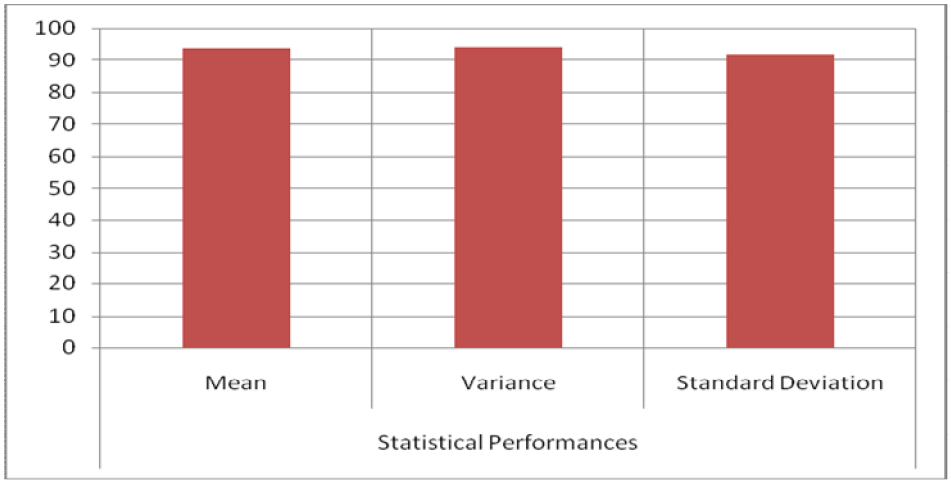

Fig. 1. Statistical performances of the privacy preserving algorithm

The feature selection stability values of the privacy preserved datasets Census and Coil 2000 are computed utilizing the stability measure Kuncheva Index KI and the outcome is appeared in the Fig.2. On account of KI, the bigger value of cardinality won't influence the selection stability and thus it is utilized as a part of the analyses as a stability measure. The selection stability is contrarily corresponded with the variation of the dataset i.e., perturbation of the dataset samples. The privacy preserving algorithm has created relatively stable feature selection outcomes on account of the statistical properties for the numerical attributes of the perturbed datasets are unswerving. 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 stable than the dataset Census as it includes just numeric traits.

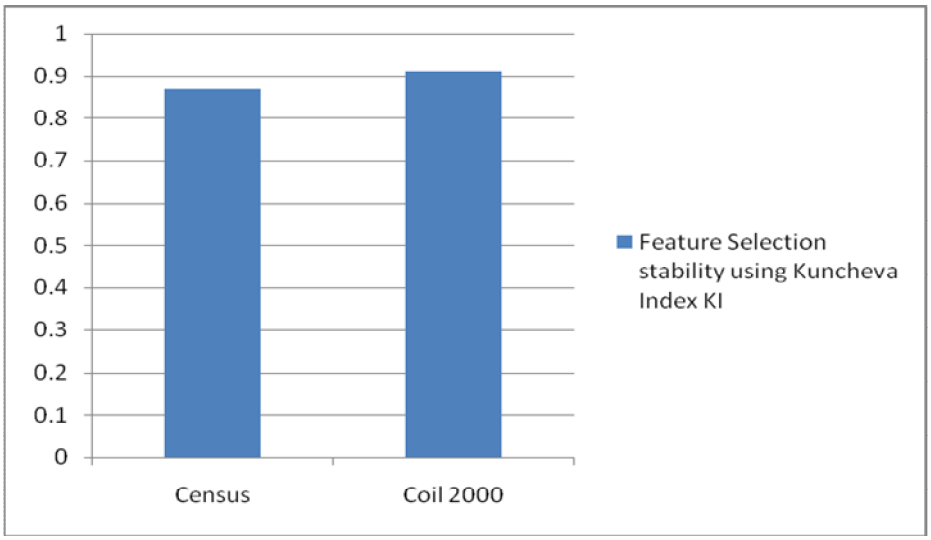

Fig. 2. Feature Selection stability using Kuncheva Index KI for the datasets Census and Coil 2000 after privacy preserving perturbation

Feature selection stability and data utility are emphatically corresponded. As the feature selection stability comes

about for the privacy preserving algorithm are great, the accuracy of the privacy preserved datasets are relatively same as before perturbation. The accuracy results are appeared in the Fig.3.

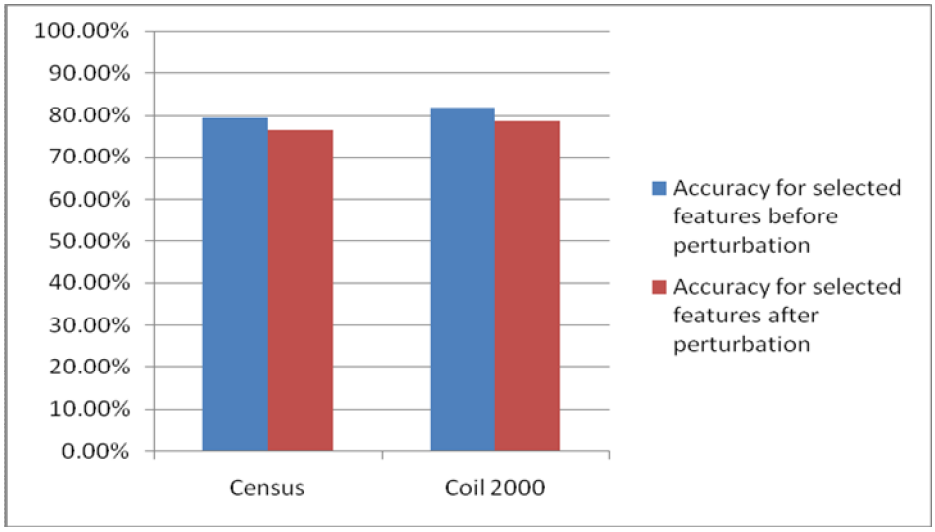

Fig. 3. Accuracy for selected features for the datasets Census and Coil 2000 before and after privacy preserving perturbation

In this way, the proposed privacy preserving algorithm has been tried utilizing two distinctive exploratory datasets for its performance in privacy preservation, feature selection stability and data utility. The exploratory outcomes have demonstrated that the utilization of the algorithm on test datasets result in stable feature selection with relatively reliable accuracy. The Table 2 abridges the statistics of the directed experiment on 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	0.88	0.92
Overall accuracy before perturbation	74.59%	76.89%
Overall accuracy after perturbation	71.79%	73.59%
Accuracy of selected features before perturbation	78.89%	82.16%
Accuracy of selected features after perturbation	75.98%	79.13%

8. CONCLUSION

The fundamental expectation 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 divulged to unapproved gatherings or interloper. Safeguarding the privacy-sensitive data of people and furthermore digging out supportive information from microdata is an extremely complicated issue. There will be tradeoffs between privacy preservation, feature election stability and accuracy. From the test comes about, it has been reasoned that the recommended privacy preserving algorithm which is used to perturb the quasi identifier attributes and sensitive attributes of the test datasets will save the privacy of the people. For the numerical attributes of the original and adjusted datasets, the statistical measures of mean, variance and standard deviation gave relatively comparative outcomes. The experiments have determined that the recommended privacy preserving algorithm gave relatively stable feature selection outcomes. In the meantime there will be least change in the accuracy because of the perturbation of the datasets. In this way, the recommended privacy preserving algorithm used in the analyses has safeguarded the privacy of the people and in the meantime gave great feature selection stability and furthermore the decline in accuracy is relatively unimportant.

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] 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] S.V. Vassilios , B. Elisa, N.F. Igor, P.P. Loredana, S. Yucel and T. Yannis, 2004, “State of the Art in Privacy Preserving Data Mining” Published in SIGMOD Record, 33, 2004, pp: 50-57.
[18] Gayatri Nayak, Swagatika Devi, "A survey on Privacy Preserving Data Mining: Approaches and Techniques", ternational Journal of Engineering Science and Technology, Vol. 3 No. 3, 2127-2133, 2011.
[19] Wang P, "Survey on Privacy preserving data mining", International Journal of Digital Content Technology and its Applications, Vol. 4, No. 9, 2010.
[20] Dharmendra Thakur and Prof. Hitesh Gupta,” An Exemplary Study of Privacy Preserving Association Rule Mining Techniques”, P.C.S.T., BHOPAL C.S Dept, P.C.S.T., BHOPAL India, International Journal of Advanced Research in Computer Science and Software Engineering ,vol.3 issue 11,2013.

[21] C.V.Nithya and A.Jeyasree,"Privacy Preserving Using Direct and Indirect Discrimination Rule Method", Vivekanandha College of Technology for WomenNamakkal India, International Journal of Advanced Research in Computer Science and Software Engineering ,vol.3 issue 12,2013.
[22] Sweeney L, "Achieving k-Anonymity privacy protection uses generalization and suppression" International journal of Uncertainty, Fuzziness and Knowledge based systems, 10(5), 571-588, 2002.
[23] Charu C. Aggarwal, Philip S. Yu "Privacy-Preserving Data Mining Models and algorithm"advances in database systems 2008 Springer Science, Business Media, LLC.
[24] Y. Lindell and B. Pinkas, "Privacy Preserving Data Mining", Journal of Cryptology, 15(3), pp.36-54, 2000.
[25] Helger Lipmaa," Cryptographic Techniques in Privacy-Preserving Data Mining", University College London, Estonian Tutorial 2007.
[26] Aggarwal C, Philip S Yu, "A condensation approach to privacy preserving data mining", EDBT, 183-199, 2004.
[27] Benny Pinkas, "Cryptographic Techniques for Privacy preserving data mining", SIGKDD Explorations, Vol. 4, Issue 2, 12-19, 2002.
[28] Ciriani, V., De Capitani di Vimercati, S., Foresti, S., and Samarati Universitua degli Studi di Milano, P.: Micro data protection. 26013 Crema, ltalia., Springer US, Advances in Information Security (2007)
[29] 29. 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)
[30] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, Muthuramakrishnan Venkita Subramanian. : ℓ-Diversity: Privacy Beyond K-Anonymity. Proc. International conference on Data Engineering. (ICDE), 24 (2006)
[31] 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)
[32] 32. 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)

AUTHORS BIOGRAPHY

Mr. P. Mohana Chelvan is currently working as an Assistant Professor in Department of Computer Science at Hindustan College of Arts and Science, Chennai, India. His educational qualifications are MCA, NIELIT C Level (IT), MPhil. (CS) and UGC NET. He is currently Ph.D. research scholar in computer science from Madurai Kamaraj University, Madurai, India in the area of Privacy Preserving Data Mining.

Dr. K. Perumal working as an Associate Professor in Department of Computer Applications at Madurai Kamaraj University, Madurai, India since 1990. He awarded his Ph.D. degree in computer science from Madurai Kamaraj University in the area of Digital image processing. He has contributed more than 50 papers in the International Journals and Conferences and also editor of proceedings for National Conference on Contemporary Developments in Information and Communication Technologies. He has guiding 9 scholars. His research interest includes Data Mining, Big Data and Image Processing especially in Medical Image Processing.