ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Issue 4, 2017 April

A PPDM Framework to Analyze Privacy and Utility Trade-Off for MNQIA Anonymization Algorithm

K.Abrar Ahmed[1], H.Abdul Rauf[2]
Research Scholar, Department of CSE Manonmaniam Sundaranar University[1]
Acedamic Dean, Sree Sastha Institute of Engineering and Technology[2]

Abstract: The rising developments in the field of Internet have facilitated the individual to a greater extent to share their personal data for numerous applications such as analysis, mining, forecasting and prediction etc. Therefore, these have created a huge opportunities to the adversary to misuse the data's that available and probably leads to privacy threat of the individuals involved and support knowledge management and Information retrieval. Several approach have been proposed such Anonymization and perturbation, which promotes the attention of privacy preservation data mining (PPDM) among the researchers. K-Anonymization is widely used approach to prevent the adversary from being raising the privacy threat to individuals who involved in the process. Our framework purely concentrates in adopting K-Anonymization strategy that suits well utility aspects of the PPDM. The Utility is justified based on classification accuracy. Our approach uses bucketization and MMDCF methods to achieve K-anonymization and classification framework of anonymization which strives to declare the optimal K-Anonymization that suits the better utility. Our experimental results facilitate in deriving the optimal K factor that facilitates the data owner to create a utility enriched privatized database (i.e. based on information loss and classification accuracy).

Keywords: PPDM, K-Anonymization Classification, Bucketization, MMDCF.

I. INTRODUCTION

In recent year the data's about individuals are shared or exchanged for different reasons. Those data contains sensitive information of user, thereby increasing concerns about privacy. In order to achieve privacy on data, PPDM plays a major role. The main idea of PPDM is to develop a method that should modify or transforms original data in such a way that no one can identify the sensitive information of owner. There are many PPDM techniques have been proposed, some of them are secure multi party computation, cryptography and randomization. Many researchers have done outstanding work for achieving privacy on datasets using K-Anonymity. The main idea of K-Anonymity is that each record cannot be distinguished from at least (K-1) records. K-Anonymity uses two techniques such as generalization and suppression. Generalization Techniques replace QI attribute value to less-specific but semantically consistent. Suppression eliminates the records present in datasets and update the record with some special symbol such as '*' which means any value can be present. Suppression reduces quality on information of individuals.

All PPDM Techniques focus on single attribute for anonymization. At this moment only a few work concentrate on multiple QI attribute anonymization at once. In our work we use Bucketization approach combined with maximal Multi-Dimensional capacity First (MMDCF).The main idea of Bucketization is to partition the multiple QI attribute in to equivalence classes. The QI value in each equivalence is generalized to same value under K-Anonymity principle. MMDCF is linear greedy algorithm uses to choose record from bucket. We use classification Techniques to find out utility of anonymized data.

The rest of the paper organizes as follows: section 2 states about preliminaries of work.section 3 describe about the methodology used to achieve K-Anonymity. Section 4 states about Experiment and result. Section 5 concludes this paper.

II. RELATED WORKS

In Paper [Ninghui Li, Tianchengi Li, Suresh Venktasubramainan], Author's finds the problem that K-

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Issue 4, April 2017

Anonymity method suffer from Identity disclosure problem. Also they find problem with l-diversity techniques as not enough to halt attribute disclosure. So Author's gives solution to that problem by proposing techniques called t-closeness. Also they explore the motivations for t-closeness and shown its advantages through experiment.

In paper [Ji-Won Byun,Ashish Karma, Elisa Bertino,Ninghui Li], Author's problem is to reduce information loss after performing Anonymization on datasets. Author propose a method called K-member clustering which uses to minimize information loss and to guaranteed the data quality. In addition to that they provide their method as NP-hard, Also they develop a metric to calculate information loss induced by generalization.

In paper [He Zhi-qiang, Chen Gang], Author's for the first time applied K-Anonymity in mobile network for giving security in Location Based Services (LBS). Since K-Anonymity may have many drawbacks such as neighborhood attack, Identity disclosure attack, Background Knowledge attack. Author's identified the sources for the creating problems in K-Anonymity and develops a new method using hierarchical clustering. They experimentally verified and proved that their idea is improved form existing K-Anonymity.

In paper [Peter KIeseberg, Sebatian Schrittwieser, Mrtin Mulazzani, Isao Echizen], Author's propose a method that gives solution for two problems occur in exchanging sensitive information about individuals. Also author's forms a policy to find collaborative attack as well as anonymization scheme.

In Paper[Chen wang,Lianzhong Liu,Lijie Gao], Author's finds the problem that almost all the Anonymization algorithm currently used are depends on generalization hierarchy which may domain generalization hierarchy (DGH) or value generalization hierarchy (VGH) in order to make data Anonymous. This method obviously gives information loss on data. Authors develop a methods using K-Member clustering Algorithm and experimentally proved that it has less information loss compared to other K-Anonymity method.

III. METHODOLOGY

Our methodology mainly concentrates on Anonymization and classification. Here we anonymized the original datasets using K-Anonymization approach and derived information loss for various value of K (i.e K=5,6,….12) .Then we apply ZeroR classification method to device the utility of the anonymized dataset. Hence by our methodology we attempts to deliver the optimal k_value that can be used for anonymizing the original datasets through which the two conflicting goals privacy and utility is achievable to a better trade off. The factors considered for deriving optimal K-value are i) Information loss with respect to privacy ii)Classification Accuracy with respect to Utility.

A. Single QI Attribute Anonymization:

Many K-Anonymization Algorithms have been proposed in [2][3][4],which anonymized only one single attribute at a time, whose purpose is to make individual's data to be very similar in a published data. In the other word, a record of each QI has to similar for at least (k-1) other records.

B. Multiple Numerical QI Attribute Anonymization [13]:

In this division, we will explain briefly about our K-Anonymization process. In paper [13] We considered the datasets be S with 'n' number of records and every records has 'm' numerical Quasi-identifier attributes (QI).we had taken these QI attributes and mark as QI_1,QI_2,….QI_m. For each QI_i, $1 \le i \le m$, we clustered these values into multiple group based on coarse-grained level.

From the group of clusters, we need to create multi-dimensional bucket. The main idea is that 'n' tuples has to be mapped into their corresponding bucket according to their own QI attributes. Once the multi-dimensional bucket is constructed, we use maximal multi-dimensional capacity first (MMDCF) methods [5] to choose different records to form QI group. The selection priority of MMDCF is based on

$$\text{Selection}(\text{buk}<QI_0^1, QI_0^2, \ldots QI_0^d>) = \sum_{1 \le j \le d} capa(QI) + \text{size}(\text{buk}< QI_0^1, QI_0^2, \ldots QI_0^d >$$

QI_0^1, QI_0^2,….. QI_0^d >.once we select different record to make up matching QI-group, thereby we formed K=Anonymous Table.

1) Example [13]:

In this section we explain our methods via real situation. In Paper [13] we considered the following micro data.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Issue 4, April 2017

Id	Age	Zip	Salary	Bonus
T1	27	12,000	1000	1010
T2	22	22,000	2975	1010
T3	34	24,000	10,100	950
T4	26	17,000	1040	2000
T5	30	16,000	3050	2020
T6	32	14,000	5000	3035
T7	22	19,000	5120	2950
T8	37	26,000	7950	4100
T9	39	27,000	1050	6000

Table: 1 Micro data

There are two quasi-Identifier Age and zip code and two sensitive attribute such as salary and bonus. We put the age cluster into four group: A11={26,27}, A12={22,22}, A13={30,32,34} and A14={37,39}.Ultimately we put the zipcode into four cluster groups: A21={12,000,14,000}, A22={17,000,16,000,19,000}, A23={22,000,24,000} and A24={26,000,27,000}, it is shown in table 2.

Age-Group(A1i)	Zip-code Group(A2i)
A11={22,22}	A21={12,000,14,000}
A12={26}	A22={17,000,16,000,19,000}
A13={32,33,34,35}	A23={22,000,24,000}
A14={37,39}	A24={26,000,27,000}

Table: 2 Two cluster Group

We make age and zipcode to be the first dimension and second dimension respectively. Now check the tuple t1 values of age and zip code with two cluster group.

	A11	A12	A13	A14
A21			{T1,T6}	
A22	T7	T4	T5	
A23	T2		T3	
A24				{T8,T9}

Table: 3 Two Dimensional Cluster

Then tuple t1 belongs to group A13, A21.Therefore we put t1 in the corresponding cell.Similarly, we place all the other records as well. We structure a two dimensional bucket as in above table.

According to MMDCF [15], we can choose different record to make up the matching QI-Group. For example Age and Zip code are choosen as QI-Attributes. Now according to the selection priority equation is as follows.

Group 1:

Iteration-1

Selection (buk< A21, A13> = 8 tuples ⟶
{T1, T6} to break the tie tuple **T1** is selected.

There are 4 tuples in A13, 2 tuples in A21 and 2 tuples in buk<A21,A13>, Totally 8 tuples. The Priority in buk<A22,A12> is 6 tuples, and in buk<A22,A13> is 7 tuples which is rejected because tuple from A13 already selected in Iteration 1. To break tie between buk<A22,A11> and buk<A22,A12>, we select buk<A22,A11> whose tuples is T4.Therefore the highest priority is buk<A22,A11> so tuples **T4** is selected, then we shield dimension <A11>.

Iteration-2

Slection(buk< A22,A11> = 6 tuples ⟶ **T7 Selected.**
Selection(buk< A22,A12> = 5 tuples ⟶ T4
Slection(buk< A22,A13> = 8 tuples ⟶
Already tuple selected from A13.

There are 2 tuples in A11, 3 tuples in A22 and 1 tuples in buk<A22,A11>, Totally 6 tuples. The Priority in buk<A22,A12> is 5 tuples, and in

buk<A22,A13> is 8 tuples which is rejected because tuple from A13 already selected in Iteration 1.The highest priority is buk<A22,A11> so tuples **T4** is selected, then we shield dimension <A11>.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Issue 4, April 2017

Iteration-3

Selection(buk< A23,A11> = 5 tuples → Already tuple selected from A11
Slection(buk< A23,A13> = 6 tuples → Already tuple selected from A13

Iteration-4

Selection (buk< A24,A14> = 6 tuples → {T8 ,**T9**} to break the tie highest tuple **T9** is Selected

Finally in all four iteration we selected 4 tuples in a first group (T1, T7, T9}

Group 2:

The tuples which is selected in group one should be removed from Two Dimensional Cluster table and repeat this procedure.

	A1	A12	A13	A14
A21			T6	
A22		T4	T5	
A23	T2		T3	
A24				T9

Table: 4 Two Dimensional Cluster

The same procedure has to be followed to obtain second group which contains tuples {T2,T4,T6} and Third group which tuples contain {T3,T5,T8}.From these group we perform generalization on multiple Quasi Identifier attributes to get 3-Anonymization with 3-diversity on micro data as shown below.

ID	Age	Zip	Salary	Bonus
T1	22-39	12,000-27,000	1000	1010
T7	22-39	12,000-27,000	5120	2950
T9	22-39	12,000-27,000	1050	6000
T2	22-32	14,000-20,000	2975	1010
T4	22-32	14,000-20,000	1040	2000
T6	22-32	14,000-20,000	5000	3035
T3	34-37	24,000-26,000	10,100	950
T5	34-37	24,000-26,000	3050	2020
T8	34-37	24,000-26,000	7950	4100

Table: 5 Anonymized micro data

C. ZeroR classification Methods:

ZeroR is one of the easiest classification methods that depends on target and omits all predictors. ZeroR classifier [6] simply predicts the majority category (class).Although there is no predictability power in ZeroR, it is useful for determining a baseline performance as a guidelines for other classification methods. Algorithm constructs a frequency table for the target and selects it's most frequent value. Predictors' contribution is not used in this model. The Model Evaluation for zeroR only predicts the majority class correctly. As mention earlier, zeroR is only useful for determining a baseline performance for other classification methods

Significant steps in PPDM Framework:

Step:1 Choose N QI Attributes
Step:2 Cluster the each QI Attributes in cluster based on degree of approximation.
Step:3 Construct Two dimensional buckets.
Step:4 Then apply MMDCF [15], and choose different record to make up the matching QI-Group.
Step:5 Repeat step3,4 untill K-Anonymization and L-diversity is achieved.
Step:6 Apply ZeroR Claasification Algorithm to find out the Uitiliy and Privacy of MNQIA Method.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Issue 4, April 2017

D. Information Loss [13]:

Since we use bucketization and MMDCF method to modify data and form clusters. This Anonymization suffer from information loss because some original values of QI in every record are either replaced with less specific values or are totally removed. Our goal is to anonymized a dataset that should own privacy while maintain data utility on other hand. We use following metric to calculate the information loss.

According to paper [7], Let D* be anonymized dataset of D.D* corresponds to a set of clusters $c=\{c_1,c_2,c_3,\ldots,c_p\}$ which is group of cluster. All records in a given cluster c_j are anonymized. Information loss occur in anonymizing a data Sets D to D* is given in [7] as

$$IL(D,D^*) = \frac{1}{|D|}\sum_{j=1}^{p} IL(C_j) \qquad (1)$$

Where $IL(C_j)$ is the number of information loss of cluster C_j, Which is defined as the sum of information loss occur in anonymizing every sequence S in C_j as in [7].

$$IL(C) = \sum_{i=1}^{|C|} IL(S_i, S_i^*) \qquad (2)$$

Where |c| is the sum of sequence in the cluster C_j, and IL(S,S*) is the information loss occur in anonymizing the sequence S t the sequence S*.Each sequence is Anonymizd by generalizing or suppressing some QI's values in some of it's events as in [7]. So we define information loss of a sequence based on the information loss of it's events.

Let H be generalization hierarchy of the attribute A.We use the loss metric (LM) measure [7] to capture the amount of information loss occurred by generalizing the value a of the attribute A to one of it's ancestors â with respect to generalization hierarchy H as in [7].

$$IL(a,\hat{a}) = \frac{|L(\hat{a})|-|L(a)|}{|\Delta A|} \qquad (3)$$

Where |L(X)| is the number of leaves in the sub tree rooted at x. The information loss of each events e is then defined as [7]

$$IL(e,e^*)=\sum_{n=1}^{|QI|} IL(e(n), e^*(n)) \qquad (4)$$

Where e* is the ancestor of event e,e(n) is the value of n^{th} QI of the event e and e*(n) is it's corresponding value in the event e*.Hence the information loss incurred by anonymizing each sequence is as follows

$$IL(S,S^*) = \sum_{m=1}^{|S|} IL(e_m, e_m^*) \qquad (5)$$

Generalization hierarchy of attribute age is shown below.

Fig-1: Generalization Hierarchy of Age Attribute

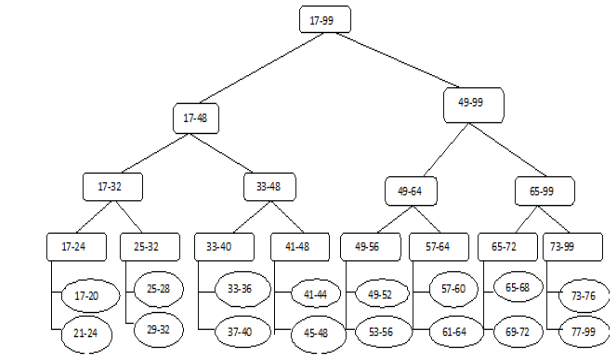

Fig-1: Generalization Hierarchy of Age Attribute

IV. Experiment And Result

We had taken adult dataset [13] from UCI Machine Learning Repository which consists of attributes age, workclass, fnlwgt, education, education number, marital status, sex, race etc. From these attributes age,sex,race have been taken and anonymized using our anonymization method. We obtained Anonymized datasets for different values of K-level.

For different K Values we calculated Information loss for MNQIA method and Datafly method using the metric explain in section 3.4, following table show information loss obtained for different K values. From the below table it is clear that the information loss we obtain for MNQIA is better than Datafly Methods for different values of K-level.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Issue 4, April 2017

K-Value	Information Loss	
	MNQIA Method	*Datafly Method*
4	0	4.25
5	5.59	6.02
6	7.37	7.5
7	8.84	9.03
8	10.09	11.5
9	12.07	12.5
10	15.17	16.8
11	18.47	19.04
12	22.38	23.4

TABLE 6: Information Loss for Different Values of K-level.

The sample implemented output screen shot, the Information Loss when k=5, 12 was shown below,

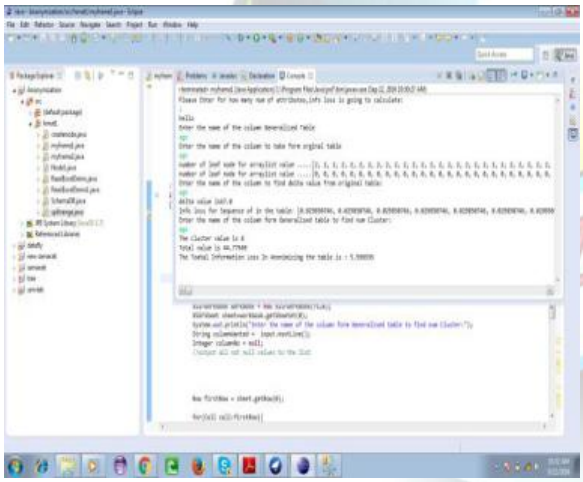

Fig-2: Information Loss MNQIA (K=5)

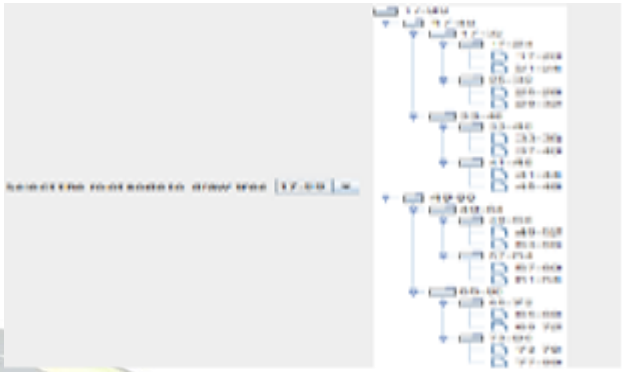

Fig-3: Implemented Output of Generalization Hierarchy for Age Attribute

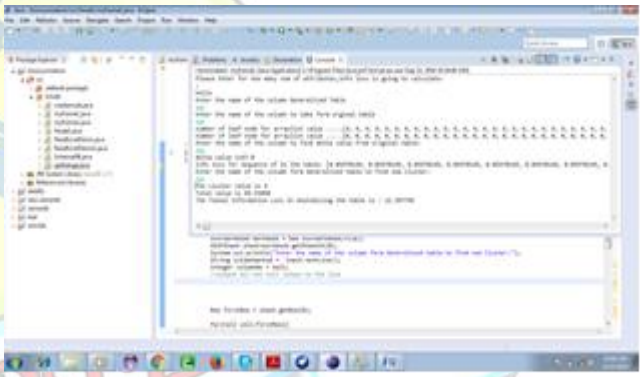

Fig-4: Information Loss of MNQIA (K=12)

We had taken Weak tool where original adult datasets have been classified using ZeroR classification methods, thereby we obtained **classification accuracy as 78.86**.Similarly we had taken different K-Anonymized datasets and applied ZeroR classification, we obtained different classification accuracy for different K values as shown below.

K-Value	Classification Accuracy for MNQIA
5	76.93
6	75.33
7	73.45
8	71.63
9	71.23
10	70.02
11	69.15
12	68.12

Table-7: Classification Accuracy for K-Anonymized Table.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Issue 4, April 2017

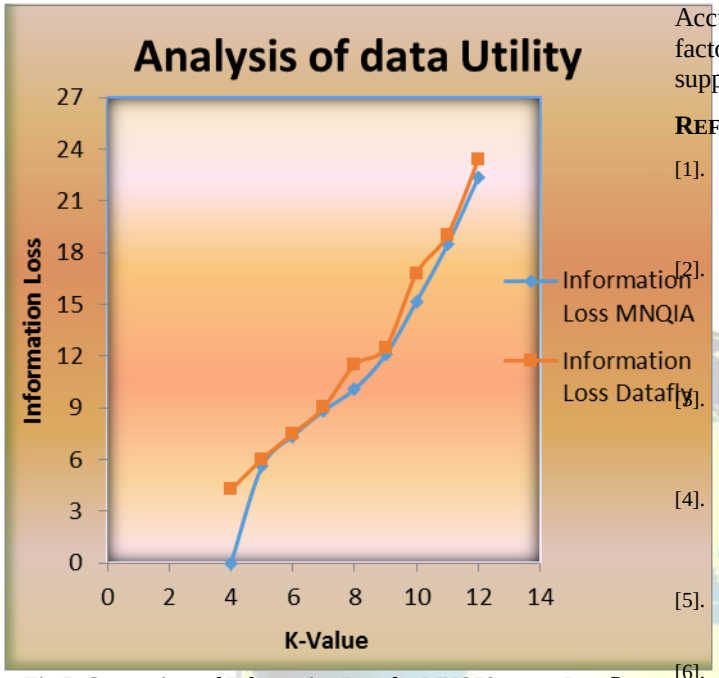

Fig-5: Comparison of Information Loss for MNQIA verses Datafly

The sample screen shot of classification accuracy is shown below

Fig-6 : Classification accuracy when K=5

V. CONCLUSION

Since several K-Anonymization Techniques have been developed for preventing the disclosure of individuals sensitive information .This paper address the framework which purely concentrate in adopting K-Anonymization strategy that suits well with utility and privacy aspects of privacy preserving Data mining. Using our Anonymization Techniques we analyze that the information loss when K=8 is having minimum deviation and better classification Accuracy. We suggest that k=8 is the best Anonymization factor to give better privacy and utility when data is supposed to publish.

REFERENCES

[1]. Ji-won Byum,Ashish Kamra,Elisa Bertino,Ninghui Li,"Efficient K-Anonymization Using clustering Techniques",DASFA 2007,LNCS-4443,pp.no-188-200,2007.

[2]. Jiuyong Li,R.chi-wing wong,Ada wai-chee Fu,Jian Oei,"Anonymization by Local Recoding in Data with Attribute Hierarchical Taxonomies",IEEE Transaction on Data And Knowledge Engineering",Vol-20,No-9,pp.no.1-14,2008.

[3]. Chen wangg,Lianzhong Liu,Lijie Gao,"Research on K-Anonymity Algorithm in privacy protection",Proc.2nd International conf.on computer and Infor.Application(ICCIA-2012),pp.0194-0196,2012.

[4]. Chitra Nasa,Suman,"Evaluation of Different classification Techniques for Web Data',International Journal of Computer Application,Vol-52,No.9,pp. 0975-8887,2012.

[5]. Ashwin MachanavajjhalaGehrke,Daniel Kifer,"L-Diversity:Privacy Beyond K-Anonymity",Proc.22nd

[6]. Ninghui Li,Tiancheng Li,Suresh Venkatasubramanian,"T-Closeness: Privacy Beyond K-Anonymity and L-Diversity", Proc.23nd International conf. on Data Engineering(ICDE),pp.24,2008.

[7]. He Zhi-qiang,chen gang,"Improvement of K-Anonymity Location privacy protrction Algorithm based on Hierarchical Clustering",Applied Mechanics and materials,vols.599-601,pp.1553-1557,2014.

[8]. Peter Kieserberg,Sebastian Schrittwieser,Martin Mulazzan,Isoa Echizen,"An Algorithm for collusion-resistant Anonymization and fingerprinting of sensitive microdata",Institute of Information Management vol.24,pp.113-124,2014.

[9]. Qinghi Liu,Hong Shen and Yingpeng sang,"Privacy preserving data publishing for multiple Numerical sensitive Attribute " ,TSINGHUASCIENCE and TECHNOLOGY,vol.20,no.3,pp.246-254,2015.

[10]. Jianmin Han,Fangwei Lue,Jianfeng Lu and Haopeng,"SLOMS:A privacy preserving Data Publishing Method for Multiple sensitive Attribute Microdata",Journal of Software,vol.8,No.12,pp. 3096-3104,2013.

[11]. Morvarid sehatakar,stan matwin,"Clustering-based multi-dimensional sequence Data Anonymization",proc. EDBT/ICDT 2014 JOINT CONFERENCE,pp. 385-390,2014.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Issue 4, April 2017

[12]. K.AbrarAhmed,H.Abdul Rauf,A.Rajesh,"Study of K-Anonymization Level qith respect to information loss",Australian Journal Basic And Applied Science,vol.10,no.2,pp.1-8,2016.

[13]. K.AbrarAhmed,H.AbdulRauf,A.Rajesh",MNQIA:A METHOD FOR MULTIPLE NUMERICAL QI ATTRIBUTE ANONYMIZATION ",IJCTA,vol.9,no. , pp.227-236, 2016.

[14]. http://storm.cis.fordham.edu/~gweiss/data-mining/ weka-data/diabetes.arff

[15]. http://archive.ics.uci.edu/ml/datasets/BankMarketing

[16]. http://archive.ics.uci.edu/ml/datasets/ILPD+%28Indian+Liver+Patient+Dataset%29

[17]. Kaufman L and Rousseeuw,P.J 1990.Finding grop in data:An Introduction to cluster Analysis.John Wiley 1990.