Volume 3, Issue 10, October 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Evaluating Clustering Performance of K-Anonymity Methods and Techniques in PPDM

M.Sampoorna*
Assistant Professor,
Department of Computer Science and Applications,
K.S.R College of Arts and Science for Women,
Tiruchengode,
India

V.Dineshkumar
Assistant Professor,
Department of Computer Applications,
Sri Ramakrishna Mission Vidyalaya College of Arts and Science,
Coimbatore, India

Abstract— Data mining is the process of extracting hidden information from the database. The current trend in business collaboration shares the data and mine results to gain mutual benefit. Privacy preserving data mining has become increasingly popular because it allows sharing of private sensitive data for analysis purposes. K-anonymity is a property that models the protection of released data against possible re-identification of the respondents to which the data refers. The main goal of this work has introduced a new k-Anonymity algorithm which is capable of transforming a non anonymous data set into a k-Anonymity data set. The proposed work uses the partitional clustering algorithm that is K-Means algorithm. As part of this approach a suitable metric has been developed to estimate the information loss introduced by Suppression, which works for both numeric and categorical data.

Keywords— Data Mining, Privacy, K-Anonymity, Clustering, Suppression.

I. INTRODUCTION

The amount of data being collected every day by private and public organizations is quickly increasing. In such a scenario, data mining techniques are becoming more and more important for assisting decision making processes and, more generally, to extract hidden knowledge from massive data collections in the form of patterns, models, and trends that hold in the data collections. While not explicitly containing the original actual data, data mining results could potentially be exploited to infer information contained in the original data and not intended for release, then potentially breaching the privacy of the parties to whom the data refer [15]. Effective application of data mining can take place only if proper guarantees are given that the privacy of the underlying data is not compromised. The concept of privacy preserving data mining has been proposed in response to these privacy concerns.

Privacy-preserving data mining (PPDM) deals with the trade-off between the effectiveness of the mining process and privacy of the subjects, aiming at minimizing the privacy exposure with minimal effect on mining results. The concept of privacy preserving data mining has been proposed in response to these privacy concerns. Privacy preserving data mining aims at providing a trade-off between sharing information for data mining analysis, on the one side, and protecting information to preserve the privacy of the involved parties on the other side. Several privacy preserving data mining approaches have been proposed, which usually protect data by modifying them to mask or erase the original sensitive data that should not be revealed. K-anonymity is an anonymizing approach proposed by Samarati and Sweeney [15], is a property that models the protection of released data against possible re-identification of the respondents to which the data refers. A data set complies with k-anonymity protection if each individual's record stored in the released data set cannot be distinguished from at least k-1 individuals whose data also appear in the data set. This protection guarantees that the probability of identifying an individual based on the released data in the data set does not exceed 1/k. Generalization and suppression are the most common methods used for de-identification of the data in k-anonymity-based algorithms.

The rest of this paper is organized as follows. In Section II K-Anonymity and the related works are discussed. In Section III Objective of the study was given. Section IV gives the general problem formulation and the basic definitions of k-anonymity. In Section V, the proposed k-anonymity clustering technique for sensitive items is given. The effectiveness of the algorithm is evaluated and the experimental results of the proposed technique are discussed in Section VI. Conclusions are given in Section VII.

II. RELATED WORK

Anonymization is a process that removes or replaces identity information from a communication or record. Communications and records may be made *pseudonymous*, in which case the same subject will always have the same replacement identity but cannot be identified as an individual. Such methods achieve data protection from a twofold perspective. First, if the data are modified, re-identification by means of record linkage or matching algorithms is harder and uncertain. Secondly, even when an intruder is able to re-identify a unit, and cannot be confident that the disclosed data are consistent with the original data. The k-anonymity protection model is important because it forms the basis on which the real-world systems known as Datafly, m-Argus and k-Similar provide guarantees of privacy protection. The

ability to collect and disseminate person-specific data increases daily. Given the sensitive nature of personal information, such as health or financial-related knowledge, it is necessary to construct techniques to protect personal privacy in shared databases. To protect privacy, the computer science community has proposed many models of protected databases. One particular model that has received considerable attention from computer scientists is called k-anonymity. Under k-anonymity, each piece of disclosed data is equivalent to at least k-1 other pieces of disclosed data over a set of attributes that are deemed to be privacy sensitive. Below are some citations we consider to be important with respect to the theory and application of k-anonymity.

In paper "Protecting Privacy when Disclosing Information: k-Anonymity and its Enforcement Through Generalization and Suppression" K-anonymity is an anonymizing approach proposed by Samarati and Sweeney. A data set complies with k-anonymity protection if each individual's record stored in the released data set cannot be distinguished from at least K-1 individuals whose data also appear in the data set. This protection guarantees that the probability of identifying an individual based on the released data in the data set does not exceed 1/k. An important method for privacy de-identification is the method of k-anonymity [25]. The motivating factor behind the k-anonymity technique is that many attributes in the data can often be considered pseudo-identifiers which can be used in conjunction with public records in order to uniquely identify the records. For example, if the identifications from the records are removed, attributes such as the birth date and zip-code and be used in order to uniquely identify the identities of the underlying records. The idea in k-anonymity is to reduce the granularity of representation of the data in such a way that a given record cannot be distinguished from at least (k − 1) other records.

In paper "On the Design and Quantification of Privacy Preserving Data Mining Algorithms" the view of k-anonymization problem from the perspective of inference attacks over all possible combinations of attributes. The author show that when the data contains a large number of attributes which may be considered quasi-identifiers; it becomes difficult to anonymize the data without an unacceptably high amount of information loss. This is because an exponential number of combinations of dimensions can be used to make precise inference attacks, even when individual attributes are partially specified within a range. The provided analysis of the effect of dimensionality on k-anonymity methods, conclude that when a data set contains a large number of attributes which are open to inference attacks, are faced with a choice of either completely suppressing most of the data or losing the desired level of anonymity. Thus, the paper shows the curse of high dimensionality could also be applied to the problem of privacy preserving data mining [14].

III. OBJECTIVE

The objective of this work is to propose a privacy preserving data mining technique for k-anonymization using clustering based on suppression. The main objective of the K-Anonymity model is to transform a table so that no one can make high-probability associations between records in the table and the corresponding entities. In order to achieve this goal, the K-Anonymity model requires that any record in a table be indistinguishable from at least (k−1) other records with respect to the pre-determined quasi-identifier. A group of records that are indistinguishable to each other is often referred to as an equivalence class. In the first phase the k-anonymization of data is done, in the second phase k-means clustering is applied to the k-anonymized dataset.

IV. PROBLEM FORMULATION

The objective of this research work is to propose a privacy preserving data mining technique for k-anonymization using clustering based on suppression. The main objective of the K-Anonymity model is to transform a table so that no one can make high-probability associations between records in the table and the corresponding entities. In order to achieve this goal, the K-Anonymity model requires that any record in a table be indistinguishable from at least (k−1) other records with respect to the pre-determined quasi-identifier. A group of records that are indistinguishable to each other is often referred to as an equivalence class.

The main goal of this research work is to introduce a new k-Anonymity algorithm which is capable of transforming a non anonymous data set into a k-Anonymity data set. The transformation is aimed to achieve good performance of a clustering algorithm trained on the transformed data set as similar as possible to the performance of a cluster trained on the original data set. Clustering is the problem of partitioning a set of objects into groups such that objects in the same group are more similar to each other than objects in other groups with respect to some defined similarity criteria.

A. K-ANONYMITY MODEL

K-anonymity [3, 4, and 18] is a property that captures the protection of released data against possible re-identification of the respondents to whom the released data refer. K-anonymity is an anonymizing approach proposed by Samarati and Sweeney [4]. A data set complies with k-anonymity protection if each individual's record stored in the released data set cannot be distinguished from at least k - 1 individual whose data also appear in the data set. The k-anonymity model makes two major assumptions:

1) The database owner is able to separate the columns of the table into a set of quasi-identifiers, which are attributes that may appear in external tables the database owner does not control, and a set of private columns, the values of which need to be protected. We prefer to term these two sets as public attributes and private attributes, respectively.

2) The attacker has full knowledge of the public attribute values of individuals, and no knowledge of their private data. The attacker only performs linking attacks. A linking attack is executed by taking external tables containing the identities of individuals, and some or all of the public attributes. When the public attributes of an individual match the public attributes that appear in a row of a table released by the database owner, then we say that the individual is linked to that

row. Specifically the individual is linked to the private attribute values that appear in that row. A linking attack will succeed if the attacker is able to match the identity of an individual against the value of a private attribute.

B. GENERALIZATION

Generalization consists of substituting attribute values with semantically consistent but less precise values. For example, the month of birth can be replaced by the year of birth which occurs in more records so that the identification of a specific individual is more difficult. Generalization maintains the correctness of the data at the record level but results in less specific information that may affect the accuracy of machine learning algorithms applied on the k-anonymous data set. Different systems use various methods for selecting the attributes and records for generalization as well as the generalization technique [2].

Generalization can be applied at the following levels

1) *Attribute (AG):* generalization is performed at the level of column; a generalization step generalizes all the values in the column.

2) *Cell (CG):* generalization is performed on single cells; as a result a generalized table may contain, for a specific column, values at different generalization levels. For instance, in the date of birth column.

Another method applied in conjunction with generalization to obtain k-Anonymity is tuple suppression. The intuition behind the introduction of suppression is about the additional method which reduces the amount of generalization to satisfy the k-anonymity constraint. Suppression is also used to moderate the generalization process when there is a limited number of outlier.

C. SUPPRESSION

Suppression refers to removing a certain attribute value and replacing occurrences of the value with a special value "?," indicating that any value can be placed instead. Suppression can drastically reduce the quality of the data if not properly used [17].Suppression can be applied at the following levels

1) Tuple (TS): suppression is performed at the level of row; a suppression operation removes a whole tuple.

2) Attribute (AS): suppression is performed at the level of column; a suppression operation obscures all the values of the column.

3) *Cell (CS):* suppression is performed at the level of single cells; as a result a k-anonymized table may wipe out only certain cells of a given tuple/attribute.

V. PROBLEM FORMULATION

The proposed methodology is used for analyzing K-anonymization of data using partitional hierarchical approach. The framework for the work is given in the Fig. 1.

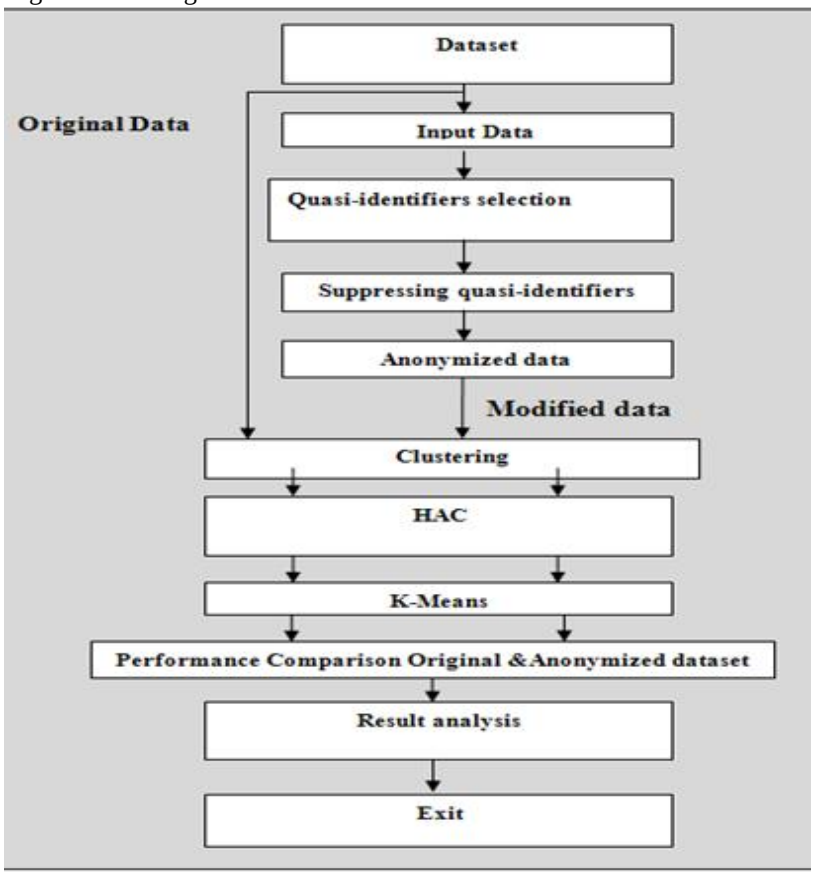

Fig.1 Framework

The framework consists of two phases. In the first phase the k-anonymization of data is done, in the second phase k-means clustering is applied to the k-anonymized dataset. The methodology consists of the following steps:

1) Dataset
2) Quasi-identifiers selection
3) Suppression
   i. Clustering for original dataset.
   ii. Clustering for anonymized dataset.

4) Compare the k-means clustering performance of the original dataset and modified dataset. Using the given performance factors.
   i. Clustering accuracy for original and anonymized dataset
   ii. Time Complexity.
   iii. Information loss.

D. Clustering

Clustering is a division of data into groups of similar objects. Representing the data by fewer clusters necessarily loses certain fine details, but achieves simplification. It models data by its clusters. Data modeling puts clustering in a historical perspective rooted in mathematics, statistics, and numerical analysis. The objective of the problem is the anonymized data can be used for knowledge extraction using data mining techniques. In this work k-means clustering data mining technique is applied, to the anonymized dataset. In order to verify the performance, k-means clustering algorithm is applied to the original dataset also. Then the clustering results of both original and anonymized data are tested.

E. HAC Algorithm

The HAC algorithm takes numeric data as the input and generates the hierarchical partitions as the output. Therefore it is applied in first clustering step to group data into subsets. In HAC, initially each object is considered as a cluster. Then by merging the closest clusters iteratively until the termination condition is reached, or the whole hierarchy is generated. It generates different levels of clusters bottom-up.

1) Calculate the distance between every two objects.
2) View each object as an individual cluster.
3) Merge the closest two clusters.
4) Update the distance between clusters.

Repeat 3-4 until reaching a stopping criterion or generating the whole hierarchy.

F. K-means Algorithm

The k-means algorithm takes numeric data as input, and generates crispy partitions (i.e., every object only belongs to one cluster) as the output. It is one of the most popularly used clustering algorithms in the research community. It has been shown to be a robust clustering method in practice. Therefore, the k-means algorithm is applied in second clustering step to cluster data sets. K-means starts by randomly selecting or by specifically picking k objects as the centroids of *k* clusters. Then k-means iteratively assigns the objects to the closest centroid based on the distance measure, and updates the mean of objects in this cluster as the new centroid until reaching a stopping criterion. This stopping criterion is based on either non-changing clusters or a predefined number of iterations.

1) Select first k objects randomly as the centroid of each cluster.
2) Assign each object to the closest cluster based on Euclidean distance or cosine similarity.
3) Update the centroid of each cluster.

Repeat steps 2-3 until stopping criterion is reached.

Algorithm

The k-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster

Input

- K : the number of clusters
- D : a data set containing n objects

Output: Set of k clusters

Method

1) Arbitrarily choose k objects from D as the initial cluster centers;
2) Repeat
3) (Re) assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster
4) Update the cluster means, i.e. calculate the mean value of the objects for each cluster
5) until no change

Apply the k-means clustering for both original and modified data set to get the clusters.

VI. EXPERIMENTAL EVALUATION

G. DATASET

The Adult dataset is downloaded from UC Irvine Machine Learning Repository. The Donor of the dataset is Ronny Kohavi and Barry Becker. This dataset contains the census data and has become a commonly used benchmark for K-Anonymity. The Adult dataset consists of fifteen fields with six continuous attributes and eight categorical attributes. The class attribute is income level with, two possible values, <= 50K or > 50K. The Adult data contains about 32,561 records totally.

Age	Work class	Flnwgt	Edu	Edu-num	Marital status	Occupation	Relation ship	Race	Sex	Capital gain	Capital loss	Hours per week	Native country	Salary
39	State-gov	77516	Bachelor	13	Never married	Clerk	Not in family	White	M	2174	0	40	Us	<=50 k
50	Self-not inc	83311	Bachelor	13	Married civ spouse	Manager	Husband	White	M	0	0	13	Us	<=50 k
38	Private	21564	Hs grad	9	Divorced	Cleaner	Not in family	White	M	0	0	40	Us	<=50 k
53	Private	234721	Hs grad	7	Married civ spouse	Cleaner	Husband	Black	M	0	0	40	Us	<=50 k
28	Private	338409	Bachelor	13	Married civ spouse	Professor	wife	Black	F	0	0	40	Cuba	<=50 k
37	Private	284582	Master	14	Married civ spouse	Manager	wife	White	F	0	0	40	Us	<=50 k
49	Private	160187	9th	5	Married civ spouse	Other service	Not in family	Black	F	0	0	16	Jamaica	<=50 k
52	Self-not inc	209642	Hs grad	9	Married civ spouse	Professor	Husband	White	M	0	0	45	Us	>50k
31	Private	45781	Master	14	Never married	Manager	Not in family	White	F	14084	0	50	Us	>50k
42	Private	159449	Bachelor	13	Married civ spouse	Manager	Husband	White	M	5178	0	40	Us	>50k

TABLE. I. Sample Dataset

H. Quasi-identifiers

Quasi-identifiers are set of features whose associated values may be useful for linking with another data set to reidentify the entity that is the subject of the data in the adult dataset, *{age, work class, education, marital status, occupation, race, gender,* and *native country}* are considered as the quasi-identifiers. Among these, *age* and *education* were treated as numeric attributes while the other six attributes were treated as categorical attributes. The quasi-identifiers are highlighted in the table below.

Age	**Work class**	Flnwgt	**Edu**	Edu-num	**Marital status**	**Occupation**	Relation ship	**Race**	**Sex**	Capital gain	Capital loss	Hours per week	**Native country**	Salary
39	**State-gov**	77516	**Bachelor**	13	**Never married**	**Clerk**	Not in family	**White**	**M**	2174	0	40	**Us**	<=50 k
50	**Self-not inc**	83311	**Bachelor**	13	**Married civ spouse**	**Manager**	Husband	**White**	**M**	0	0	13	**Us**	<=50 k
38	**Private**	21564	**Hs grad**	9	**Divorced**	**Cleaner**	Not in family	**White**	**M**	0	0	40	**Us**	<=50 k
53	**Private**	234721	**Hs grad**	7	**Married civ spouse**	**Cleaner**	Husband	**Black**	**M**	0	0	40	**Us**	<=50 k
28	**Private**	338409	**Bachelor**	13	**Married civ spouse**	**Professor**	wife	**Black**	**F**	0	0	40	**Cuba**	<=50 k
37	**Private**	284582	**Master**	14	**Married civ spouse**	**Manager**	wife	**White**	**F**	0	0	40	**Us**	<=50 k
49	**Private**	160187	**9th**	5	**Married civ spouse**	**Other service**	Not in family	**Black**	**F**	0	0	16	**Jamaica**	<=50 k
52	**Self-not inc**	209642	**Hs grad**	9	**Married civ spouse**	**Professor**	Husband	**White**	**M**	0	0	45	**Us**	>50k
31	**Private**	45781	**Master**	14	**Never married**	**Manager**	Not in family	**White**	**F**	14084	0	50	**Us**	>50k
42	**Private**	159449	**Bachelor**	13	**Married civ spouse**	**Manager**	Husband	**White**	**M**	5178	0	40	**Us**	>50k

TABLE. II. Quasi-identifier selection

I. Suppression

Privacy can be preserved by simply suppressing all sensitive data before any disclosure or computation occurs. Given a database, we can suppress specific attributes in particular records as dictated by the privacy policy. To achieve k-anonymity, quasi-identifier attributes are completely or partially suppressed. A particular suppression policy is chosen to maximize the utility of the k-anonymized data set. Suppression refers to removing a certain attribute value and replacing occurrences of the value with a special value "?," indicating that any value can be placed instead.

Age	Work class	Flnwgt	Edu	Edu-num	Marital status	Occupation	Relationship	Race	Sex	Capital gain	Capital loss	Hours per week	Native country	Salary
39	***	77516	Bachelor	13	Never married	***	Not in family	White	M	2174	0	40	Us	<=50k
39	Self-not inc	83311	Bachelor	13	Married civ spouse	Manager	Husband	White	M	0	0	13	Us	<=50k
***	Private	21564	Hs grad	9	***	Cleaner	Not in family	White	M	0	0	40	Us	<=50k
***	Private	234721	Hs grad	7	Married civ spouse	Cleaner	Husband	Black	M	0	0	40	Us	<=50k
***	Private	338409	Bachelor	13	Married civ spouse	Professor	wife	Black	F	0	0	40	***	<=50k
42	Private	284582	Master	14	Married civ spouse	Manager	wife	White	F	0	0	40	Us	<=50k
***	Private	160187	***	5	Married civ spouse	***	Not in family	Black	F	0	0	16	***	<=50k
***	Self-not inc	209642	Hs grad	9	Married civ spouse	Professor	Husband	White	M	0	0	45	Us	>50k
***	Private	45781	Master	14	Never married	Manager	Not in family	White	F	14084	0	50	Us	>50k
42	Private	159449	Bachelor	13	Married civ spouse	Manager	Husband	White	M	5178	0	40	Us	>50k

TABLE. III. Suppressed Data

J. Clustering

Clustering is a method of grouping data into different groups, so that the data in each group share similar trends and patterns.

1) Clustering on original data

Clustering on original test data consists of 1000 records. Clustering technique is applied for both numerical and categorical data for quasi-identifier attributes using k-means clustering. Clustering on original data consists of 5 groups of clusters which are derived using k-means clustering technique.

2) Clustering on modified data

Clustering on modified data also consists of 1000 records. Clustering for modified data is based on the concept of k-anonymization. The anonymized data is released for clustering and accuracy of the dataset is tested here. The k-anonymity requirement can be naturally transformed into a clustering problem where the aim is to find a set of clusters. Each of which contains at least k records. In order to maximize data quality, the records in a cluster must be as similar to each other as possible.

3) Performance factors for clustering

The performance factors for clustering the original and modified data can be calculated by the following.

i. Accuracy.
ii. Information loss
iii. Time complexity.

i. Accuracy

The accuracy of the dataset is calculated based on the suppressed k-value. The accuracy of the original and modified dataset is implemented and tested by the subsets of the Adult dataset with different sizes such as 500, 1000, 5000 and 10000 records namely. The accuracy for 1000 records is defined in fig.2 below.

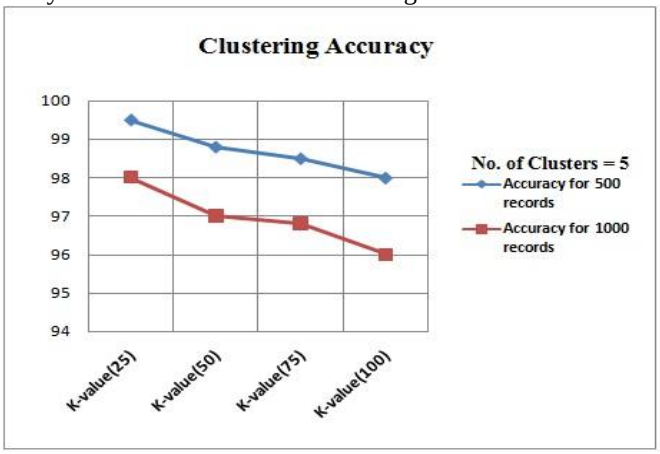

Fig. 2 Clustering Accuracy

ii. Information Loss

The ultimate goal of the clustering problem is the k-anonymization of data, the cost function is formulated to represent the amount of distortion (i.e., information loss) caused by the suppression process. Recall that, records in each cluster are

suppressed to share the same quasi-identifier value that represents every original quasi-identifier value in the cluster. By assuming, that the numeric values are suppressed into a range [min, max] and categorical values into a set that unions all distinct values in the cluster. With these assumptions, the developed metric is, referred to as Information Loss metric (IL), that measures the amount of distortion introduced by the suppression process to a cluster.

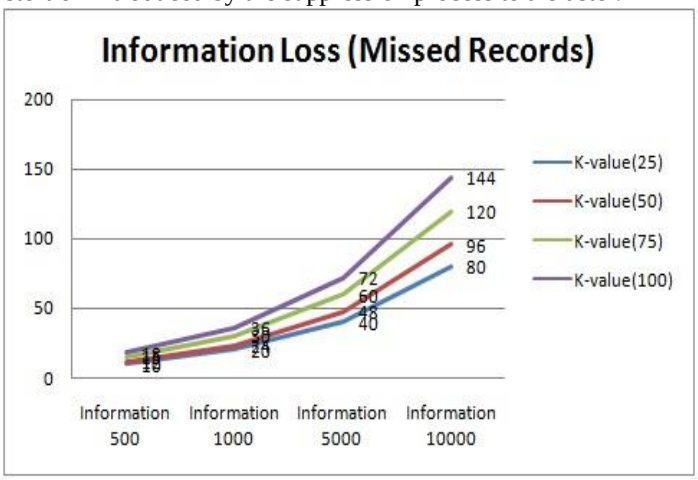

Fig. 3 Information Loss

iii. Time Complexity

The time complexity of the dataset is calculated based on the suppressed k-value. The time complexity of the original and modified dataset is implemented and tested by the subsets of the adult dataset with different sizes such as 500, 1000, 5000 and 10000 records namely. The time complexity for 500 records is defined in fig.4 below.

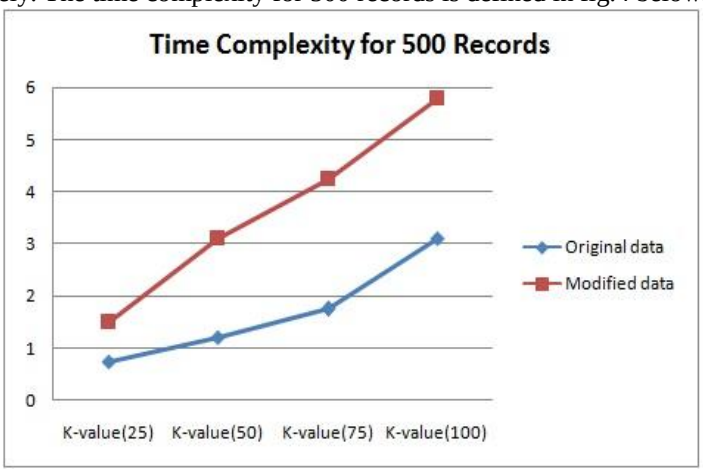

Fig. 4 Time Complexity

As expected, the k-anonymization algorithm is modified to minimize k-means clustering errors after anonymizing the original data also. The execution time of the algorithms for original dataset and anonymized dataset is based on different k values.

VII. CONCLUSIONS

K-anonymity is one category which has recently been investigated as an interesting approach to protect microdata undergoing public or semi-public release from linking attacks. In this research work, the quasi-identifiers available in the data base are modified using suppression based on the threshold k-value. In the modified data base, a clustering algorithm is applied to verify whether the anonymized data is used for data analysis or not. Clustering algorithms has been widely applied to various domains to explore the hidden and useful patterns inside data. Because the most collected data in real world contain both categorical and numeric attributes, the traditional clustering algorithm cannot handle this kind of data effectively. The proposed approach uses the idea of clustering to minimize information loss and thus ensure good data quality. As part of the approach a new suitable metric is developed to estimate the information loss introduced by Suppression, which works for both numeric and categorical data. In future a new clustering technique can be developed to overcome the information loss and time complexity of k-means clustering algorithm.

REFERENCES

[1] A. Friedman, R. Wolff, and A. Schuster, "*Providing k-Anonymity in Data Mining*," Int'l J. Very Large Data Bases, vol. 17, no. 4, pp. 789-804, 2008.

[2] A. Friedman, R. Wolff, and A. Schuster, "*Providing k-Anonymity in Data Mining,"* Int'l J. Very Large Data Bases, vol. 17, no. 4, pp. 789-804, 2008.

[3] 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.

[4] B.C.M. Fung, K. Wang, and P.S. Yu, "*Anonymizing Classification Data for Privacy preservation,"* IEEE Trans. Knowledge and Data Eng., vol. 19, no. 5, pp. 711-725, May 2007.

[5] Fayyad, U.M. (2003). Editorial. *SIGKDD Explorations*, 5(2).

[6] J. Domingo-Ferrer, "*A new privacy homomorphism and applications. Source",* Information Processing Letters archive. Volume 60, Issue 5 (December 1996)

[7] J. Roberto, Jr. Bayardo, and A. Rakesh, "*Data Privacy through Optimal k-Anonymization,"* Proc. Int'l Conf. Data Eng., vol. 21, pp. 217-228, 2005

[8] K. LeFevre, D.J. DeWitt, and R. Ramakrishnan, "*Incognito: Efficient Full Domain k-Anonymity*," Proc. 2005 ACM SIGMOD, pp. 49-60, 2005.

[9] K. LeFevre, D.J. DeWitt, and R. Ramakrishnan, "*Mondrian Multidimensional k-Anonymity*," Proc. 22nd Int'l Conf. Data Eng., p. 25, Apr. 2006.

[10] K. Wang, P.S. Yu, and S. Chakraborty, "*Bottom-Up Generalization: A Data Mining Solution to Privacy Protection*," Proc. Fourth IEEE Int'l Conf. Data Mining, pp. 205-216, 2004.

[11] L. Sweeney, "*Achieving k-Anonymity Privacy Protection Using Generalization and Suppression,"* Int'l J. Uncertainty, Fuzziness, and Knowledge-Based Systems, vol. 10, no. 5, pp. 571-588, 2002.

[12] L. Sweeney, "*k-Anonymity: A Model for Protecting Privacy*," Int'l J. Uncertainty, Fuzziness, and Knowledge-Based Systems, vol. 10, no.5, pp. 557-570, 2002.- generalization.

[13] M.S. Wolf and C.L. Bennett, *"Local Perspective of the Impact of the HIPAA Privacy Rule on Research,"* Cancer-Philadelphia Then Hoboken, vol. 106, no. 2, pp. 474-479, 2006.

[14] *"On k-Anonymity and the Curse of Dimensionality"*. Charu C. Aggarwal. IBM T. J. Watson Research Center. Route 134 & Taconic State Parkway. Yorktown Heights, Year, 2005, Journal, In VLDB.

[15] P. Samarati and L. Sweeney, *"Generalizing Data to Provide Anonymity When Disclosing Information,"* Proc. 17th ACMSIGACT-SIGMOD-SIGART Symp. Principles of Database Systems, vol. 17, p. 188, 1998. PPDM.

[16] S.Kisilvich, L.Rokach, B.Shapira, Y.Elovici, "*An Efficient Multimensional Suppression for k-Anonymity"* IEEE Trans. Knowledge and Data Eng., vol. 22, no. 3, pp. 334-347, March 2010.

[17] S.V. Iyengar, "*Transforming Data to Satisfy Privacy Constraints*," Proc. Eighth ACM SIGKDD, pp. 279-288, 2002.

[18] Vijay S. Iyengar, "*Transforming Data to Satisfy Privacy Constraints,"* Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and Data mining *portal.acm.org/* Pages: 279 - 288 Year of Publication: 2002 ISBN: 1-58113-567-X

[19] Yehuda Lindell and Benny Pinkas, "*Privacy preserving data mining,"* year [2000].

