Survey on Anonymization using k-anonymity for Privacy Preserving in Data Mining

Binal Upadhyay*, Dr. amit ganatra[2].

*Department of Computer Engineering, Parul University
Limda, Waghodia Road, Vadodara, Gujarat 391760,India
[1] Email: binal1994upadhyay@gmail.com

Department of Computer Engineering, Changa
Charotar university of science and technology
[2]Email: amitgantra.ce@charusat.ac.in

Abstract—**K-Anonymity widely used in protecting privacy. By the definition we can say anonymization means a nameless for that take one example like a person should not identifiable untraceable or unreachable. Anonymization using k-anonymity for privacy preserving gives the best privacy for the data and more protective for the whole datasets. In this paper we used Hybrid anonymization for mixing some type of data. In this case we show that this model applied to various data mining problems and also various data mining algorithms. In the many paper we show that using the k-anonymity we reduce the more information loss but here issue is that not satisfied with multiple sensitive attributes. Using the k-anonymity for privacy preserving the main motivation is removing or transferring personally identifiable information.**

Keywords— Data Mining, K-anonymity, Classification, SVM, Privacy Preserving, Hybrid Anonymization

I. Introduction

Now a days fast development age, more and huge amount of data which is used by a people, at that same timing privacy issue in that published data have drawn more and more people's attention. K-anonymity is the anonymization approach which is proposed by the Samarati and Sweeny attacks [13]. Here we can say that the classification techniques can be applied for the figure that relationship between the quantity and features of sold items [9]. In the k-anonymity many attacks are apply on data sets like Linking Attack for this we get one example in that case we have two data set table like medical database and voter database here we compare the same data like age now we show that in both table and if we find that age 39 than show both table and find the person name and his disease like that. Second attack is Homogeneity attack in that case all the sensitive values in each equlience class are identical in such case even though the data is anonymized the sensitive value of that an individual can be predicted. In third case it is a Background knowledge attack in this attack that entire sensitive attribute can be identified based on the association between on ore more quasi identifier attributes.[4] Here the concept of the privacy preserving in data mining is that extend the main traditional data mining techniques to work with modify related data and hide sensitive information. For that PPDM that support the cryptographic and anonymized based approach. In the Cryptographic approach carry out the data mining task using secure multi party computational (SMC). Another approach is the anonymization approach in that case it replaces original value of attribute with modified related value in data base for privacy preservation. In our concept we use anonymized approach because cryptographic approach is used only on distributed database and anonymized approach used on both distributed and centralized database. Providing the k-anonymity the main purpose divide patterns into common patterns into a special pattern, with a special pattern Is a one type of pattern which is not shared by more than k peoples and that time only blur and remove that special pattern before it release. One another thing of k-anonymity model that is emphasizes upon the existence of a minimum of k vertices in the anonymized network a node that cannot be re-identified with confidence more than 1/k. In this paper we used many data mining algorithms which used for the k-anonymity. In the case of the PPDM we can use multiple methods like that the classification methods based on the some different area like data distribution for the dataset, data distortion for the data, data mining algorithms for the whole dataset, data or rules for hiding the data and that gives the most security for the datasets [1].

II. RELETED WORK

An easy Privacy protection technology is one type of firstly popular academic research which has many applications which are famous in many areas in recent years [1]. K-anonymity is one of the techniques which help in releasing a large amount of data.

A. Anonymized Approaches

There are many anonymized approaches are available like k-anonymity, l-diversity for k-anonymity, p-sensitive k-anonymity, (α, k)-anonymity, t-closeness, (k, e)-anonymity, (c, k)-safety, m-confidentiality, skyline privacy etc. These all approaches are firstly focus on achieving the anonymized data and do not consider on how much data to be anonymized. Here we can say that such type of anonymized may be of good quality and preserve the privacy. However, it leads to unnecessary information loss. For this reason we need to restrict the amount of data which is allowed for generalization. In all that approaches some approaches are Generalization based approaches and some are Permutation based approaches. In this paper we mainly focus on the k-anonymity approaches. And k-anonymity approach is generalized based approach and generlize approach is better than the Permutation based approach because it provides protection against the presence privacy and association privacy. Here observing the k-anonymity we find that the k-anonymity does not maintain diversity of the sensitive attribute in each equivalence class.

B. PPDM Techniques

In the previous work there is many PPDM techniques are available. There is two main area for the PPDM Techniques.

- Data Modification
- SMC

Here data modification and SMC is the two many part of the PPDM techniques. In the case of data modification there is four part one is perturbation, condensation, K-anonymity, Differential Privacy and in the SMC there is classification and cryptography based anonymization is available. We show the one table for the merits and demerits for the PPDM techniques [3].

TABLE I
MERITS AND DEMERITS OF THE PPDM TECHNIQUES [3]

Techniques	**Merits**	**Demerits**
Perturbation	Different attributes are preserved like that is separate. It has high data utility.	Privacy preservation is very less. If we want to reconstruct the original data that is not possible.
Condensation	It is good performed with the stream data sets.	There is large amount of information loss occur.
Anonymization	There is individual privacy is maintained.	Use linking attack. Heavy information loss occurs.
Differential Privacy	Accuracy of results and improved utility.	There is problem is that Scalability level is still a question
Evolutionary Algorithms	It is more secure and effective.	High uncertainty.
SMC	Accuracy of results Effective. Transformed data are exact and more protected.	Complicated when more than two parties are involved. And it is more Expensive.

B. K-anonymity Technique

K-anonymity is technique which gives the new and more efficient ways for anonymized data and it preserve patterns during whole anonymization. The k-anonymity model defines the whole privacy of output of process and that process is not by itself. It is simple and well understood model [10,12,16]. K-anonymity is main privacy protection model. K-anonymization provides joining attacks by using the suppression and generalization for released micro data so that no individual can uniquely distinguished from the size of k. There are three k-anonymity algorithms are available first is incognito algorithm second is Samarati's Algorithm and third is Sweeney's Algorithm. In the incognito algorithm it produces all the possible k-anonymous full domain generalization of a relation with optional tuple of suppression threshold. The main advantages of incognito mode which is it finds all k-anonymous generalization of full domain and it select optimal solution selected using the different area. The main disadvantage of this incognito algorithm is that it uses the breadth first search method for the traverse solution shape. The second algorithm is Samarati's Algorithm which search for searches for the possible k-anonymous solutions by jumping at different DGH level [15]. And it uses the binary search to obtain the solution in less time. Maximum number of tuples allows achieving the k-anonymity. And finally third algorithm Sweeny algorithm that gives the best solution attained after generalizing the variables with the unique values and find that this approach is much more efficient. Using the k-anonymity model it gives the result very fast. Sometimes this released data may not be suitable for research purpose as it provides very little information. There is some risk in the k-anonymity. We have shown that the actual risk of re- identification of individual records is which is often lower than the worst-case risk for most of the records so long as the adversary that has knowledge of some or all quasi-identifier attributes. However, we have also shown that the risk may be dramatically higher with the knowledge of other attributes beyond all the quasi-identifier [18,19]. A database satisfies K-anonymity if every record is in-distinguishable on quasi-identifier from at least k-1 other records. K-anonymity attends much anticipated popularity. K-anonymity algorithms and semantic

.

III. ALGORITHMS

The k-anonymity model work with some algorithm and gives the different type of result. K-anonymity is the one most privacy preservation model that provide the protection on the anonymized data sets. Some algorithms are emphasized here.

A. SVM

Vapnik and colleagues (1992) groundwork from Vapnik & Chervonenkis' statistical learning theory in 1960. SVM classifying data in common task in machine learning. A support vector machine constructs a hyper plane or set of training data point of any class since in general the larger the margin the lower the generalization error of the classifier.

The entire research on the support vector Machine is a one type of supervised machine learning scheme that divides the data sets in different two parts.

- linear classification
- Non-linear classification.

We can define SVM as a support vector machine that construct one line between parameters and find the maximum distance between the parameters and maximum result find the better results. SVM is used for classification, regression, or other tasks. In the linear classification in SVM it is possible to find linear hyper plane divides the data into two classes in the whole training process. In the non-linear classification all the data distributed in non-linear situation and it can hardly find a linear hyper plane to divide the data perfectly in original dimension. SVM is the supervised learning model. SVM is best supervised classifier [2].

B. *Clustering Algorithm*

Requirement of clustering is generate result based on data generated cluster by itself. It is called as the internal evaluation. These methods usually assign the best score to the algorithm which produces clusters with high similarity within a cluster and low similarity between clusters. Clustering is one type of un-supervised algorithm. The main goal of clustering algorithm is it determines the intrinsic grouping in a set of un-labelled data. Clustering algorithm has some requirement like it dealing with different type of attribute. It discover cluster with arbiter sets. There is some problems occurs like

Clustering algorithm for k-anonymity has shown great deal. We get one example like KNN which is call nearest neighbour which search can be done efficiently for low dimensional spaces with kd-trees or similar structures, in high dimensional spaces for query times with these structures reduce the linear search. There is another clustering algorithm is incremental clustering that improve the k-anonymized data set. Greedy clustering algorithm also available which but it does not lend itself but it improves the quality of a solution by that clusters.

C. *KAMP*

In the KAMP algorithm two pattern are include which are,

- generalization
- suppression

Basically in case of generalization there we will predict the value because in that value put in the one range like one person who age 35 in that case in the generalization we put the value in range like 33-36 like that and in the suppression the value of attribute are replace with some special value like "*" for that we get one example age with value [39] is generalized as [3*].In that case we don't find the exact value that gives high privacy. K-anonymity of multi-pattern (KAMP) to protect data from re-identifying users by using the combination of patterns[8].

D. *Hybrid anonymization*

In hybrid anonymization there is s-hybrid and multi-dimensional hybrid for the k-anonymized dataset. Here we can say that hybrid anonymization is that in which a limited number of data elements can be relocated. We can say that relocation is potentially increasing the utility of anonymization at the cost of truthfulness. Hybrid technique can also be evaluated with respect to different cost metric and real application show that utility gain can better quantified [9].

D. *The greedy algorithm*

In the case of the greedy algorithm is that is the instead of striving to build a k-regular generalization graph over the data at once, we can do one thing is that we can set the data in a sequence of k distinct iterations and adding a whole single assignment to the graph under the construction at each iteration. We can say that this algorithm is designed to achieve optimum solution for a given problem in the data set. We can say that in greedy algorithm approach, whole the decisions are made from the given solution from the given domain. As being greedy, the closest solution of that seems to the main aspects is this algorithm provide an optimum solution which is chosen. Greedy algorithms trying to find a localized optimum solution for the data set, which are may eventually lead to globally optimized solutions. However, generally greedy algorithms do not provide globally optimized solutions. We find a greedy algorithm have not time complexity by the experiment we find that the data utility is gain up 41% and also gives the efficiency advantages [19]. By the result we show that greedy algorithm works for the practical values of the k used which is used in the real world settings and also find that linear-time–back-tracking for the greedy process which does not affect the complexity $O(n^3)$ which is the iteration complexity. We find that overall complexity of the iteration is O(kn2). We show the advantages apply on the time efficiency [6].

E. *Comparison table for algorithm*

TABLE II
COMPARISON

Algorithm	**Parameters**		
	Complexity	**Efficiency**	**Time cost**
SVM	$O(n^3)$	Low	High
KAMP	$O(log n)$	High	Low
Hybrid anonymization	$O(n\ log n)$	Need Improve	High
K-anonymous Decision tree	$O(n\ log n)$	High	Low

IV. EXPERIMENTAL RESULT

A. *Comparison graph*

Here we show that for algorithm which have all different complexity efficiency data utility, and time cost now we show the result.

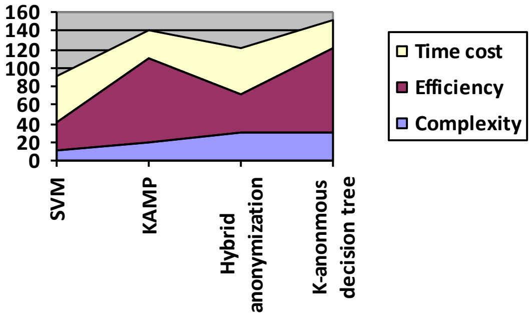

Fig. 1 A chart of time cost, Efficiency, Complexity of SVM and Hybrid anonymization.

Every here we can see that in the result that the complexity of the SVM, KAMP, Hybrid anonymization, and k-anonymous decision tree in the chart that shows that the SVM has the lower complexity in the group of algorithm. KAMP's complexity is not lower but also not high so we find that the KAMP has needed to improve that complexity. In the case of the of hybrid anonymization that have high complexity and k-anonymous decision tree have also a high complexity

For the Efficiency we find that the SVM have higher efficiency KAMP conduct lower efficiency. In the case of hybrid anonymization efficiency is good but that have to improve it. And in the last k-anonymous decision tree have the higher complexity.[5,10]

In the time cost we can find that the SVM need high timing so procedure is going to slow. So in that case we need to improve the time cost. In the second KAMP have low time cost so it is the faster than the SVM. In the case of hybrid anonymization that have high time cost that means it needed more timing than the KAMP in the last decision tree that need the low time cost so it is the faster in the case and it gives the fast result.

B. *Line chart for data utility*

We can say that It is a more important issue for utility of data privacy protection. We can say that in order to hide sensitive information, false information should insert the database, or block data values. Although sample Techniques do not modify the information stored in the database, but that, since their information is incomplete, still reduces data utility. More changes to the database, less data utility of the database. So estimated parameters of data utility is data information loss applied privacy protection. Of course, the estimate of information loss related with the specific data mining algorithms.

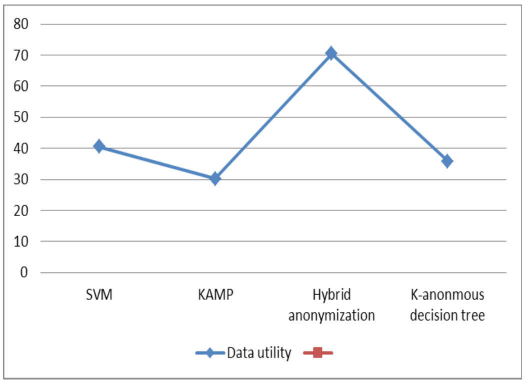

Fig. 2 A sample line graph for the data utility.

Here line graph shows that the all the different algorithm's data utility graph from that graph we find the hybrid anonymization have the highest data utility and KAMP have lowest data utility and SVM and k-anonymous decision tree need to improve that data utility that we find hybrid Anonmization is the better in the performance using the hybrid anonymization we get the better result in k-anonymity and also in the multi-dimensional k-anonymization..

V. Conclusions

Finally, I conclude that privacy preserving in data mining ids the main aspect to provide the privacy. Privacy is necessary to protect people in competitive situations. Using the k-anonymity with anonymization and using the suppression and generalization method for the more secure database it provide the security to the different type of datasets. K-anonymity is important privacy preserving model for the data mining. We also show the complexity, time cost, efficiency and complexity of our experiments. Privacy in data stream mining, Efficiency and minimum computation cost in distributed PPDM, Privacy and accuracy with minimal loss.

References

[1] Xinjun Qi , Mingkui Zong School of Technology,Harbin University An Overview of Privacy Preserving Data Mining, 2011 International Conference on Environmental Science and Engineering (ICESE 2011)

[2] Qi Jia∗, Linke Guo∗, Zhanpeng Jin∗, Yuguang Fang† ∗Department of Electrical and Computer Engineering, Binghamton University, Binghamton, NY 13902, USA †Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611, USA Email: {qjia1, lguo, zjin}@binghamton.edu, fang@ece.ufl.edu Privacy-preserving Data Classification and Similarity Evaluation for Distributed Systems 2016 IEEE 36th International Conference on Distributed Computing Systems.

[3] G. Arumugam Senior Professor and Head, Department of Computer Science Madurai Kamaraj University Madurai, Tamilnadu, India. V. Jane Varamani Sulekha Research Scholar, Department of Computer Science Madurai Kamaraj University Madurai, Tamilnadu, India. IMR based Anonymization for Privacy Preservation in Data Mining.

[4] T. Pranav Bhat, C. Karthik∗ and K. Chandrasekaran Department of Computer Science and Engineering, NITK Surathkal 575 025, Karnataka, India A Privacy Preserved Data Mining Approach Based on k-Partite Graph Theory T. Pranav Bhat, C. Karthik∗ and K. Chandrasekaran Eleventh International Multi-Conference on Information Processing-2015 (IMCIP-2015).

[5] Fran Casino, Joshep Domingo-ferrer, constantinos, Domenec puing, Agusti solans, A k-anonymous approach to privacy preserving collaborative filltwering Journel pf computer and system science 2014.

[6] Aggarwal, C., Yu, P. S. A condensation approach to privacy preserving data mining. In proceedings of International Conference on Extending Database Technology (EDBT), pp.183–199, 2004. 746.

[7] Kun Liu, Chris Giannella, and Hillol Kargupta . A Survey of Attack Techniques on Privacy-Preserving Data Perturbation Methods. Privacy-Preserving Data Mining, volume 34 of Advances in Database Systems, Springer, (2008).

[8] Chia-Hao Hsu Department of Electrical Engineering National Chung HsingUniversityTaichung,Taiwan,R.O.C.Email:w100064001@mail.nchu.edu.tw Hsiao-Ping Tsai Department of Electrical Engineering National Chung Hsing University Taichung, Taiwan, R.O.C. Email: KAMP: Preserving k-anonymity for Combinations of Patterns 2013 IEEE 14th International Conference on Mobile Data Management.

[9] Mehmet Ercan Nergiz, Muhammed Zahit gok Hybrid k-anonymity journalhomepage: http://dx.doi.org/10.1016/j.cose.2014.03.006.

[10] Arik Friedman · Ran Wolff · Assaf Schuster Providing k-anonymity in data mining Received: 30 September 2005 / Revised: 24 May 2006 / Accepted: 2 August 2006 / Published online: 10 January 2007 © Springer-Verlag 2007.

[11] Matthew Andrews Bell Labs, Murray Hill, NJ a Gordon Wilfong Bell Labs, Murray Hill, NJ Lisa Zhang Bell Labs, Murray Hill, NJ Analysis of k-Anonymity Algorithms for Streaming Location Data The Third International Workshop on Security and Privacy in Big Data (BigSecurity 2015).

[12] Adeel Anjum∗ ,Adnan Anjum† ∗Comsats Institute of Information Technology, CIIT, Park Road Chak Shahzad Islamabad adeel.anjum@comsats.edu.pk †National University of Science and Technology IslamabadDifferentially Private K-anonymity. 12th International Conference on Frontiers of Information Technology 2 2014 IEEE.

[13] L. Sweeney, “k-anonymity: A model for protecting privacy,” International Journal of Uncertainty Fuzziness and Knowledge Based Systems, vol. 10, no. 5, 2002, pp. 557-570, doi: 10.1142/S021848850 2001648.

[14] Jaimain Han, Jaun Yu, Yuchang Mo, Jianfeng Lu, Huawen Liu, MAGE: A Semantics retaining k-anonymization method for mixed data. journal homepage: www.elsevier.com/locate/knosys 0950-7051/$ - see front matter 2013 Elsevier B.V. All rights reserved.

[15] Zhao FeiFei 1,Dong LiFeng2, Wang Kun2,Li Yang2 Study on Privacy Protection Algorithm Based on K-Anonymity 2012 International Conference on Medical Physics and Biomedical Engineering.

[16] 1School of Computer Science and Technology Tianjin University Tianjin, China 2Postgraduate Training Brigade Military Transportation University Tianjin, China Study on Privacy Protection Algorithm Based on K-Anonymity 2012 International Conference on Medical Physics and Biomedical Engineering.

[17] Xiangwen Liu School of Computer Science and Communication Engineering Jiangsu University Zhenjiang, China Qingqing Xie School of Computer Science and technology, Anhui University Hefei, China Liangmin Wang* School of Computer Science and Communication Engineering Jiangsu University Zhenjiang, China A Personalized Extended (α, k)-Anonymity Model 2015 Third International Conference on Advanced Cloud and Big Data.

[18] Anirban Basu, Toru Nakamura, Seira Hidano, Shinsaku Kiyomoto KDDI R&D Laboratories Inc., 2-1-15, Ohara, Fujimino-shi, Saitama 356-8502, Japan k-anonymity: risks and the reality 2015 IEEE Trustcom/BigDataSE/ISPA.

[19] Katerina Doka NTUA, Greece, Mingqiang Xue I2R, Singapore ,Dimitrios Tsoumakos Ionian University, Greece ,Panagiotis Karras Skoltech, Russia k-Anonymization by