Accuracy and Utility Balanced Privacy Preserving Classification Mining by Improving K-Anonymization

Naga Prasanthi Kundeti [1], Chandra Sekhara Rao MVP [2]

[1] *Department. of Computer Science and Engineering*, Acharya Nagarjuna University, Guntur, Andhra Pradesh, India.
email:prasanthi.kundeti@gmail.com

[2] *Department. of Computer Science and Engineering*, RVR & JC College of Engineering, Guntur, Andhra Pradesh, India.
email: manukondach@gmail.com

***Abstract* - The data available is vast and data is being analyzed to improve businesses. This data analysis also contributes to society in different ways. Now there are new challenges to protect privacy of data. So, Privacy Preserving Data Mining (PPDM) techniques have evolved which protect the privacy of data while carrying out data analysis. Privacy Preserving Data Publishing (PPDP) is a part of PPDM which is a major research area. As part of PPDP several anonymization algorithms are proposed. K-anonymization is one among them. In this paper a new method for privacy preserving data mining is proposed which is better than applying k-anonymization alone. The present research work focuses on the approach which decreases the risk of various attacks and at the same time provides more utility of data.**

Keywords - K-anonymization, data mining, privacy preserving data mining.

I. INTRODUCTION

Currently, there is exponential increase in volumes of data generated every year. This data contains large amounts of personal information also. Businesses and organizations collect data for various data analysis tasks. Entities release data collections publicly or to third parties for data analysis. While sending data to organizations, there is a chance of misusing personal information during this process. This has brought a new challenge to the researchers to protect privacy of data. So, data is anonymized before being given for publication. This way of preserving privacy during publishing of data is called Privacy Preserving Data Publishing (PPDP).

Generally privacy threats can be of different types and they are given below:

Membership Disclosure: This kind of privacy threat allows an assailant to check whether an individual's data is present in a data set or not and to deduce some meta-information about an individual. This kind of privacy threat handles only implicit sensitive attributes not explicit sensitive attributes.

Attribute disclosure: In this an individual need not be linked to a specific entity in data set. But still Attribute disclosure attack may occur. Sensitive attributes are those attributes in a dataset which the individuals desirous to be kept unrevealed. If these sensitive attributes are revealed it will cause damage to data owners' privacy. Information can be inferred by linking set of data entries that contain same sensitive attribute value.

Identity disclosure (or re-identification): In this a particular person can be directly associated with a specific data entry in a data set. An assailant can discover all the sensitive information present in the dataset about an individual. This kind of attack has many legal consequences and is considered to be illicit.

In order to avoid above disclosure risks in a dataset, the attributes of the input dataset are categorized into different types. They are:

Identifying Attributes: These attributes have a high re-identification risk. These attributes will be discarded from the dataset. For ex: employee Id, employee name, SSN etc.

Quasi-identifying attributes: These attributes can be joined with some external information and can be used to re-identify an individual. So, these attributes need to be transformed in order to avoid de-identification risks. For ex: profession, gender, ZIP codes and date of birth.

Sensitive attributes: These attributes include some important information about individuals which should not be leaked. Attackers are interested in such important information and try to misuse that information and harm the data owners. These attributes are not modified but they are applied with some constraints like t-closeness or l-diversity. For ex: disease.

Insensitive Attributes: These attributes do not cause any privacy risks and are unchanged.

One of the possible solution is to exclusively remove attributes that identify users or that contain some sensitive information about users from data before publishing the data. But this approach is not effective [1]. Even though individual identifiers are removed, there is still a chance to combine different data sets or obtain background knowledge about people and make some inferences from them. A quasi identifier is not like an identifier attribute that

explicitly identifies a user. A quasi identifier is merged with other data available from public repositories to re-identify the owner of the record. This is termed as linkage attack [2]. Using quasi identifiers (QIDs) and linking them to some background knowledge enables them to identify particular users. For example, attributes like gender, ZIP code etc. are called as quasi identifiers. Selecting the appropriate algorithm for given data is tedious task for several algorithms are available in the area of PPDP.

From 1990 U.S. Census Data Sweeney was able to especially recognize a person in the US using quasi identifiers in a survey conducted in 2000[3]. Sweeney was able to extract the Massachussetts governor's medical record by linking the voters list anonymized medical records in Group Insurance Commission (GIC) using quasi identifiers [4].

Data in a database can be anonymized by applying various privacy preserving techniques. Some of them are Generalization, Suppression, Anatomization and Perturbation.

- Generalization: In this method a data value is replaced with a more generalized one. For numerical attributes, a particular data value may be replaced with a range of values as a generalized one. For categorical attributes generalization is performed using a hierarchy. For example, engineer and lawyer are some of the data values for occupation which can be replaced with a more generalized value of 'professional'.
- Suppression: This method prevents information disclosure by eliminating some attribute values. Generally replacing the original data value with("*").
- Anatomization [5]: In this, sensitive attributes and quasi identifiers are placed in two different Tables so that linking QIDs to sensitive attributes become very difficult.
- Perturbation: In this, original data values are replaced with synthetic values with the same statistical information.

Samarati and Sweeny [6], [7] proposed the most popular privacy model namely k-anonymization. According to [8] k-anonymity for a table is defined as follows [8]:

"Let T(A1,...,An) be a table.
Let QI be the set of quasi-identifiers corresponding to table T.
T fulfils k-anonymity property with respect to QI if and only if each sequence of values in T [QI] appears at least with k occurrences in T[QI]".

In k-anonymity model, QIDs are altered through generalization and suppression and create record groups (also called Equivalence classes) that have same QID values. In order to generalize the attribute, Value Generalization Hierarchy (VGH) is used. For example, consider the VGH for age as given below.

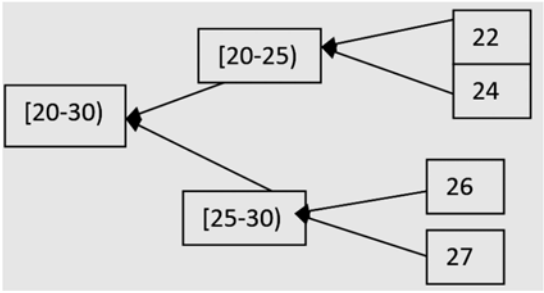

Fig.1: VGH for age attribute of Adult dataset

For example, consider a table of criminal records. The fields are name, marital status, age, zip-code and crime. Among these fields, name is the identifier attribute (ID), marital status, age and zip-code are quasi identifiers (QIDs) and crime is sensitive attribute (SA). In order to perform anonymization, ID is removed and QIDs are generalized using a single-dimensional generalization scheme. Marital status field is replaced with more generic value, age is replaced with ranges of values and last digit of ZIP code is replaced with"*".

TABLE 1. CRIMINAL RECORDS DATA

	Identifier	Quasi Identifiers			Sensitive Attribute
Record No.	Name	Marital Status	Age	ZIP Code	Offence
1	Annie	Separated	28	81042	Assassination
2	Bob	Single	21	81021	Burglary
3	Dennis	Wid0wed	22	81024	Peddling
4	John	Separated	29	81046	Beat
5	Lily	Wid0wed	26	81045	Illegal Copying
6	Simon	Single	23	81027	Obscenity

TABLE II: ANONYMOUS TABLE FOR CRIMINAL RECORDS DATA

		Quasi Identifiers			Sensitive Attribute
Record No.	EQ	Marital Status	Age	ZIP Code	Offence
1	1	Not Married	[25-30)	8104*	Assassination
4		Not Married	[25-30)	8104*	Beat
5		Not Married	[25-30)	8104*	Illegal Copying
2	2	Not Married	[20-25)	8102*	Burglary
3		Not Married	[20-25)	8102*	Peddling
6		Not Married	[20-25)	8102*	Obscenity

II. ANONYMIZATION ALGORITHMS

There are several algorithms specified for k-anonymization in literature. Among these three algorithms utilizing generalization and suppression are chosen. They are: (i) Sweeney's algorithm Datafly [9], (ii) Incognito Algorithm [10] and (iii) Mondrian Algorithm [11].

There are several tools available for automatically carrying out k-anonymization. ARX Data Anonymization Tool is one of those software tools used in privacy preserving data publishing. It provides a wide range of privacy models and many statistical disclosure control

methods. The user can specify threshold values for these methods. The data is transformed using a combination of two methods namely (a) global recoding with full domain generalization of attribute values and (b) local recoding with record suppression. Using ARX tool it is possible to calculate optimal solution with minimal loss of data quality.

Instead of traditional k-anonymization algorithms, ARX implements a new globally optimal anonymization algorithm called, Flash. Flash constructs a search space containing different transformations on data. It identifies a transformation that has minimal loss. As it constructs the complete solution space, different solutions for anonymization problem can be inspected by users. This algorithm traverses the generalization lattice in bottom-up breadth-first manner and generates paths based on following key ideas:

Flash algorithm uses Predictive tagging. Predictive tagging is suitable when generalization lattice is traversed in vertical manner.

Flash implements a stable strategy which avoids the difficulties in traversing a lattice vertically and its execution time becomes unpredictable with respect to input data set representation.

The algorithm checks all transformations that enable applying multiple optimizations to achieve maximum performance.

Proposed Privacy Preserving Data Mining Approach:

Input: Dataset D

Output: Privacy enabled Dataset D'

The attributes in given dataset were categorized into four categories namely Identifiers, Quasi Identifiers, Sensitive Attributes and Insensitive Attributes.

Identifiers were removed from the given dataset.

The set of Quasi Identifiers were identified. For numerical quasi identifiers geometric data perturbation was applied. Intermediate data set Dm is obtained. For categorical quasi identifiers, k-anonymization algorithm was applied on data set Dm.

The privacy enabled dataset D'm was obtained by applying k-anonymization on Dm.

Same k-anonymization algorithm was applied on dataset D and anonymized dataset D'k was obtained.

Classification algorithms like naive Bayes, J48 etc. were applied on data sets D'm and D'k and then accuracy of classification of both datasets were compared.

vii) It was observed from the results, proposed approach performs better than existing approach.

III. IMPLEMENTATION

In this paper, Adult data set from UCI Machine Learning Repository [12] is considered for evaluation. Adult data set contains following attributes Age(Numeric), Fnlwgt (Numeric), Work class(Text), Education(Text), Education num(Numeric), Marital status(Text), Occupation(Text), Relationship(Text), Race(Text), Sex(Text), Capital gain(Numeric), Hours per week(Numeric), Native country(Text), Capital loss(Numeric) and Class label(Text). Among these attributes Class label attribute is the sensitive attribute. Work class, Education, Age and Native country are marked as quasi identifiers. Value generalization hierarchies (VGH) are defined for these quasi identifiers as given below.

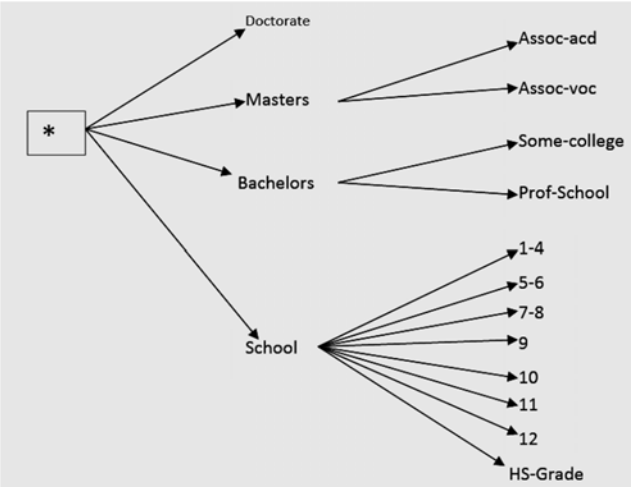

Fig2: VGH for Education attribute of Adult dataset.

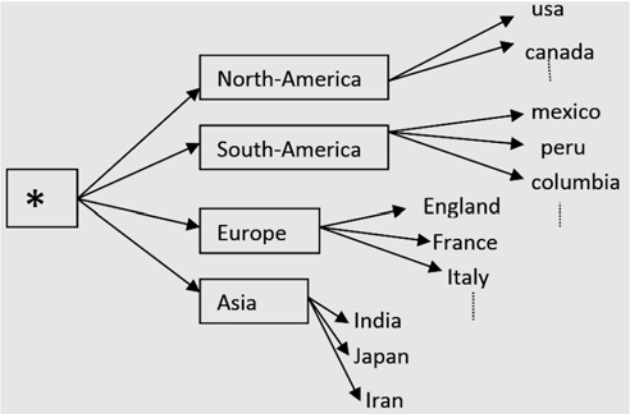

Fig 3: VGH for native country attribute of Adult dataset

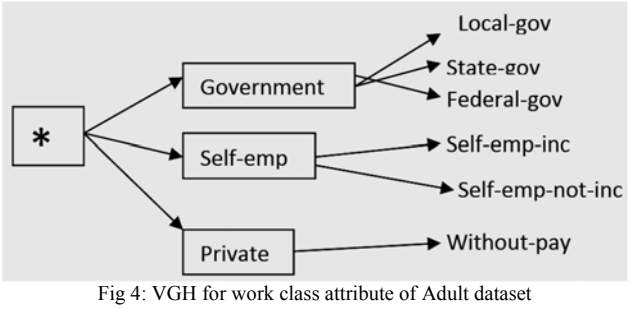

Fig 4: VGH for work class attribute of Adult dataset

First geometric data perturbation [13] was applied on age attribute and the modified Adult Dataset was obtained.

On the modified Adult dataset, K- anonymization was performed for different values of k using ARX anonymization tool [14] using the generalization hierarchies as shown above. k-anonymization is performed on this data with different values for k like 5,15,25,35,45,55,65,75,85. After anonymizing, data classification was applied. Accuracy of classification results varied with varying values of k. Classification algorithms like naive Bayes and J48 were applied on anonymized data using Weka tool. It was observed that accuracy of classification did not vary much with small increase in k-value.

Original Adult dataset was considered, same set of quasi identifiers were used. Now k-anonymization was applied using generalization hierarchies which are specified above. For age attribute, generalization hierarchy was defined [10-20),[20-30),....[80-90). Using ARX tool k-anonymization was applied for different values of k like 5, 15, 25, 35, 45, 55, 65, 75, 85. After anonymization classification algorithms of naive Bayes and J48 were applied and accuracy of classification was tabulated as below.

IV. RESULTS AND DISCUSSION

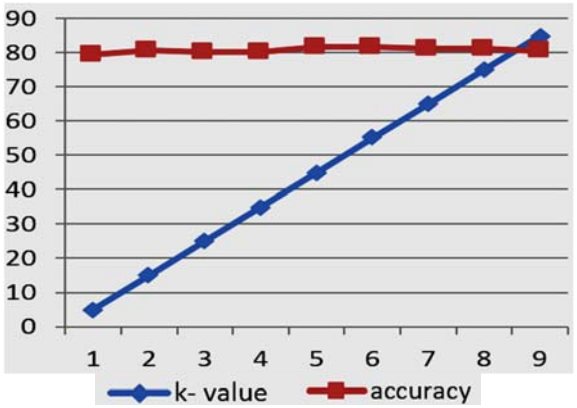

Fig 5: naive Bayes classification (accuracy vs k-value) on original Adult data set

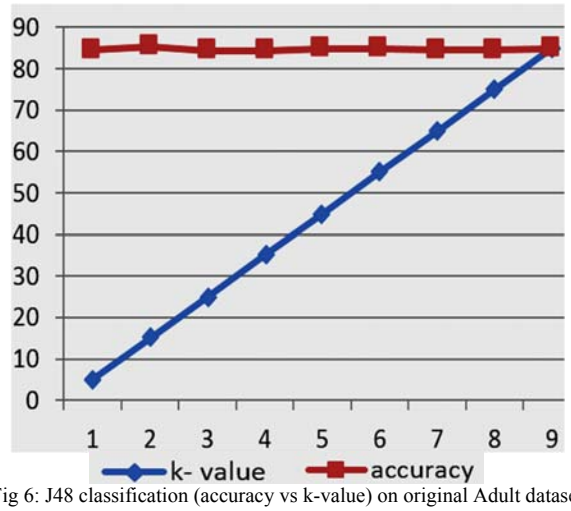

Fig 6: J48 classification (accuracy vs k-value) on original Adult dataset

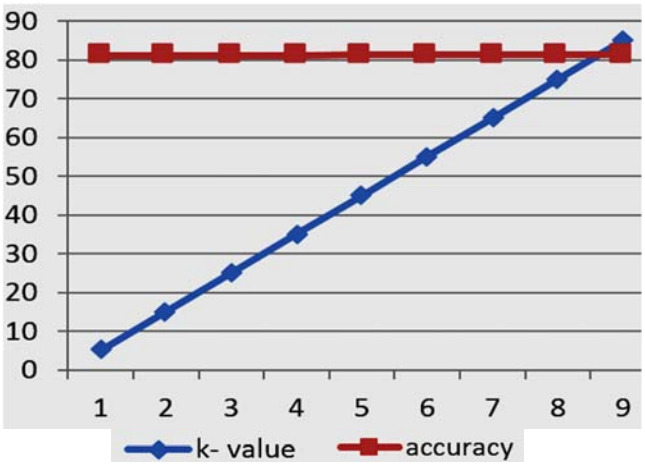

Fig 7: naive Bayes classification (accuracy vs k-value) on perturbed Adult data

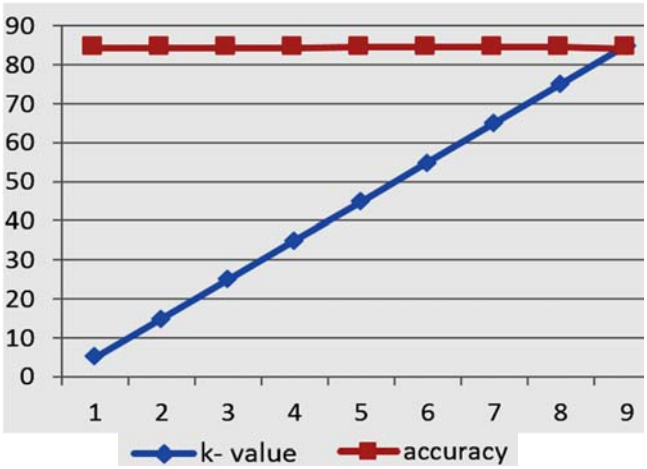

Fig 8: J48 classification (accuracy vs k-value) on perturbed Adult dataset

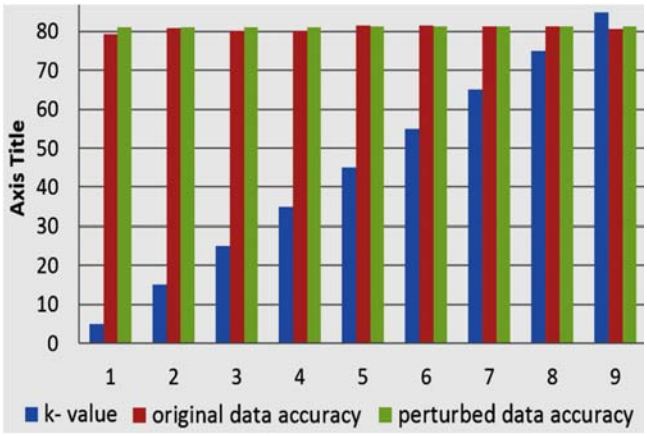

Fig 9: Comparing accuracy of k-anonymization and perturbed k-anonymization results

The figures below show that the Re-identification risks of data are reduced after anonymization.

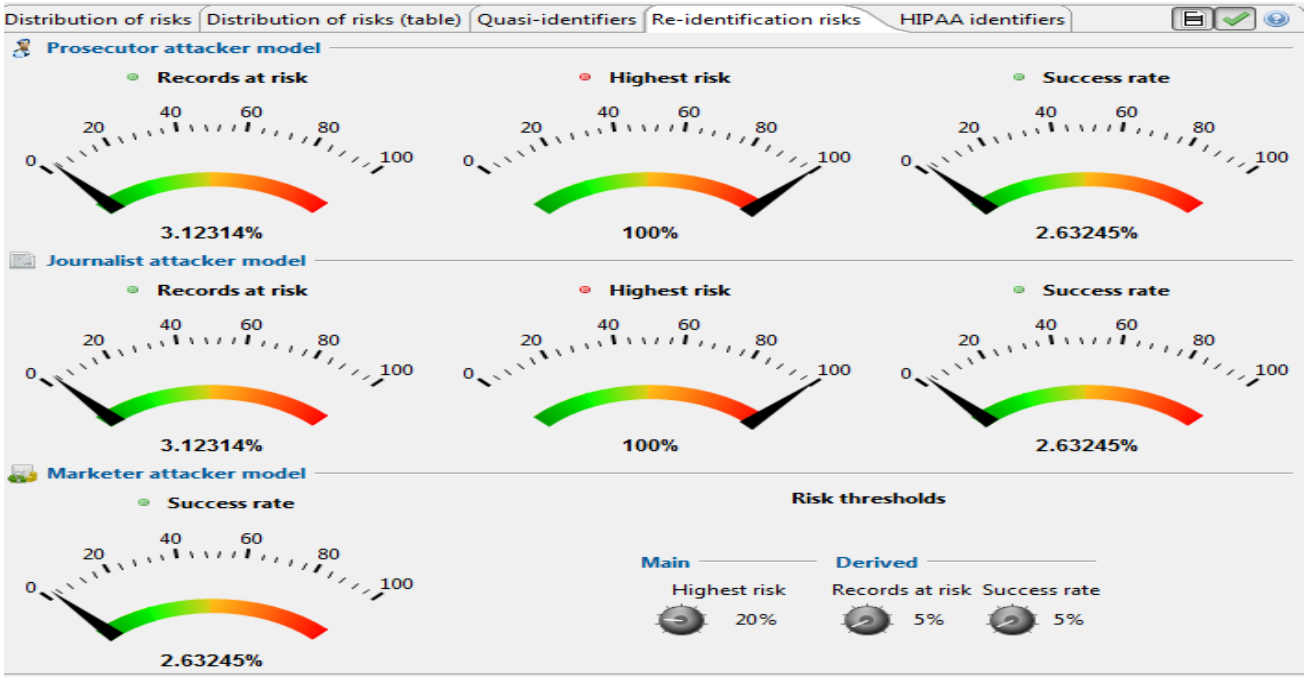

Fig10. Re-identification Risks after perturbation before anonymization (Utility metric = Loss)

It is observed from the above picture that after perturbing the data using geometric data perturbation, the 'Highest risk' is 100% in case of prosecutor attacker model and Journalist attacker model. Even the success rate of attack is 2.63%. The records that are at risk are 3.12%, 3.12% and 2.63% respectively for each type of attacker model.

The Records at risk can be reduced using our novel approach which is depicted in the following figures.

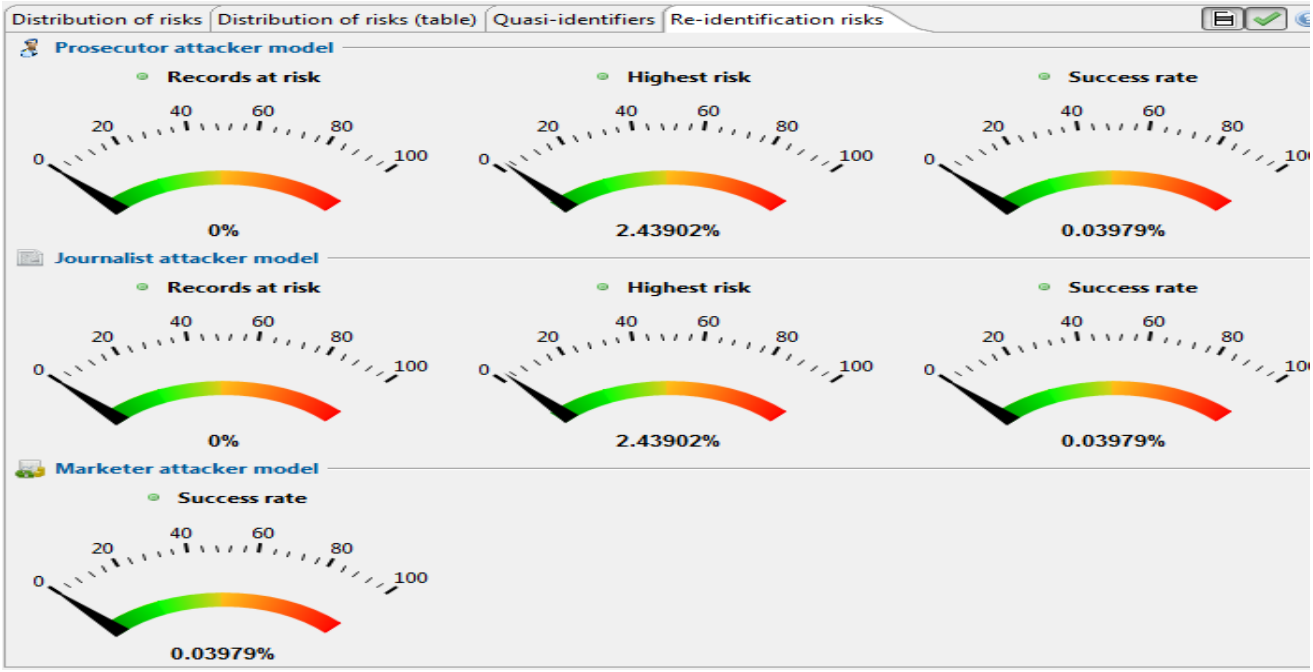

Fig 11. Re-identification risks after perturbation and after anonymization(Utility metric= Loss)

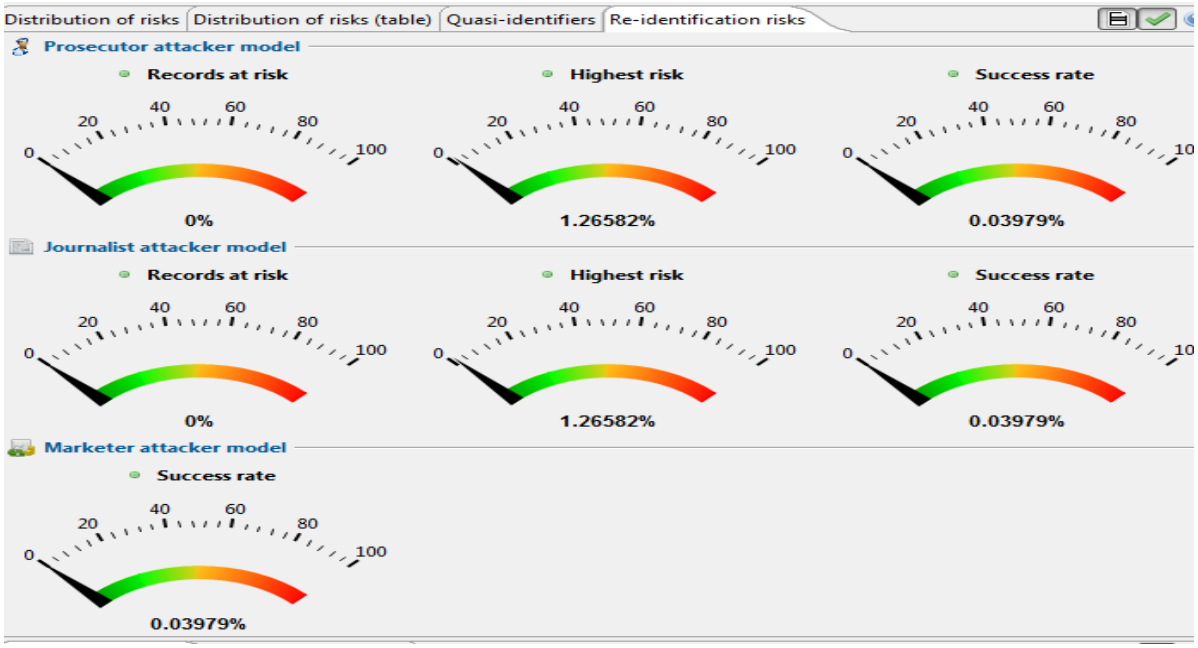

Fig.12 Re-identification risks after applying anonymization on perturbed data (utility metric=Discernibility matrix)

When the novel approach was applied for privacy preserving data mining, the records at risk for various types of attacker models have reduced to 0% as shown in fig11 and fig12. The highest risk was reduced from 100% to 2.4% with 'Loss' as utility metric. The highest risk was reduced from 100% to 1.26% with 'Discernibility matrix' as utility metric. The attacker's success rate in re-identifying or linking the records in anonymized data was reduced to 0%.

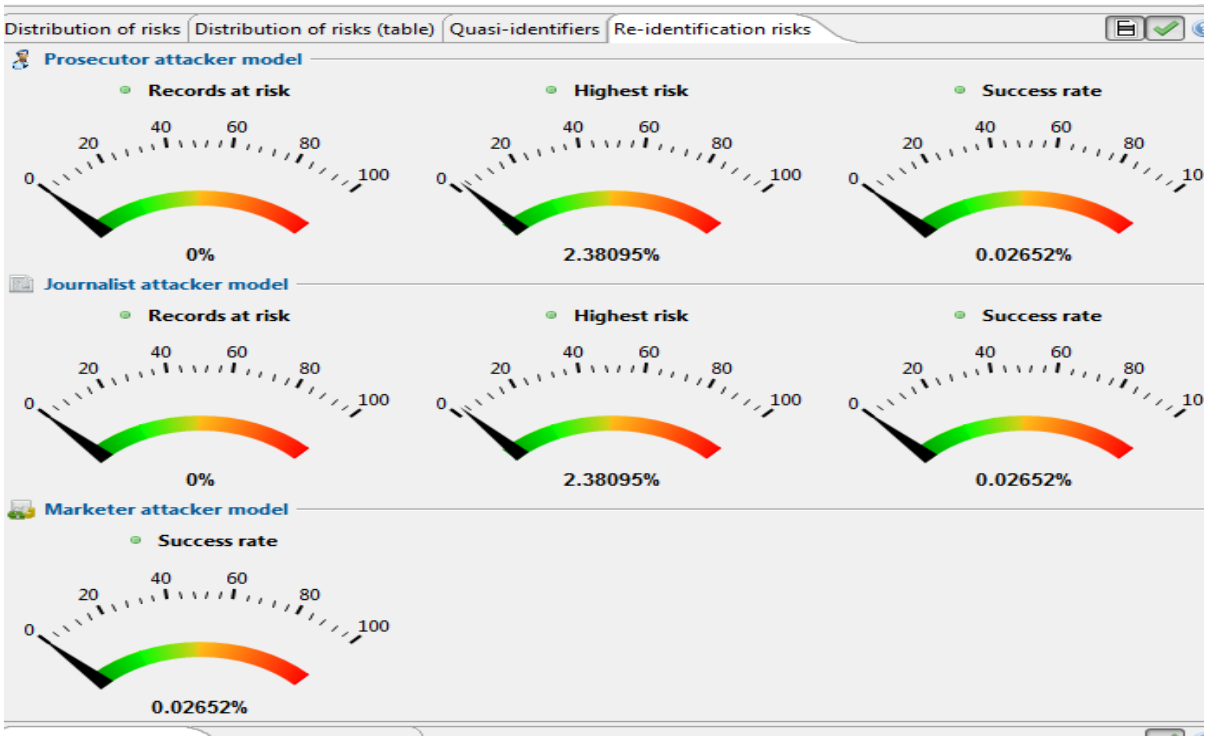

Fig 13 Re-identification risks after applying anonymization on original data(Utility metric=Discernibility matrix)

From fig13, it is observed that records at highest risk are 2.38% when anonymization was applied directly on original data. But using our novel approach the highest risk was reduced to 1.26% as shown in fig12. Hence it is concluded that the novel approach is better than the existing methods of privacy preservation in terms of utility and privacy.

V. CONCLUSION

K-anonymization is confirmed as one of the best privacy preserving data publishing techniques. K-anonymization is performed here without compromising the accuracy of classification. When k-anonymization is applied on geometric perturbed data, accuracy of data is more when compared to applying k-anonymization on original data.

REFERENCES

[1] Shmatikov V and Narayanan A, " How to break anonymity of the netflix prize dataset", arXiv preprint cs/0610105,2006.

[2] Benjamin C.M.Fung, Ke Wang, Rui Chen and Philip S.Yu, "Privacy preserving data publishing", Proc. Workshops of 26th Int. Conf. on Data Engineering, vol.42, no.4, pp. 305-308, 2010.

[3] Latanya Sweeney, " Simple Demographics often identify people uniquely", Health(San Francisco), vol.671, pp.1-34,2000.

[4] Latanya Sweeney, "k-anonymity: A model for protecting privacy", International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol.10, no.05, pp.557-570,2002.

[5] Xiaokui Xiao, Yufei Tao "Anatomy: Simple and Effective privacy preservation", Proceedings of e 32nd international conference on very large database, ACM,pp.150,139,2006.

[6] Pierangela Samarati and Latanya Sweeney, "Protecting Privacy when Disclosing Information: k- anonymity and its Enforcement Through Generalization and Suppression", Proc. of the IEEE Symposium on Research in Security and Privacy, pp. 384-393,1998.

[7] Pierangela Samarati and Latanya Sweeney, "Generalizing data to provide anonymity when disclosing information", in PODS, vol. 98, p. 188, 1998.

[8] Pierangela Samarati, "Protecting respondents identities in microdata release", IEEE Transactions on Knowledge and Data Engineering , Volume: 13, issue 6, Nov/Dec 2001.

[9] Latanya Sweeney, "Achieving k-anonymity privacy protection using generalization and suppression", International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, volume.10, issue 5,p.571-588, 2002.

[10] LeFevre K., DeWitt D.J., and Ramakrishnan R. "Incognito: Efficient Full-domain K-Anonymity", In Proceedings og the 2005 ACM SIGMOD International Conference on Management of Data, SIGMOD'05, pages49-60,2005.

[11] LeFevre K., DeWitt D.J., and Ramakrishnan R."Mondrian Multidimensional K-Anonymity". In Proceedings of the 22nd International Conference on Data Engineering, ICDE'06,page 25,2006.

[12] Ronny Kohavi and Barry Becker, UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Adult, CA: University of California, School of Information and Computer Science.

[13] Keke Chen, Ling Liu, "Geometric data perturbation for privacy preserving outsourced data mining", Knowledge Information and Systems,2010

[14] https://arx.deidentifier.org/anonymization-tool/

[15] Prasser F., Kohlmayer F., Lautenschlaeger R., and Kuhn K.A. ARX - a comprehensive tool for anonymizing biomedical data. In Proceedings of the AMIA Annual Symposium, pages 984–993, 2014.