On the Relationship between Feature Selection Stability and Change in Statistical Properties of Datasets

Mohana Chelvan P¹, Dr. Perumal K²

¹Assistant Professor, Department of Computer Science, Hindustan College of Arts and Science, Chennai – 603 103, India.
²Associate Professor, Department of Computer Applications, Madurai Kamaraj University, Madurai – 625 021, India.

ABSTRACT:

Every organization stores massive volume of data about individuals due to the day-to-day online transactional activities. Data mining extracts previously not known knowledge from gigantic amount of stored operational data of organizations which can be utilized for administrative decision making. The datasets are for the most part high dimensional because of the headways in data and communication technologies. Feature selection is an imperative dimensionality reduction technique to deal with the “curse of dimensionality”. The subset of features selected in consequent iterations of feature selection algorithms must be same or at least similar for the small perturbations of the experimental dataset. The sturdiness of feature selection algorithms is called as the feature selection stability. High data quality with security/privacy is the foremost necessity of good privacy preserving data mining technique. This paper gives a description of the a range of privacy preserving data mining techniques and moreover explores the amend in statistical properties of datasets due to perturbations of datasets by the privacy preserving data mining techniques and their effects in feature selection stability.

Keywords: selection stability; feature selection; stability measures; privacy preserving; data publishing; data mining; Kuncheva Index

1. INTRODUCTION

There will be advancements in internet technologies which results in the accumulation of high dimensional microdata in the organizations. In data mining, small subset of relevant features works better than full set of features [1]. Feature selection is the dimensionality reduction technique which selects small subset of relevant features and eliminates irrelevant and redundant features. Feature selection improves accuracy, efficiency and model interpretability in data mining. Selection stability is the robustness of the feature selection algorithms i.e., the subsequent iterations of the feature selection algorithm must select same or similar subset of features and otherwise the researcher would not rely on his experimental results. Unstable feature selection reduces the confidence of the researcher as the research becomes inconclusive. Selection stability becomes an important criterion for feature selection algorithms and recently it becomes an emerging topic of research. Selection stability is mostly dataset dependant but is not completely algorithmic independent [2] [3].
The organizations gather, store, process and analyze huge amount of individual data for their business needs and at the same time they must protect the privacy of individuals. The privacy preserving data mining techniques will perturb the datasets for preserving the privacy of the individuals. The transformation of datasets for preserving privacy of the individuals results in change in statistical properties of numerical fields. It will result in alteration in selection stability as the selection stability is mostly dataset dependent. There is a positive correlation between privacy preserving data mining which results in change in statistical properties of the dataset due to the perturbation of attribute values and the feature selection stability.

2. PRIVACY PRESERVING DATA MINING TECHNIQUES

Data mining is imperative for the organizations for getting edge over the contenders and the information mined from the huge stockpiling of transactional data can be utilized for decision making. The information got from client data by data mining is monetarily significant for the organizations and in the meantime securing sensitive data of people is additionally essential since information in crude shape regularly contain touchy data about people. Divulgence of the delicate data like illness, MasterCard points of interest, and so forth., will ruins the notoriety of the organization and furthermore may brings about substantial budgetary misfortunes for the people. Despite the fact that the distinguishing qualities like ID numbers and names can be expelled from the dataset without influencing most data mining, sensitive data about people may in any case spill because of linking attacks that depend on the quasi-identifiers a.k.a., public attributes. Such linking attacks may join the quasi-identifiers of a distributed table with freely available different tables like a public datasets, and by which uncover private data of particular people. This has prompt two related research regions called as privacy preserving data mining (PPDM) and privacy preserving data publishing (PPDP).

A mathematical definition of privacy called as differential privacy offers the likelihood of accommodating these contending interests i.e., getting information from dataset by data mining and furthermore saving the privacy of the people. This privacy preserving technique is imagined by the researcher of Microsoft Research Lab, named as Cynthia Dwork in 2006. Differential privacy can make us conceivable to take in the general qualities of populations and in the meantime ensuring the privacy of any person's records. Differential privacy is an applied condition on the discharge instrument and not on the dataset. Differential privacy tends to the contradiction of learning valuable information about a population and in the meantime nothing around a person.

Privacy preserving data mining alludes to the area of data mining that looks to shield privacy-sensitive information from unsanctioned or spontaneous divulgence thus ensuring individual data records and their privacy. This strategy secures privacy of the people and in the meantime helpful information can be extricated from the data. There are a few techniques which can be utilized to empower privacy preserving data mining. The vast majority of the methods utilize some type of change or alteration. These strategies alter the gathered dataset trying to shield individual records from being re-recognized. An intruder or malicious data
miner can't be sure about the rightness of a re-recognizable proof even with extra learning. There are number of methods for doing data change, for example, suppression, swapping, aggregation and noise addition. For the most part, data publishing strategies, for example, k-anonymity, l-diversity, t-closeness and different anonymous release utilize generalization methods, which decrease accuracy and data utility to a great extent. Microdata publishing incorporate a few privacy preserving methodologies like randomization, sampling, suppression, swapping and perturbation [5]. Suppression replaces the recognizing characteristics with a few images like "*". Generalization changes the quasi identifier attributes in each bucket into summed up values. The basic dangers for microdata publishing incorporate attribute disclosure, membership disclosure and identity disclosure. The best approach to shield the data from the dangers is anonymising the data. The coordinating various buckets were imperative for securing attributes disclosure. An identity of a person through different attacks has been induced by membership information. The selection criteria which were not sensitive attributes will prompts the membership disclosure. The exposure of identifying attributes of the distinctive individual is called as identity disclosure [6].

2.1. THE K-ANONIMITY MODEL

The k-anonymity property will be possessed by the anonymous data. In this technique the published records can't be re-recognized and in the meantime the data can for all intents and purposes important. A release of data will have the k-anonymity property if the data for every individual incorporated into the release can't be separated from in any other k-1 individuals whose data in addition shows up in the release [6]. There will be two techniques to accomplish k-anonymity as takes after:

- Suppression: In this technique, a few attributes are substituted with values like "*". All or a few estimations of general are replaced by "*".
- Generalization: In this technique, individual values of attributes are substituted by more extensive classification.

This model is developed because of the probability of backhanded distinguishing identification of records from public databases. It is possible by the adventure of blend of records ascribes to recognize individual record. In this technique there will be lessening of the granularity of information portrayal by generalization and suppression. This granularity is diminished enough that any given record will be mapped into at least k different records of data.

2.2. THE L-DIVERSITY MODEL

For safeguarding privacy in datasets, the l-diversity method is the strategy for group based anonymization that is used by diving the granularity of a data representation. There will be a trade off among the outcomes in some loss of effectiveness of data management or mining algorithms and to get some privacy. This model is an augmentation of the k-anonymity model. Here it decreases the granularity of data representation utilizing procedures together with generalization and suppression to such an extent that any given record maps onto in any other k different records in the data. Especially when the sensitive values inside a group uncover homogeneity, the l-diversity model shows has the advantage by which a portion of the imperfections in the k-anonymity model where protected characters to the level of k-individuals isn't comparable to protecting the relating sensitive values that were generalized or suppressed. The l-diversity model will put in the help of intra-group diversity for sensitive attributes in the anonymization strategy.
The k-anonymity is helpless against various attacks in which background knowledge learning to an assailant will influence the attacks to happen to much more useful. The sorts of such attacks are as per the following.

- **Homogeneity Attack:** In this sort of attack, there will be the values for a sensitive data contained by an arrangement of k records are vague. In such cases, the sensitive value for the arrangement of k records might be precisely conceived.

- **Background Knowledge Attack:** At least one quasi-identifier attribute is associated with the sensitive attribute and this will lessen the arrangement of conceivable values for the sensitive attribute. For example, to limit the scope of values for a sensitive attribute of a patient's illness will be used by realizing that heart assaults occur at a lessened rate in Japanese patients.

The l-diversity method was delivered to assist k-anonymity by in addition saving the assorted variety of sensitive fields as the sensitive attributes might be construed for k-anonymity data [5].

In the class that there are in any class l “well-represented” values for the sensitive attribute, at that point a equivalence class is said to have l-diversity. In the class that each equivalence class of the dataset has l-diversity, at that point the dataset should have l-diversity.

### 2.3. THE T-CLOSENESS MODEL

The t-closeness is an improvement of l-diversity group based anonymization that is used to protect privacy in data sets by diving the granularity of a data representation. In this method there will be a trade off that outcome in some loss of effectiveness of data management or mining algorithms keeping in mind the end goal to increase some privacy. The t-closeness model is the expansion of the l-diversity model by treating the values of an attribute plainly by considering the dissemination of data values for that attribute.

The attribute values will be semantically comparative or especially skewed on account of real world data sets; the value dispersions may cause multifaceted nature in making practical l-diverse representations. The l-diversity method will be profitable in that it might hamper an attacker utilizing the worldwide conveyance of an attribute's data values so as to construe information about sensitive data values. In genuine datasets, only one out of every odd value may uncover attributes with sensitivity. The l-diversity might be hard and avoidable to fulfil while ensuring against attribute disclosure [7]. Here sensitive information uncovering may happen in light of the fact that it doesn't distinguish that values might be the semantically close while l-diversity necessity guarantees “diversity” of sensitive values in each gathering. For example, an attacker could translate a stomach illness applies to an individual if a specimen including the individual just recorded three distinctive stomach infections.

The distribution of values for l-diverse data brings about the reasoning of sensitive attributes and by besides saving the dissemination of sensitive fields, the t-closeness technique was produced to advance l-diversity. Privacy beyond k-anonymity and l-diversity portrays t-closeness.

In the t-closeness standard, an equivalence class should have t-closeness if the separation among the dissemination of a sensitive attribute in the class and the distribution of the attribute in the entire table is close to a threshold t. At this crossroads, a table should have t-closeness if all equivalence classes have t-closeness.
2.4. THE SLICING MODEL

For privacy preserving microdata publishing, there will be various anonymization systems including generalization and bucketization has been used. As of late, it has been discovered that on account of high-dimensional data, generalization loses generous amount of information. On the off chance that there have been no certain segment between quasi-identifying attributes and sensitive attributes, the bucketization technique does not matter for that data and besides does not avoid membership disclosure. Another strategy called as slicing gives better answer for the above issues by partitioning the data both on a level plane and vertically. From the exploration tests it have been exhibited that slicing preferred data utility over generalization and can be used for membership disclosure protection and moreover the slicing strategy can deal with high-dimensional data [8]. On account of slicing strategy, the profoundly related group of attributes are safeguarded for better data utility and uncorrelated gathering of data are sliced both on a level plane and vertically to protect privacy of the published microdata.

3. FEATURE SELECTION ALGORITHMS

The process of feature selection is mostly based on the three approaches viz. filter, wrapper and embedded [9]. The filter approach of feature selection is by removing features on some criteria or measures and in this approach, the goodness of a feature is evaluated using intrinsic or statistical properties of the dataset. A feature is adjudged as the most suitable feature based on these properties, and is selected for machine learning or data mining applications [10]. In the wrapper approach the subset of features is generated and then goodness of subset is evaluated using some classifier. The purpose of some classifier in this approach is to rank the features in the dataset and based on this rank, a feature is selected for the required application. The embedded approach tries to make use of the advantages of both the filter and wrapper methods. The main idea behind these algorithms is the reduction of search space for a wrapper approach by the filter approach.

3.1. CORRELATION-BASED FEATURE SELECTION CFS

The worth of a subset of attributes is evaluated by CFS by considering the degree of redundancy between them along with the individual predictive ability of each feature. Subsets of features that are having low inter-correlation between the classes but that are highly correlated within the class are preferred [11]. CFS determines the best feature subset and can be combined with search strategies such as genetic search, best-first search, backward elimination, forward selection and bi-directional search. Authors have GA as search method with CFS as fitness function. CFS is given in (1).

\[ r_{sc} = \frac{k r_{zi}}{\sqrt{k + (k - 1) r_{ii}}} \]  \hspace{1cm} (1)

where \( r_{sc} \) is the correlation between the class variable and the summed subset features, \( k \) is the number of subset features, \( r_{zi} \) is the average of the correlations between the class variable and the subset features and \( r_{ii} \) is the average inter-correlation between subset features [11].

4. SELECTION STABILITY MEASURES

The three main categories of the stability measures based on the representation of the output of the selection method are stability by index, stability by rank and stability by weight. Let \( A \) and \( B \) be subsets of features and \( A, B \subset X \), of the same size or cardinality, \( k \) and let \( r = |A \cap B| \) be the cardinality of the intersection of the two subsets. The important properties for each stability measurement [12] are as follows:
• **Monotonicity.** For a number of features, m and fixed subset size, k, and the larger the intersection between the subsets, the higher the value of the consistency index.

• **Limits.** The index should always be bound by constants which do not depend on m or k. The maximum value of the limit should be reached when the two subsets are identical, i.e., when r = k.

• **Correction for chance.** The index should always have a constant value in which the independently drawn subsets of features will have same cardinality, k.

In addition to these requirements, there are some important properties which have impact on the selection stability result should be taken into consideration [2] [3].

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

4.1. KUNCHEVA INDEX KI

In most of the stability measures, there will be overlap between the two subsets of the features due to chance. The larger cardinality of the selected features' lists positively correlated with the chance of overlap. To overcome this drawback, the Kuncheva Index KI which is proposed in [12] contains correction term to avoid the intersection by chance. The stability index is a measure based on correction for chance and cardinality of the intersection. KI is the only measurement that obeys all the requirements appeared in [12] i.e., Monotonicity, Limits and Correction for chance. The correction for chance term was introduced in KI and so it becomes desirable. On comparing selection stability measures, KI is the desirable measure because of the correction of the chance term and so larger value of cardinality will not affect stability.

\[
KI (F'_1, F'_2) = \frac{|F'_1 \cap F'_2| \cdot m - k^2}{k \cdot (m - k)}
\]  

(2)

In (2), \(F'_1\) and \(F'_2\) are the subset of selected features in subsequent iterations of feature selection, k is the number of selected features and m is the total number of features of the experimental dataset. KI's results bounds between the ranges of \([-1, 1]\), where −1 means k = m/2, i.e., there is no intersection between the two subsets of features. KI becomes 1 when the cardinality of the intersection set equals k, i.e., \(F'_1\) and \(F'_2\) are identical. KI becomes close to zero for dissimilarly drawn lists of subset of features.

5. EXPERIMENTAL RESULTS

The two datasets used in the experiments are Census-Income (KDD) dataset and Insurance Company Benchmark (COIL 2000) dataset. The datasets are obtained from the KEEL dataset repository [13]. Table 1 shows the characteristics of the datasets. In the listed datasets, the Census dataset has both categorical and numeric values while the Coil 2000 dataset has only numeric values.

The privacy preserving data mining techniques used in the experiments are data publishing techniques i.e., k-anonymity, l-diversity, t-closeness and slicing. The data sets are perturbed by the application of the privacy preserving techniques on the experimental datasets. This modification on the datasets affects the statistical properties of the numerical fields which
Table 1. Characteristics of The Datasets Census And Coil 2000

| S. No. | Dataset Characteristics | Dataset |
|--------|-------------------------|---------|
|        |                         | 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          | Categorical, Numerical | Numerical |

will leads to change in feature selection stability as selection stability is mostly dataset dependent. The feature selection was done before and after perturbation of the dataset by privacy preservation and feature selection stability was measured using the measure Kuncheva index. The feature selection was done by the attribute evaluator CFS and the search method BestFirst. The BestFirst method searches the space of attribute subsets by greedy hillclimbing improved with a backtracking facility. BestFirst may search forward after starting with the empty set of attributes or search backward after starting with the full set of attributes or search in both directions by considering all possible single attribute deletions and additions at a given point after starting at any point.

The result of the experiments are summarised in Table 2 for census dataset and in Table 3 for coil 2000 dataset. From the analysis of the results, it has been proved that there is positive correlation between change in statistical properties of the dataset and feature selection stability. For getting stable feature selection, the researcher must also consider for minimum change in statistical properties of the numerical fields of the dataset. The experiments have proved that statistical properties of experimental datasets are most affected by k-anonymity and least affected by the slicing technique which will result in unstable selection stability values for k-anonymity and stable selection stability values for slicing technique. From the results, it has been proved that slicing is the better technique of data publishing in comparing with k-anonymity, l-diversity and t-closeness in terms of feature selection stability as the change in statistical properties is at minimum for numerical attribute values. If there are more changes in the statistical properties of the experimental datasets due to privacy preservation, the corresponding selection stability measures becomes lesser and so there will be positive correlation between change in statistical properties.
Table 2. Stability of Statistical Properties in % and Kuncheva Index Values for Census Dataset

| Data Publishing Techniques | Stability in % for census dataset in terms of statistical properties measures after perturbation | Kuncheva Index |
|---------------------------|-------------------------------------------------------------------------------------------------|----------------|
|                           | Mean   | Standard Deviation | Variance |                        |
| k-anonymity               | 84.63  | 81.25              | 58.99    | 0.638                   |
| l-diversity               | 86.14  | 82.36              | 61.60    | 0.673                   |
| t-closeness               | 87.97  | 83.34              | 63.89    | 0.712                   |
| Slicing                   | 94.82  | 90.48              | 80.05    | 0.941                   |

Table 3. Stability of Statistical Properties in % And Kuncheva Index Values for Coil 2000 Dataset

| Data Publishing Techniques | Stability in % for coil 2000 dataset in terms of statistical properties measures after perturbation | Kuncheva Index |
|---------------------------|-------------------------------------------------------------------------------------------------|----------------|
|                           | Mean   | Standard Deviation | Variance |                        |
| k-anonymity               | 78.83  | 75.00              | 43.75    | 0.618                   |
| l-diversity               | 81.05  | 76.68              | 47.41    | 0.631                   |
| t-closeness               | 83.82  | 78.95              | 53.38    | 0.722                   |
| Slicing                   | 91.51  | 87.50              | 73.44    | 0.926                   |

of the dataset and feature selection stability. So, the privacy preserving data mining techniques should preserve the privacy of the individuals and at the same time the change in statistical properties of the numerical attributes of the dataset should be at minimum for better feature selection stability and data utility. It has been graphically depicted in Fig. 1 and Fig. 2 which will show the change in statistical properties and the feature selection stability measures of data publishing techniques for census and coil 2000 datasets.
6. CONCLUSION

The statistical properties of the datasets will be changed by the appliance of the privacy preserving data mining techniques. There will be positive correlation among amendment in statistical properties of the datasets and feature selection stability. So, when the privacy preserving data mining techniques are applied on datasets, the change in statistical properties of the numerical attribute values of the datasets should be at minimum to get stable feature selection and moreover for better data utility. Privacy preservation and data utility are negatively correlated as more perturbation of the dataset for better privacy preservation results in lower data utility. The tradeoff between privacy preservation, change in statistical...
properties of the numerical attributes, feature selection stability and data utility is an interesting research problem.

REFERENCE

[1] A. K. Jain and B. Chandrasekaran. 1982, Dimensionality and sample size considerations in pattern recognition practice. In P. R. Krishnaiah and L. N. Kanal, editors, Handbook of Statistics, 835–855.

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

[3] Salem Alelyani, Zheng Zhao, Huan Liu, 2011, 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.

[4] R. Agrawal and R. Srikant. Privacy-preserving data mining. In ACM-SIGMOD Conference on Management of Data, pages 439-450, May 2000.

[5] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, Muthuramakrishnan Venkita Subramanian, 2006, “ t-Diversity : Privacy Beyond K-Anonymity”, Proc. International conference on Data Engineering.(ICDE), 24.

[6] Charu C. Aggarwal, 2005, “On k-Anonymity and the Curse of Dimensionality”, Proceedings of the 31st VLDB Conference, Trondheim, Norway, 901-909.

[7] 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), 1(8): 28-33 (2012)

[8] Tiancheng Li, Jian Zhang, Ian Molloy, “Slicing: A New Approach for Privacy Preserving Data Publishing”, IEEE Transaction on KDD, Vol. 24, No. 3, pp. 561-574, 2012.

[9] K. Sudha, J. JебamalarTamilselvi, "A Review of Feature Selection Algorithms for Data Mining Techniques", International Journal on Computer Science and Engineering (IJCSE) ISSN: 0975-3397, Vol. 7, No.6, pp. 63-67, June 2015.

[10]K. Mani, P. Kalpana, “A review on filter based feature selection”, International Journal of Innovative Research in Computer and Communication Engineering (IJIRCE) ISSN: 2320-9801, Vol. 4, Issue 5, May 2016

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

[12]L. I. Kuncheva, 2007, 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.

[13]J. Alcalá-Fdez, A. Fernández, J. Luengo, J. Derrac, S. García, L. Sánchez, and F. Herrera. (2010) “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.
AUTHORS BIOGRAPHY

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 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.