*Impact Factor (SJIF): 3.632*

**International Journal of Advance Research in Engineering, Science & Technology**

*e-ISSN: 2393-9877, p-ISSN: 2394-2444*

***(Special Issue for ITECE 2016)***

# A Review on l-diversity with clustering in privacy preserving

**Dharmik Vasiyani[1], Jay Gandhi[2]**

[1]*Computer Engineering,B.H.Gardi College Of Engineering and Technology*
[2] *Computer Engineering,B.H.Gardi College Of Engineering and Technology*

***Abstract — detailed personal data is regularly collected and sharing in very huge amount. All of these data are useful for data mining application. These all data include shopping habits, criminal records, medical history, credit records etc. On one side such information is a very useful resource for business association and governments for choice analyzing. On the other side privacy regulations and other privacy concerns may prevent data owners from sharing information for data analysis. Two basic handling systems used to accomplish anonymization of a data set are generalization and suppression. Generalization refers to replacing a value with a less particular however semantically predictable value, while suppression refers to not releasing a value at all. This generalizes and suppression based Anonymity is good choice to hide personal information from the attackers but it also suffers from information loss. In propose method here use clustering approach to the l-diverse anonymity, this algorithm first clustering instances based on sensitive attribute and then arrange them with less information loss and as per the criteria of l-diversity. With this approach information loss is minimized. Proposed algorithm makes this approach efficient.***

***Keywords- Data mining; l-diversity; privacy preserving; k-anonymity; clustering.***

## I. INTRODUCTION

Knowledge Discovery in Databases (KDD) is a data mining process it is the method of recognizing genuine, novel, helpful, and justifiable examples from broad data sets. Data Mining is the focal point of the KDD methodology, including calculations that explore the data, make models, and find significant examples. Data mining has developed as an issue instrument for a wide blend of employments, running from national security to market examination. Large parcels of these applications include mining data that fuses private and delicate information about clients. For example, medicinal exploration might be coordinated by applying data mining calculations on patient therapeutic records to perceive illness outline.

Privacy preserving data mining is to make data mining frameworks without growing the threat of misuse of the data used to create those methods. The subject of privacy preserving data mining has been broadly inspected by the data mining bunch starting late. Different convincing schedules for privacy preserving data mining have been proposed. Most frameworks utilize some kind of progress on the main data to perform the privacy preservation. The changed dataset is made accessible for mining and should meet privacy necessities without losing the benefit of mining [1].

### 1.1 Classification Of Privacy Preserving Data Mining

There are numerous methodologies which have been received for privacy preserving data mining. We can classify them in view of the following measurements:

**Data distribution:** This dimension refers to the dispersion of data. A percentage of the methodologies have been developed for centralized data, while others refer to a scattered data situation. Circulated data situations can also be classified as horizontal data appropriation and vertical data circulation. Horizontal conveyance alludes to these situations where distinct database records stay in better places, while vertical data dissemination, refers to the situations where all the values for diverse characteristics live in better place.

**Data modification:** data modification is utilized as a piece of solicitation to change the original values of a database that should be released to the public and along these lines to ensure high privacy affirmation. It is vital that a data change technique should be working together with the privacy policy got by an association[1].

**Data mining algorithm:** This estimation indicates to the data mining algorithm, for which the data change is taking place. This is actually something that is not known ahead of time, notwithstanding it facilitates the analysis and design of the data concealing algorithm. We have included the problem of concealing data for a mix of data mining algorithms, into our future exploration plan. For the present, distinctive data mining algorithms have been considered in isolation of each other. Among them, the most essential considerations have been developed for classification data mining

algorithms, like choice tree inducers, association rule mining algorithms, clustering algorithms, rough sets and Bayesian network[1].

**Data or rule hiding:** The estimation refers to whether rough data or totaled data should be hidden. The complexity for covering up collected data as rules is obviously higher, and along these lines, mostly heuristics have been developed. The lessening of the measure of public data causes the data excavator to deliver weaker deduction rules that will not allow the induction of confidential values. This strategy is also known as "rule perplexity".

**Privacy preservation:** The estimation which is the most basic refers to the privacy assurance framework utilized for the selective change of the data. Selective change is required in order to accomplish higher utility for the balanced data given that the privacy is not risked.

The procedures that have been applied hence are:

Heuristic-based techniques like versatile change that modifies only selected values that minimize the utility loss instead of all available values.

Cryptography-based frameworks like secure multiparty handling where a calculation is secure if toward the end of the calculation no social affair knows anything except for its own info and the results[2].

Reconstruction based techniques where the original allotment of the data is reproduced from the randomized data It is central to realize that data conformity results in corruption of the database execution. To evaluate the debasement of the data, we mainly utilize two estimations. The main, measures the confidential data protection, while the second measures the loss of functionality.

Samarati and Sweeney proposed k-anonymity which is an anonymizing approach [2]. An information set consents to k-anonymity security if every individual's record set away in the released information set can't be perceived from in any occasion k - 1 individual whose information also appear in the information set. This security guarantees that the likelihood of recognizing an individual 2 centered around the 'released information in the information set does not surpass 1/k. Generalization and suppression are the most widely recognized techniques utilized for de-distinguishing proof of the information in k-anonymity-based calculations [3,4,5].

L-diversity is a kind of anonymization that is utilized to protect security in information sets by diminishing the granularity of an information representation. This lessening is a trade off those results in a couple loss of sufficiency of information organization or mining algorithms to build some protection. The l-diversity model is an expansion of the k-anonymity model which lessens the granularity of information representation utilizing procedures including generalization and concealment such that any given record maps onto at any rate k distinctive records in the information. The l-diversity model handles a rate of the weaknesses in the k-anonymity model where secured personalities to the level of k-individuals is not equivalent to guaranteeing the relating delicate values that were generalized or stifled, especially when the sensitive values inside a social occasion display homogeneity[6]. The l-diversity model includes the progression of intra-social affair diversity for delicate values in the anonymization part.

### 1.2 Problem Definition

Generalization and suppression based anonymity suffers from the information loss[7]. Clustering approach to the anonymity can reduce information loss[9]. Proposed approach uses clustering technique to anonymization to reduce information loss. This approach has less information loss with compare it to traditional l-diversity approach of anonymity.

In this paper section1 contains the introduction of privacy preserving data mining, section2 contains related work in anonymity with clustering, section 3 contains proposed work, section 4 contains conclusion, section 5 contains references.

## II. RELATED WORK

### 2.1 Clustering-Based k-Anonymity

Proposed algorithm first divide all tuples into more solid clusters efficiently and correctly. For clustering here used modified K means algorithm. After clustering then anonymized the dataset this approach applied to the dataset where number of QI attributes are less than three. For future work we can extend this technique where size of QI-attributes is larger than 3[7].

### 2.2 Systematic Clustering Method for l-diversity Model

Technique first grouping the similar data together as per the criteria of the l-diversity and then it anonymizes all of them individually. For clustering the instances, here used systematic clustering algorithm. This approach makes algorithm faster and for future we can show experimental study on efficiency and effectiveness of this algorithm [8].

### 2.3 L-diverse Anonymity Algorithm Based on Clustering Techniques

This algorithm is based on clustering techniques and carries out experimental verification on real dataset. With this approach we have less information loss. Proposed algorithm first create buckets of similar sensitive attributes then picks instance from the buckets such a way that they have less information loss and also as per the criteria of L-diverse algorithm. In future we can make this algorithm efficient [9].

### 2.4 An Enhanced l-Diversity Privacy Preservation

This algorithm based on clustering. It minimize information loss as well as assure data quality. It first updates the centroid of the cluster whenever a record is added to the cluster so that each centroid accurately reflects the current center of a cluster for quality improvement. Identical tuples are assigned to the same cluster. We can apply this approach to the real world examples [10].

### 2.5 Clustering-Based Frequency l-Diversity Anonymization

Bucket clustering proposed here to minimize information loss. Bucket clustering algorithm first creates buckets by gathering all tuples as per to their sensitive value then it selects one tuple from biggest bucket as a cluster center, then it chooses best diverse tuple to guarantee l-diverse principle with least information loss [11].

## III. PROPOSED WORK

Very few research works has done in l-diversity with clustering but still there is no any algorithms that have feasible results. L-diverse anonymity algorithm based on clustering techniques[9] was applied on real world data set and also have less information loss but it is not efficient in time complexity. We will make this algorithm efficient with our proposed algorithm of l-diversity with clustering.

Main purpose behind this approach is to give utility friendly anonymization. Suppression and generalization based l-diversity suffers from information loss[9]. With this approach the generalization and suppression techniques comes at the end of the algorithm so that we can get utility friendly version of l-diversity. Proposed algorithm first create buckets of similar sensitive attributes then picks instance from the buckets such a way that they have less information loss and also as per the criteria of L-diverse algorithm of sensitive attribute. After collecting the same sensitive attributes in buckets, then sorting them as per their size. Then k-mean clustering technique runs on them. Then after it picks the records from clusters and makes a set of instances as per criteria of l-diversity. When the cluster is created as per the l-diversity then after anonymized all the instances and after then measures the utility.

### 3.1 Flowchart of Proposed Algorithm

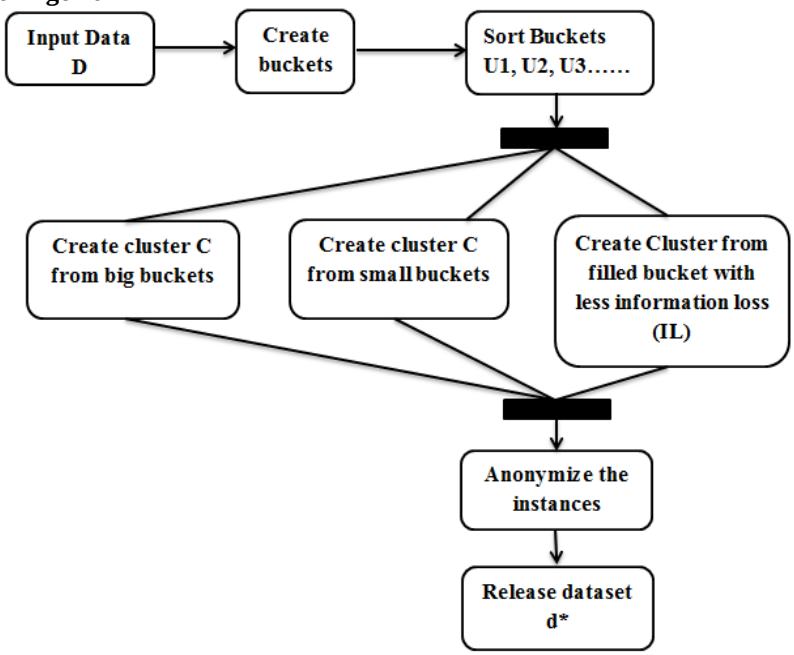


*Figure 1. Flowchart of algorithm*

Here gives the general description of proposed algorithm as per the flow diagram as under. First dataset d given as an input of this algorithm then it creates the buckets such a way that each buckets contains same sensitive attributes and then sorting those buckets as per their size. Second it picks the records from the big bucket u randomly and put them into the cluster c till the bucket is less than l, here l represents the number for diversity of sensitive values. Then it picks the record from the smaller buckets and till |c|<l put them into the cluster c. Then after it picks the record randomly from the cluster u till it becomes empty and put them into the cluster with less information loss. At the end anonymize the instances. That final cluster is l-diverse and also has less information loss.

**3.2 Proposed algorithm:**

Input: dataset D and a diverse anonymity threshold value l.
Output: dataset D*, l different sensitive attribute values of each equivalence class.
1.Built the set of buckets that contains same sensitive attribute
Values and sort them based on size U={u1,u2,u3,}.
2 if the number of buckets<l then return:
3. Assume result=$\Phi$.
4. Create Thread-1: Create cluster from big buckets
4.1 While (the number of filled buckets is
equal to or more than l)
4.1.1 Randomly pick a record $r_i$ from big
non empty bucket u and add it to
cluster c={ $r_i$};
4.1.2 u = u – { $r_i$ };
4.1.3 Create Thread-2: Create cluster from
small buckets
4.1.3.1 While (|c| < l);
a) pick a record $r_j$ from the small
filled bucket so that IL(c U { $r_j$ })
is minimum;
b) u = u – { $r_j$ };
5. Create Thread-3: Create Cluster from filled bucket with less
information loss(IL);
5.1 While (the number of filled buckets is more than zero)
5.1.1 Randomly pick a record $r_k$ from filled bucket u;
5.2.2 u = u – { $r_k$ };
5.2.3. pick a cluster c so that IL(c U { $r_k$});
5.2.4. c = c U { $r_k$ };
6. Produce anonymous equivalence class by applying local recoding techniques on each cluster.
7. Return a releasable anonymized dataset d*.

As per the proposed algorithm dataset D taken as a input and also the threshold value of l have to decide. In (line 1) creates the set of buckets such a way that each bucket contains same sensitive attribute and sort them as per their size. In (line 2) it checks the minimum requirements of this algorithm there should be more than l buckets. In (line 3) it takes one variable result and assumes it to empty. In (line 4) till filled buckets is equal to or more than l first it randomly takes a record $r_i$ from bigger filled bucket u and add it to the cluster c and till number of instances in cluster c less than zero it takes the records from small filled buckets and add it to the cluster so that the information loss is minimum. In (line 5) till the number of filled buckets is there it takes the records from filled buckets randomly and add it to the cluster that create less information loss with addition to that record. In (line 6) it anonymizes the instances. And at the end we have l-diverse dataset with less information loss.

## IV. CONCLUSION

Every privacy preserving techniques provides the privacy to micro data at different level of data mining. But that all have one common problem is information loss. Anonymity is good choice to hide personal information from the attackers but it also suffers from information loss. Generalize and suppression approach to the anonymity is mainly reasons for the information loss in anonymity. From this technique with use of clustering approach to l-diversity we can reduce information loss. And also here used thread to different process and that makes it efficient in terms of time.

## REFERENCES

[1] Pingshui WANG. "Survey on Privacy Preserving Data Mining". International Journal of Digital Content Technology and its Applications Volume 4, Number 9, December 2010.

[2] Rokach, Lior, Roni Romano, and Oded Maimon. "Negation recognition in medical narrative reports." Information Retrieval 11.6: 499-538, 2008.

[3] J. Uncertaillty, Fuzziness, and Knowledge-Based Systems, vol. 10, no.5, pp. 557-570, 2002.

[4] L. Sweeney, "Achieving k-Anonymity Privacy Protection Using Generalization and Suppression," int'l J. Uncertaillty, Fuzizness, and knowledge-Based Systems, vol. 10, no. 5, pp. 571-588, 2002.

[5] B.C.M. Fung, K. Wang, and P.S. Yu, "Top-Down Specialization for Information and Privacy Preservation," Proc. 21st IEEE Int'l Conf. Data Eng. (ICDE '05), pp. 205-216, Apr. 2005.

[6] K. Wang, p.s. Yu, and S. Chakraborty, "Bottom-Up Generalization:A Data Mining Solution to Privacy Protection," Proc. Foulrth IEEE int'l Conf. Data Mining, pp. 205-216, 2004.

[7] Xianmang He, HuaHui Chen, Yefang Chen,Yihong Dong, PengWang, and Zhenhua Huang"Clustering-Based k-Anonymity" Springer-Verlag Berlin Heidelberg 2012

[8] Md Enamul Kabir, Hua Wang, Elisa Bertino & Yunxiang Chi"Systematic Clustering Method for l-diversity Model" ACM 2010

[9] Pingshui Wang, Jiandong Wang "L-diverse Anonymity Algorithm Based on Clustering Techniques" Binary Information press-2012

[10] Gaoming Yang, Jingzhao Li, Shunxiang Zhang, Li Yu"An Enhanced l- Diversity Privacy Preservation" IEEE-2013

[11] Mohammad-Reza Zare-Mirakabad1, Aman Jantan2, and St´ephan Bressan3"Clustering-Based Frequency l-Diversity Anonymization" Springer-2009