© 2018 IJSRSET | Volume 4 | Issue 8 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section : Engineering and Technology

Heuristic Based Approach for Privacy Preserving in Data Mining

Patel Shreya[1], Aniket Patel[2]

[1]CE Department Silver oak College of Engineering & Technology, Gota, Ahmedabad, Gujarat, India

[2]IT Department, Silver oak College of Engineering & Technology, Gota, Ahmedabad, Gujarat, India

ABSTRACT

Data Mining has been the foremost researched space for researchers because of the possibilities of extension at every application of it. Once the information becomes massive in volume, several issues strike for security and privacy breach. If the data changes, it'd be mandatory to rescan the database that results to long computation time and inability to promptly reply to the user. Some applications like sharing of such knowledge to a selected user that have threats of preserving the original data so that the injection of such data can be prohibited. Thus it's a timely need to secure the information whereas handling them to the known or unknown users. Such troubles prompted the advancement of Privacy Preserving Data Mining (PPDM) Techniques. Primary objective is to accomplish harmony between privacy preservation and knowledge discovery and hiding the data from attacker. Privacy Preserving has become a crucial issue within the development progress of Data Mining techniques. Methods like k-anonymity along with the hybrid approach of l-diversity and t-closeness. Experimental outcomes shows that the approach not solely preserve's data privacy however one will get better accuracy with minimum loss of data.

Keywords: Data Mining, Privacy Preserving Data Mining, Heuristic Based Approach.

I. INTRODUCTION

All around the world-network community, there is greater demand by community for individual-specific data, yet the boundless availability of information makes it very difficult to release any information about people without breaching privacy. Data mining would be newly appearing field, connecting the 3 worlds of Databases, AI and Statistics. The data age has enforced many organizations to collect huge number of data. Yet, the utility of this data is unimportant if "meaningful information" or "knowledge" cannot be extracted from it. Data mining, otherwise known as data discovery, tries to answer this need. In comparison to standard analytical methods, data mining methods search for interesting information without challenging previous conclusion. Data mining tools conclude future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automatic, possible reasoning offered by data mining move on the far side of the reasoning of past actions enforced by recollecting tools typical of decision support systems. Data mining tools will answer business queries that usually were too time intense to resolve. They scour databases for hidden patterns, finding predictive information that specialist might miss as a results of it lies outside their expectations. Most organizations already collect and refine huge quantities of information. Data mining methods are enforced speedily on existing software and hardware platforms to reinforce the worth of existing information resources, and may be integrated with new merchandise and systems as they're brought on-line.

II. REQUIREMENT FOR PRIVACY IN DATA MINING

However, ancient data mining techniques and algorithms directly operated on the original data set, which is able to cause the leakage of privacy data. And in today's world data is the most significant resource. At an equivalent time, massive amounts of information. Implicates the sensitive knowledge that their disclosure can't be unnoticed to the aggressiveness of enterprise. These issues challenge the standard data mining, thus privacy-preserving data mining (PPDM) has become one in all the most recent trends in privacy and security and data mining analysis.

III. PRIVACY PRESERVING IN DATA MINING (PPDM)

Privacy means that it's the flexibility of an private or cluster of user to isolate themselves, or information concerning themselves, and thereby specify themselves by selection. Which contains Information privacy which is the right to have some control over how your personal information is collected and used. It's flexibility of a private or cluster of user to prevent information concerning themselves from turning into renowned to individuals aside from those they like better to offer the knowledge to. Privacy sometimes related to anonymity though it's usually most extraordinarily valued by people who are publicly known. Privacy is also seen as a reality of security. Various people have various mindset for privacy, for a few individuals personal information is privacy whereas for few individual just some of the sensitive attribute is privacy. In easy words, the Facebook user with one thousand friends and fifty cluster memberships could be a ton, a lot of possible to be injured by a breach than somebody who barely uses the site. Framework for PPDM is shown in Figure 1 and is performed in 3 levels. Where level1 is traditional DB, level2 is Data Mining algorithms are enforced on it for the generation of information/knowledge [1] and level3 shows the results.

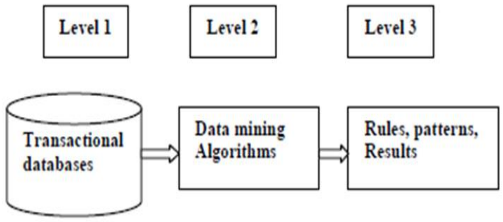


Figure 1. PPDM Framework [2]

PPDM is mostly used to extract knowledge from huge amount of data with the help of data hiding and rule hiding [3]. The PPDM algorithm are specifically on the tasks of classification, association rule mining and clustering classification [4].PPDM is a model used for sensitive data. The main goal is to keep the information private is to stop misuse the personal information. Once necessary information is disclosed then it's unimaginable to stop the misuse of knowledge. If data owner printed their data, they need worry of misuse. So, this prevents them to share their data.There are many different approaches based in Privacy Preserving in Data Mining basically the techniques are divided into three major groups such as Heuristic based approach, Reconstruction based approach and Cryptographic based approach which are as shown in the Figure 2.

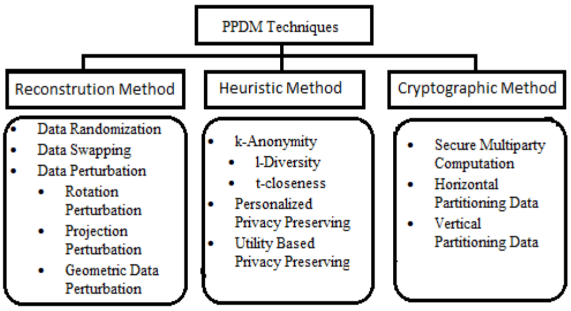


Figure 2. PPDM Techniques

IV. RELATED WORK

A. Heuristic Based Approach

A Heuristic Based mostly could be a technique designed for solving a problem more quickly once classic strategies are too slow, or for finding a depth solution when classic strategies don't find any actual solution. The goal of a heuristic based approach is to provide an answer in an exceedingly cheap time frame that is good enough for solving the problem that is important now. This solution might not be the best of all the particular solutions to this problem, but it is merely equivalent to the precise answer.

However it's still valuable because finding it does not require a way too much long time. This approach process the records in "**group based**" manner [5]. It protects the database by anonymizing the data so that the attacker cannot understand which data belongs to whom. By different anonymization the data is changed and it holds good enough utility and that can be released to other parties safely. This whole process is called as privacy-preserving data publishing. Information is hold on in the main within the tabular kind[6]. And mostly information is discovered in 2 ways which is microdata and macrodata. In past information are published mostly in precomputed statistical and tabular format. Such variety of information is called macrodata [6]. Numerous organizations (e.g., Medical authorities and Government agencies) are in need of releasing a someone's specific information which is often called micro data for public health researchers and numerical analysis [1].Database contains numerous kinds of attributes a group of non-sensitive attributes {Q1, …,Qp} of a table known as quasi-identifier if these attributes may be connected with external information to unambiguously establish (can be known as candidate key) a minimum of 1 individual inside the final population[7]. Age, Gender, State is a set of QI attributes.[8].Sensitive attributes example Medical records, salaries, etc and these attributes are what the researchers would like, so they are always released directly [9][15]. An attribute K consists of values that are most unique/original value for to identify the individual from set S. Denote by K. Key attributes which will used to identify a record, like Name and Social Security Number [5][7]. Equivalence Class (EC):- Each group that shares the same values on every QI for example Birthdate and Age [5]. While releasing the sensitive information it must required to preserve them from disclosure. There are mainly two kinds of Information Disclosure. Identity disclosure: a personal is connected to a selected record within the revealed information. Attribute disclosure: Sensitive attribute data of a personal is disclosed [7][10][11].

k-Anonymity: It is an anonymizing approach proposed by Samarati and Sweeney [12].This technique is employed for restricting the disclosure risk. K - anonymity necessities claims that, a data set is k anonymous (k 2: I) if every record within the data set is indistinguishable from a minimum of (k-l) alternative records among the constant data set [13]. This k-Anonymity demand is mostly acomplished by using generalization and suppression [14]. K-Anonymity includes mainly two types of attack Homogeneity attack and Background Knowledge attack.

Homogeneity attack:- Here all the value of sensitive attributes in an EC are same. So it is easy for the adversary to predict that the person is in which equivalence class[9].

Background Knowledge Attack:- Here attacker link the quasi-attribute which they know to the Sensitive attribute to get the information[9].

While k-Anonymity protects against identity disclosure, it is insufficient to prevent attribute disclosure [16]. To handle this limitation of k-Anonymity, Machanavajjhala et al. [17] recently introduced a replacement notion of privacy, known as L-Diversity.

l-Diversity: This method enhances K-Anonymity. This method removes specific/explicit identifiers and generalizes the QID values to make sure that the information users cannot specify every individual's sensitive values with a confidence larger than 1/l [18].This technique is used for maintaining the minimum size of k and for preventing the homogeneous attack. Machanavajjhala et al. [19] gave a no of judgement of the term "well represented" in this principle. Here two forms of attacks are addressed they're skewness attack and similarity attack.

Skewness Attack:- If a record have 1000 number of patient with and without cancer then that sensitive attribute is 2-diverse and there will be 50% of chances for the adversary to understand that whether that person have cancer or not.

Similarity Attack:- In a record if the value of sensitive attributes is l-diverse but semantically similar so there is chances of similarity attack.

t-closeness: An equivalence category is claimed to own t-closeness if the space between the distribution of a sensitive attribute during this category and therefore the distribution of the attribute within the whole table isn't any over a threshold t [19].k-Anonymity does not protect against attribute disclosure while t-closeness seeks to prevent attribute disclosure from happening [20]. The t parameter in t-closeness permits one to exchange between utility and privacy. There are mostly two ways to find the probability distribution, first is variational distance formula and second is earth mover distance formula. While EMD formula satisfies the two properties of t-closeness they are generalization and subset property.

B. Personalized privacy preservation

This technique performs the minimum generalization for satisfying everybody's needs, and thus, retains the most important quantity of data from the microdata [20].

C. Utility based privacy preservation

A utility based method generally capture 2 aspects: the data loss caused by the anonymization and therefore the priority of attributes. Such utility-aware anonymization might facilitate to boost the standard of research later [22].

V. FRAMEWORK

Here figure 3 describes the framework of heuristic based approach for privacy preserving data mining. The dataset D is used by the ARX tool where it is used to create Semantic Hierarchy tree of different attributes like sensitive attributes, quasi attributes etc… and the output is stored in R with the help of data mining system. Then with the help of heuristic based approach here hybrid approach is been used which is l-diversity and t-closeness on the attributes using semantic hierarchy tree. So one gets Dataset D' which with the help of ARX tool and Data mining system gives result R'. Hence at last compares the original result R with the produced result R'.

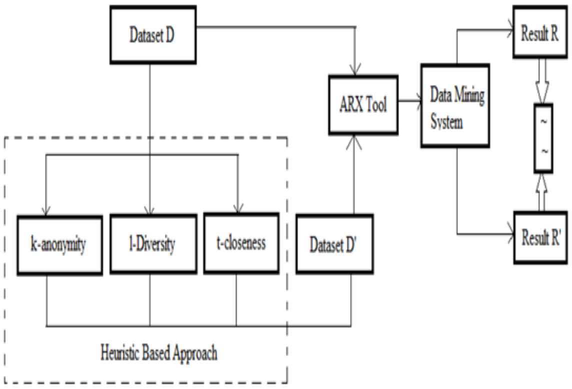


Figure 3. Framework for heuristic based approach for privacy preserving in data mining

ALGORITHM

Algorithm: l-Diversity added t-closeness algorithm based on privacy measurements.

Input: a releasable dataset Tn-1, an incremental dataset ΔTn−1, and a diversity threshold value l.

Output: a releasable dataset Tn, which ensures that each EC has the same sensitive attribute values set before and after update and has minimum information loss.

Step 1: Take dataset D.

Step 2: Convert the Quasi-identifiers into semantic Hierarchical tree for classification. Let t= {v1, v2,…, vk } be a tuple with k QI values and t'={ v1', v2',…, vk'} be a generalized tuple of t.

Step 3: Form initial equivalence class, S= E1, E2, E3,…, En}

Step 4: Check for Numerical & Non-Numerical attribute from the dataset D.

Step 5: If **Numerical Attribute** Found

1) Calculate EMD between two attribute

   Here P and Q are signature with m and n clusters having cluster representative Pi and Qj respectively. D= [di,j] be the ground distance and F=[fi,j] is the flow between Pi and Qj.

$$EMD(P,Q) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{i,j} d_{i,j}}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{i,j}}$$

2) Hellinger's Distance to calculate for 2 distributions. For two discrete probability distributions P and Q

$$1 - H^2(P,Q) = \sum_{i=1}^{k} (\sqrt{p_i q_i})$$

Step 6: If **Non-Numerical Attribute** Found

For multiple sensitive attributes **n** in S the smallest distance between EC1n and EC2n is consider using l-diversity with the entropy of an EC (E) which is defined as

$$Entropy(E) = -\sum_{s \in S} p(E,s) \log p(E,s)$$

in which S is the domain of the sensitive attribute, and p(E, s) is the fraction of records in E that have sensitive value s.

Step 7: MSA Calculation

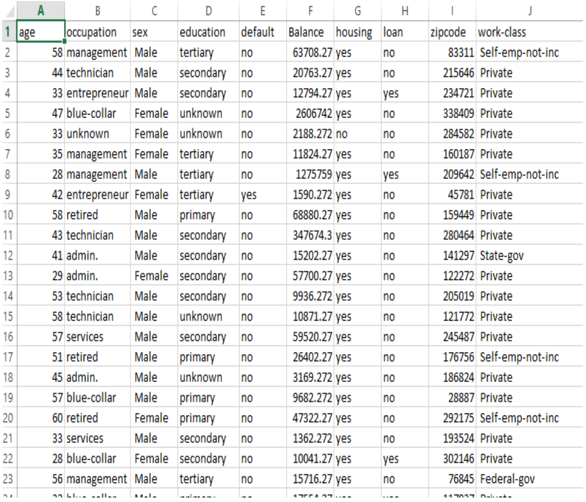
(Measurement System Analysis- Insert the record r into a selected candidate EC which results the minimal information loss which includes overall record data)

Step 8: Display Results

Experiments were performed to measure the accuracy of data while protecting the sensitive attributes at the same time. Here we use ARX Tool to calculate or measure the accuracy of data. This tool is one of the powerful anonymization tool. In this tool, analysis risk is a perspective where various analysis risks are measured on the data.It includes re-identification risks for the user that specify what is the measure of risk in disclosure of data.

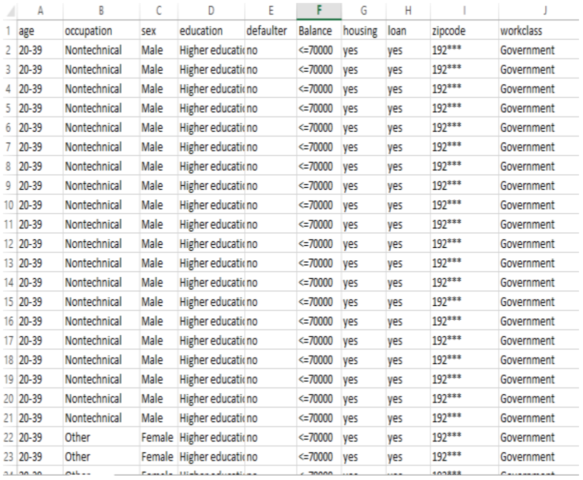
Table 1. Dataset attributes description

Dataset= 34000	
Attributes	Types
Age	Quasi-attribute
Occupation	Quasi-attribute
Sex	Quasi-attribute
Education	Quasi-attribute
Default	Insensitive attribute
Balance	Sensitive attribute
Housing	Sensitive attribute
Loan	Sensitive attribute
Zipcode	Quasi-attribute
Work-class	Quasi-attribute



	A	B	C	D	E	F	G	H	I	J
1	age	occupation	sex	education	default	Balance	housing	loan	zipcode	work-class
2	58	management	Male	tertiary	no	63708.27	yes	no	83311	Self-emp-not-inc
3	44	technician	Male	secondary	no	20763.27	yes	no	215646	Private
4	33	entrepreneur	Male	secondary	no	12794.27	yes	yes	234721	Private
5	47	blue-collar	Female	unknown	no	2606742	yes	no	338409	Private
6	33	unknown	Female	unknown	no	2188.272	no	no	284582	Private
7	35	management	Female	tertiary	no	11824.27	yes	no	160187	Private
8	28	management	Male	tertiary	no	1275759	yes	yes	209642	Self-emp-not-inc
9	42	entrepreneur	Female	tertiary	yes	1590.272	yes	no	45781	Private
10	58	retired	Male	primary	no	68880.27	yes	no	159449	Private
11	43	technician	Male	secondary	no	347674.3	yes	no	280464	Private
12	41	admin.	Male	secondary	no	15202.27	yes	no	141297	State-gov
13	29	admin.	Female	secondary	no	57700.27	yes	no	122272	Private
14	53	technician	Male	secondary	no	9936.272	yes	no	205019	Private
15	58	technician	Male	unknown	no	10871.27	yes	no	121772	Private
16	57	services	Male	secondary	no	59520.27	yes	no	245487	Private
17	51	retired	Male	primary	no	26402.27	yes	no	176756	Self-emp-not-inc
18	45	admin.	Male	unknown	no	3169.272	yes	no	186824	Private
19	57	blue-collar	Male	primary	no	9682.272	yes	no	28887	Private
20	60	retired	Female	primary	no	47322.27	yes	no	292175	Self-emp-not-inc
21	33	services	Male	secondary	no	1362.272	yes	no	193524	Private
22	28	blue-collar	Female	secondary	no	10041.27	yes	yes	302146	Private
23	56	management	Male	tertiary	no	15716.27	yes	no	76845	Federal-gov

Figure 4. Original Data D.



	A	B	C	D	E	F	G	H	I	J
1	age	occupation	sex	education	defaulter	Balance	housing	loan	zipcode	workclass
2	20-39	Nontechnical	Male	Higher educati	no	<=70000	yes	yes	192***	Government
3	20-39	Nontechnical	Male	Higher educati	no	<=70000	yes	yes	192***	Government
4	20-39	Nontechnical	Male	Higher educati	no	<=70000	yes	yes	192***	Government
5	20-39	Nontechnical	Male	Higher educati	no	<=70000	yes	yes	192***	Government
6	20-39	Nontechnical	Male	Higher educati	no	<=70000	yes	yes	192***	Government
7	20-39	Nontechnical	Male	Higher educati	no	<=70000	yes	yes	192***	Government
8	20-39	Nontechnical	Male	Higher educati	no	<=70000	yes	yes	192***	Government
9	20-39	Nontechnical	Male	Higher educati	no	<=70000	yes	yes	192***	Government
10	20-39	Nontechnical	Male	Higher educati	no	<=70000	yes	yes	192***	Government
11	20-39	Nontechnical	Male	Higher educati	no	<=70000	yes	yes	192***	Government
12	20-39	Nontechnical	Male	Higher educati	no	<=70000	yes	yes	192***	Government
13	20-39	Nontechnical	Male	Higher educati	no	<=70000	yes	yes	192***	Government
14	20-39	Nontechnical	Male	Higher educati	no	<=70000	yes	yes	192***	Government
15	20-39	Nontechnical	Male	Higher educati	no	<=70000	yes	yes	192***	Government
16	20-39	Nontechnical	Male	Higher educati	no	<=70000	yes	yes	192***	Government
17	20-39	Nontechnical	Male	Higher educati	no	<=70000	yes	yes	192***	Government
18	20-39	Nontechnical	Male	Higher educati	no	<=70000	yes	yes	192***	Government
19	20-39	Nontechnical	Male	Higher educati	no	<=70000	yes	yes	192***	Government
20	20-39	Nontechnical	Male	Higher educati	no	<=70000	yes	yes	192***	Government
21	20-39	Nontechnical	Male	Higher educati	no	<=70000	yes	yes	192***	Government
22	20-39	Other	Female	Higher educati	no	<=70000	yes	yes	192***	Government
23	20-39	Other	Female	Higher educati	no	<=70000	yes	yes	192***	Government

Figure 5. 3-Anonymized Data D'.

Re-identification risks

It is the practice of matching undisclosed data with publicly available information, in order to discover the individual to which the data belongs to. Re-identification risks may be analyzed based on sample characteristics or on the concept of uniqueness. Re-identification risk are defined by 3 models: Prosecutor model, Journalist model, Marketer model.

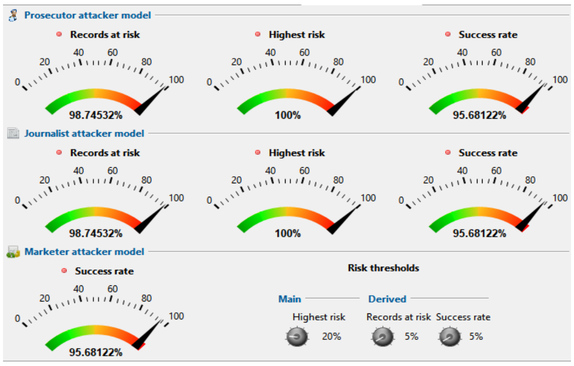


Figure 6: Re-identification risk of original data D

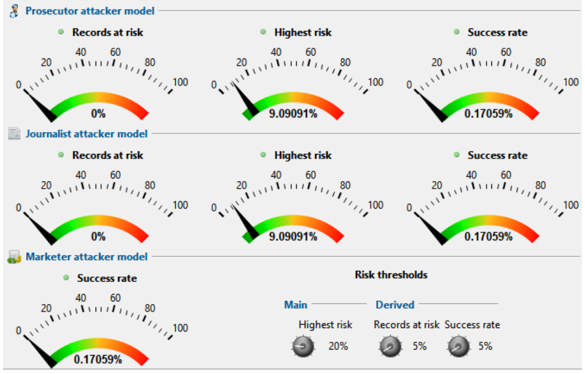


Figure 7. Re-identification risk of anonymized data **D'**

Distribution of equivalence class sizes

Here the distribution of sizes of equivalence classes can be analyzed. The distribution is represented for both input and output data, and shown as either histogram or table. Here the diagram shows the distribution regarding the equivalence class of age attribute of original data

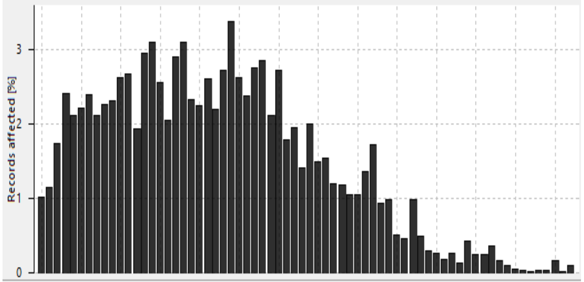


Figure 8. Distribution of class size of original data D

This figure shows the distribution of class after anonymization is performed. This distribution is shown after performing 3-anonymization. Once anonymization is performed it is represented in the range form and distribution is done accordingly so now attacker won't understand any individual's identity.

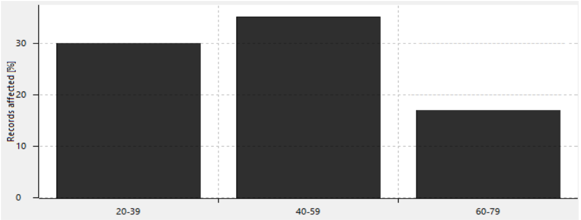


Figure 9. Distribution of class size of Anonymized data D'.

Risk estimation

This table shows the risk estimation of any individual record that the attacker can easily find any person record. Here the risk estimation of original data is shown.

Table 2. Risk estimation of original data D

Estimate	Value%
Lowest re-identification risk	0.23148
Average re-identification risk	95.6812
Highest re-identification risk	100.00
Individual Affected by High risk	92.7245
Entire Uniqueness within population	51.9950

This shows the risk estimation of 3- Anonymized data. And risk factor of 3-Anonymized data is much less than original data.

Table 3. Risk estimation of original data **D'**

Estimate	**Value%**
Lowest re-identification risk	0.03236
Average re-identification risk	0.17059
Highest re-identification risk	9.09091
Individual Affected by High risk	0.03235
Entire Uniqueness within population	0.0000

Population uniqueness

Population uniqueness means that the total records which are uncommon in the data or sample. Which are not always easy to verify but can with the help of simple program which is sufficient to decide sample uniqueness.

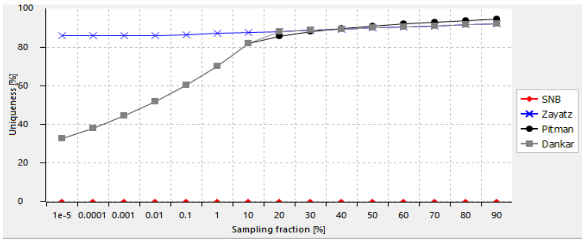


Figure 10. Analysis of Population uniqueness of original data D

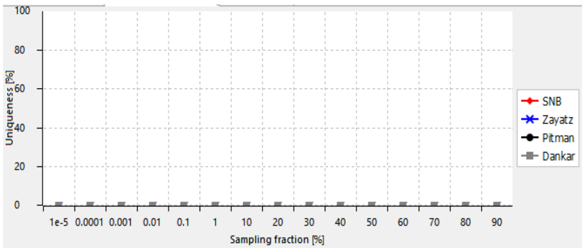


Figure 11. Analysis of Population uniqueness of Anonymized data D'.

V. CONCLUSION

Heuristic Based approach as an alternative method to privacy preserving data mining has been reviewed. The heuristic based technique includes k-anonymization method, l-diversity, t-closeness methods. Out of with it considers sensitive attribute as dependent attribute and remaining attributes of dataset except as independent attributes. Data mining desire to excerpt useful data from multiple sources, whereas privacy preservation in data mining deserves to preserve these data against disclosure or loss. Here we discovered that by eliminating the weakness of t-closeness and developing a new heuristic approach with k-Anonymity and hybrid approach of l-Diversity & t-closeness, we can achieve better results which depicts minimum information loss and better accuracy.

VI. REFERENCES

[1]. Hina Vaghashia,Amit Ganatra,"A Survey: Privacy Preservation Techniques in Data Mining "International Journal of Computer Applications (0975 – 8887) Volume 119 – No.4, June 2015.

[2]. A.S.Shanthi,Dr. M. Karthikeyan,"A review on privacy preserving Data Mining " 978-1-4673-1344-5/12 ©2012 IEEE.

[3]. Sarra Gacem, Djamila Mokeddem and Hafida Belbachir,"Privacy Preserving Data Mining: Case of association rules" IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 3, No 1, May 2013.Golab, L. And Ozsu, M., "Issues in Data Stream Management," ACM SIGMOD Record, Vol. 32, pp. 5-14(2003).

[4]. N.Punitha,R.Amsaveni,"Methods and Techniques to Protect the Privacy Information in Privacy Preservation Data Mining " N Punitha et al,Int. J. Comp. Tech. Appl., Vol 2 (6), 2091-2097.

[5]. Tapasya Dinkar, Aniket Patel and Dr. Kiran R. Amin ," Preserving The Sensitive Inofrmation Using Heuristic Based Approach " 978-1-4673-9802-2/16 © 2016 IEEE.

[6]. Christy Thomas, Diya Thomas,"An enhanced method for privacy preservation in data publishing" 4th ICCCNT – 2013 July 4 - 6, 2013, Tiruchengode, India.

[7]. Nagendra kumar.S,Aparna.R "Sensitive Attributes based Privacy Preserving in Data Mining using k-Anonymity" International

Journal of Computer Applications (0975 – 8887) Volume 84 – No 13, December 2013.

[8]. Pu Shi,Li Xiong, Benjamin C. M. Fung,"Anonymizing Data with Quasi-Sensitive Attribute Values" CIKM'10, October 26–30, 2010, Toronto, Ontario, Canada. Copyright 2010 ACM 978-1-4503-0099-5/10/10.

[9]. R.Indhumathi, S.Mohana, "Data Preserving Techniques for Collaborative Data Publishing" International Journal of Engineering Research & Technology (IJERT) Vol. 2 Issue 11, November – 2013.

[10]. Pierangela Samarati,Latanya Sweeney, "Protecting Privacy when disclosing information: k-Anonymity and its enforcement through generalization and suppression"F30602-96-C-0337.

[11]. W.T. Chembian,Dr. J.Janet,"A Survey on Privacy Preserving Data Mining Approaches and Techniques " Proceedings of the Int. Conf. on Information Science and Applications ICISA 2010 6 February 2010, Chennai, India.

[12]. Pierangela Samarati,Latanya Sweeney,"Generalizing Data to Provide Anonymity when Disclosing Information" The work of Pierangela Samarati was supported in part by National Science Foundation and by DARPA.

[13]. Manish Shanna ,Atul Chaudhar,Manish Mathuria,Shalini Chaudhar,Santosh Kumar,"An Efficient Approach for Privacy Preserving in Data Mining"978-1-4799-3140-8/14 ©2014 IEEE.

[14]. S.Vijayarani, A.Tamilarasi, M.Sampoorna,"Analysis of Privacy Preserving K-Anonymity Methods and Techniques" Proceedings of the International Conference on Communication and Computational Intelligence – 2010, Kongu Engineering College, Perundurai, Erode, T.N.,India.27 – 29 December,2010.pp.540-545.

[15]. Supriya Borhade Researcher, Department of Computer Engineering, Pune University, Pune, India "A Survey on Privacy Preserving Data Mining Techniques" IJETEA International Journal of Emerging Technology and Advanced Engineering Volume 5, Issue 2, February 2015.

[16]. M V R NarasimhaRao, J.S.VenuGopalkrisna, R.N.V. Vishnu Murthy, Ch. Raja Ramesh," Closeness Privacy Measure For Data Publishing Using Multiple Sensitive Attributes" IJESAT] International Journal Of Science & Advanced Technology Volume-2, Issue-2, 278 – 284.

[17]. Ashwin Machanavajjhala ,Johannes Gehrke, Daniel Kifer,"l-Diversity: Privacy Beyond k –Anonymity" Proceedings of the 22nd International Conference on Data Engineering (ICDE'06) 8-7695-2570-9/06 © 2006 IEEE.

[18]. Yuichi Sei, Takao Takenouchi, Akihiko Ohsuga,"(l1,lq)-diversity for Anonymizing Sensitive Quasi-Identifiers" 978-1-4673-7952-6/15 © 2015 IEEE DOI 10.1109/Trustcom - BigDataSe-.

[19]. Ninghui Li,Tiancheng Li,Suresh Venkatasubramanian,"t-closeness: Privacy Beyond k-Anonymity and l –Diversity" 1-4244-0803-2/07 ©2007 IEEE.

[20]. Jordi Soria-Comas, Josep Domingo-Ferrer, David S´anchez and Sergio Mart´ınez,"t-closeness through Microaggregation: Strict Privacy with Enhanced Utility Preservation" 978-1-5090-2020-1/16 2016 IEEE.

[21]. Xiaokui Xiao Yufei Tao,"Personalized Privacy Preservation" SIGMOD 2006, June 27–29, 2006, Chicago, Illinois, USA. Copyright 2006 ACM 1-59593-256-9/06/0006.

[22]. Jian Xu, Wei Wang, Jian Pei, Xiaoyuan Wang,Baile Shi, Ada Wai-Chee Fu, "Utility-Based Anonymization Using Local Recoding" KDD'06, August 20--23, 2006, Philadelphia, Pennsylvania, USA. Copyright 2006 ACM 1-59593-339-5/06/0008.

[23]. UCI Machine Learning Repository- http://archive.ics.uci.edu/ml/index.php

[24]. ARX Tool- http://arx.deidentifier.org/