ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2016

Managing Privacy and Security of Healthcare Data Using Enhanced Slicing

Ronica Raj, Veena Kulkarni, Anand Khandare

M.E. Student, Dept. of Computer Science, Thakur College of Engineering and Technology, Mumbai, India

Assistant Professor, Dept. of Computer Science, Thakur College of Engineering and Technology, Mumbai, India

Assistant Professor, Dept. of Computer Science, Thakur College of Engineering and Technology, Mumbai, India

ABSTRACT: Privacy and Security of data is a major concern for all kinds of industries. Be it medical, I.T sector, public institutions or government organizations, all of them needs to store data and keep them safe and secure from data hackers. Hence there have been various privacy preservation and encryption algorithms proposed to keep the data secure. Privacy preserving data mining (PPDM) protects the privacy of sensitive data without losing the usability of the data. This paper discusses about various PPDM techniques proposed and also proposes a new PPDM technique known as Enhanced Slicing which protects the data against membership and attribute disclosure and also prevents identity disclosure and is a better version of Overlap Slicing in terms of privacy. We also analyze the execution time, performance and accuracy of both the algorithms based on the cardinality of the data. The results of the algorithm and the execution time are shown in a graphical format. The database is encrypted using the RSA algorithm.

KEYWORDS: Privacy, Security, PPDM, Overlap Slicing, Enhanced Slicing, RSA.

I. INTRODUCTION

There has been tremendous and enormous growth in healthcare data with the development of electronic patient records. These data are shared with various other people such as physician and data analyst for analysis of data and many other purposes. Such sharing of data creates privacy and security issues as these records can be exposed to third parties and to unauthorized parties as well. Hence to assure the privacy and security of such sensitive data there have been many techniques introduced in privacy preserving data mining

Privacy and security plays an important role in data mining. This paper implements a secure and private data management framework for both security and privacy of healthcare data. The security of data is ensured by encrypting the data by using RSA algorithm and privacy is assured by using annonymization based PPDM techniques.

The three most widely used techniques of PPDM are generalization, bucketization and Slicing. Bucketization does not prevent membership disclosure and it does not apply for data that do not have a clear distinction between quasi-identifiers and sensitive attribute. Generalization loses high amount of data and does not preserve identity disclosure Slicing provides better data utility but still its prone to attacks. Slicing protects the data against membership and attribute disclosure but it does not provide any details about identity disclosure. Overlap slicing does not prevent identity disclosure.

Hence to overcome these disadvantages, an efficient technique has been introduced in this paper known as Enhanced Slicing which protects the data against membership and attribute disclosure and also prevents identity disclosure.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2016

Table 1: Original Dataset

PatientID	NAME	AGE	GENDER	WEIGHT	HEIGHT	DIEASE	SYMPTOMS	ALLERGY	BLOODGROUP
2278392	Sudhakar	55	Male	85	5.5	Cancer	Weight loss	Skin	B+
149190	George	60	Male	88	5.4	Asthma	Cough	Food	A+
64410	Satish	45	Male	80	5.2	Candidiasis	Red Rash	Dust	AB+
16680	Gopal	68	Male	90	5.7	Calculi	Nausea	Pet	O-
35754	Nataraj	70	Male	81	5.1	Cancer	Fatigue	Drug	A+
55842	Jhinkoo	53	Female	70	5.3	Breast cancer	Sore Nipple	Cockroach	B+

II. RELATED WORK

Latanya Sweeney [1] proposed *k*-anonymity using generalization and suppression. Generalization is the most common annonymization PPDM technique which replaces quasi-identifier values with less specific but semantically consistent values. Then all quasi-identifier values in a group are generalized to the entire group extend. For the generalization to be effective, records in the same bucket must be close to each other to avoid information loss during generalization. The limitations of generalization are:

1. It fails on high-dimensional data.
2. It causes too much information loss.

Bucketization is the process of defining the several records grouping based on their sensitive values. The sensitive values of the attributes are identified and sorted based on the frequencies in ascending order. After sorting is done, the contiguous sensitive values are grouped into the similar bucket [2]. Bucketization is used mainly on high-dimensional data. The anonymized dataset in bucketization consists of set of buckets with permuted sensitive attributes.

The limitations of bucketization are:

1. It does not prevent membership disclosure
2. It requires clear separation between quasi identifier and sensitive attributes and also breaks the co-relation between them.

Another technique introduced is slicing which overcomes the disadvantages of generalization and bucketization. Slicing divides the data both horizontally and vertically. It preserves better data utility than generalization and can be used for membership disclosure protection. Another important advantage of slicing is that it can handle high-dimensional data [3].

The limitations of Slicing are:

1. It does not prevent attribute co-relation.
2. Data utility is lost.

An extension of slicing has been introduced next, known as overlap slicing. Overlap slicing duplicates attribute in more than one column and this releases more attribute correlations. Hence increases privacy and utility of data, by achieving correlation among attributes. Overlap slicing works in three main steps [4]:

1. Attribute partitioning: In attribute partitioning, correlations of the attribute are measured to form there group. To measure the correlation mean square contingency coefficient is used
2. Tuple partitioning: In this step tuples are grouped to form bucket. Mondrian algorithm is used for tuple partitioning.
3. Column Generalization: Column generalization is required for preserving membership disclosure. It would be useful to apply column generalization to ensure that each column value appears with at least some frequency.

The limitation of overlap slicing is:

1. It does not prevent identity disclosure.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2016

Table 2: Overlap Slicing

AGE_PATIENTID_DIEASES	NAME_WEIGHT_HEIGHT_BLOODGROUP	GENDER_SYMPTOMS_ALLERGY
[55,149190, Breast cancer]	[Nataraj, 80 , 5.2, O-]	[Male, Cough, Dust]
[45, 64410, Asthma]	[Gopal, 70 , 5.4, AB+]	[Female, Red Rash, Drug]
[53, 35754, Candidiasis]	[George, 90, 5.1, A+]	[Male, Sore Nipple, Pet]
[68, 16680, Calculi]	[Sudhakar, 88, 5.7, B+]	[Male, Weight loss, Food]
[60, 55842, Cancer]	[Satish, 85, 5.5, A+]	[Male, Nausea, Skin]
[70, 2278392, Cancer]	[Jhinkoo, 81, 5.3, B+]	[Male, Fatigue, Cockroach]

III. PROPOSED METHOD

Enhanced Slicing

In this paper a robust and effective technique has been introduced known as enhanced slicing. Enhanced slicing ensures that the attacker cannot learn the sensitive value and the identity of the person at any cost and also preserves privacy. Enhanced Slicing consists of mainly three steps:

1. Attribute partitioning into columns
2. Tuple partitioning into buckets
3. Enhanced slicing

The first two steps are similar to that of overlap slicing. In the last step i.e. enhanced slicing, highly co-related attributes are grouped in the same column to preserve the attribute correlations. Vertical partitioning requires grouping of co-related attributes. Next there comes horizontal partitioning which is done by grouping the tuples into buckets. After the two partitioning steps, the quasi-identifier is represented in a generalised form to prevent the identification of individual records in the data. The quasi identifier is this table is the age. The next and the final step in enhanced slicing is performing operation on the sensitive attribute in such a way that the identity of the person cannot be revealed at any cost. Here the most sensitive attribute in the table is name of the patient. Hence in the final step, shuffle operation is performed i.e. the letters of the name has been disarranged to prevent identity disclosure. Enhanced-slicing is that it can manage data with greater dimension and can completely stop membership exposure and also identity disclosure.

Table 3: Enhanced Slicing

AGE_PATIENTID_DIEASES	NAME_WEIGHT_HEIGHT_BLOODGROUP	GENDER_SYMPTOMS_ALLERGY
[[50-60],149190, Breast cancer]	[aajNatr, 80 , 5.2, O-]	[Male, Cough, Dust]
[[40-50], 64410, Asthma]	[opGal, 70 , 5.4, AB+]	[Female, Red Rash, Drug]
[[50-60], 35754, Candidiasis]	[georGe, 90, 5.1, A+]	[Male, Sore Nipple, Pet]
[[60-70], 16680, Calculi]	[uhSakdar, 88, 5.7, B+]	[Male, Weight loss, Food]
[[60-70], 55842, Cancer]	[Shisat, 85, 5.5, A+]	[Male, Nausea, Skin]
[[60-70], 2278392, Cancer]	[hJikono, 81, 5.3, B+]	[Male, Fatigue, Cockroach]

IV. STEPS TO PERFORM ENHANCED SLICING

Step 1: Upload the dataset.
Step 2: Find the sensitive attribute.
Step 3: Calculate the correlation coefficient of the attributes.
Step 4: Perform attribute partitioning.
Step 5: Perform tuple partitioning.
Step 6: Randomize the tuple in each buckets.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2016

Step 7: Perform generalization on quasi identifier and shuffle operation on sensitive attribute.
Step 8: End.

V. RESULTS AND DISCUSSION

We have conducted two experiments. In the first experiment we evaluate the performance and accuracy of our system with the existing system based on the cardinality (number of records) of the data. We evaluate the quality of the annonymized data which is computed by our system, using the privacy technique namely overlap slicing and enhanced slicing. We use Weka software package to evaluate the Classification Accuracy (CA) and Root Mean Squared Error (RMSE). Classification Accuracy helps us to compute the performance by asking the classifier to give its best guess about the classification for each instance in the test set [5]. It is represented in the percentage format in our experiment. Root Mean Squared Error is the measure of accuracy. It is the quadratic scoring rule which measure the average magnitude of error. The square root of average is taken. Since the error is squared before they are averaged, the RMSE relatively give high weight to large errors. Hence the lower the RMSE value, better the accuracy [6].
For computing the CA and RMSE we have taken the cardinality of 50 records, 100 records, 200 records, 300 records and 600 records respectively. We have generated a graph comparing the CA and RMSE values of both the systems as shown in Fig. 1,2,3,4 and 5. This experiment demonstrates that:

1. Overlap slicing has a better performance measurement than Enhanced slicing.
2. Overlap slicing has better accuracy rate than Enhanced slicing.

In the second experiment we calculated and compared the Execution Time in milliseconds through our system and generated a graph based on the cardinality of 50 records, 100 records and 300 records as shown in Fig. 6, 7 and 8. This experiment demonstrates that:

1. Enhanced slicing has less execution time compared to Overlap slicing when the cardinality=50 records.
2. Enhanced slicing has less execution time compared to Overlap slicing when the cardinality=100 records.
3. Overlap slicing has less execution time compared to Enhanced slicing when the cardinality=300 records.

Experimental Data

We used the real-time dataset which was generated manually by us. The dataset is described in Table 1. The dataset consist of 10 attributes in total. In our experiment we have performed 3-column slicing on OCC-10. The 3 columns are:

1. {AGE_PATIENTID_DISEASE}
2. {NAME_WEIGHT_HEIGHT_BLOODGROUP}
3. {GENDER_SYMPTOMS_ALLERGY}

We are using Weka software package to compute CA and RMSE using Decision tree C4.5 (J48). Default setting is used. We use training set for the experiment. In our experiment we use one attribute as target attribute. It is the attribute on which the classifier is build. We build the classifier on the sensitive attribute {NAME_WEIGHT_HEIGHT_BLOODGROUP}

Fig.1.Cardinality = 50 Records

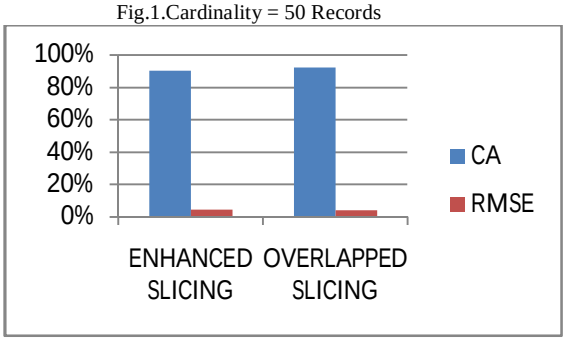
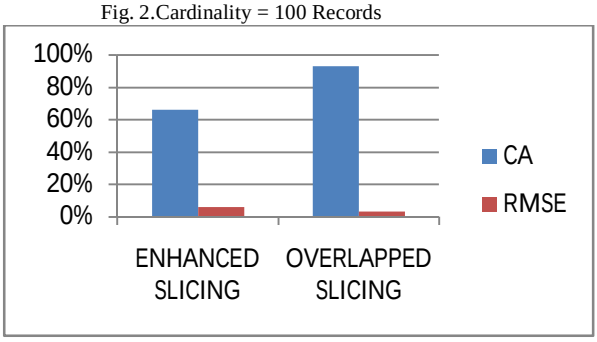


Fig. 2.Cardinality = 100 Records



ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2016

Fig. 3.Cardinality = 200 Records

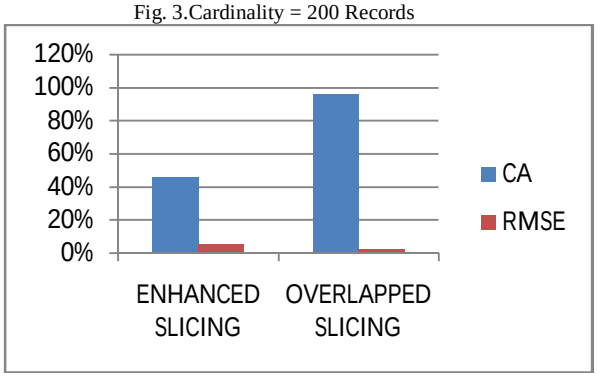


Fig 4.Cardinality = 300 Records

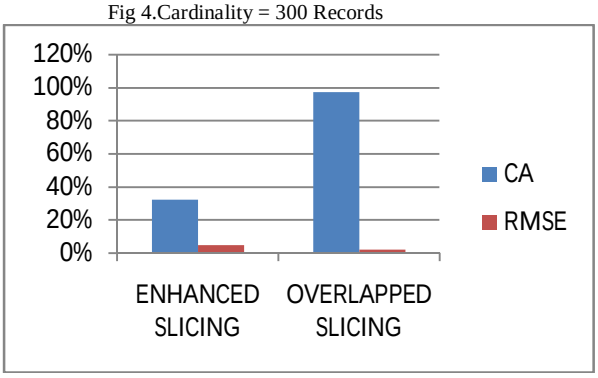


Fig 5.Cardinality = 600 Records

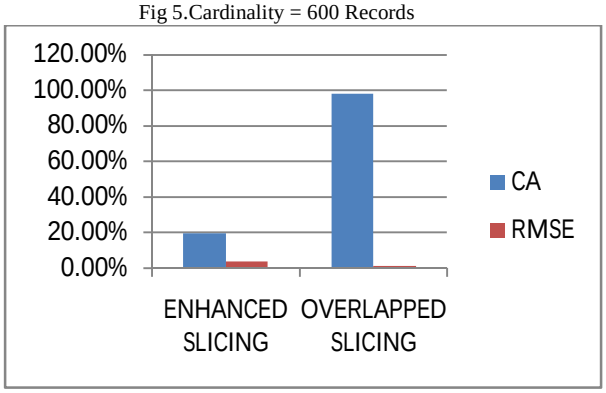


Fig 6.Execution Time [Cardinality = 50 Records]

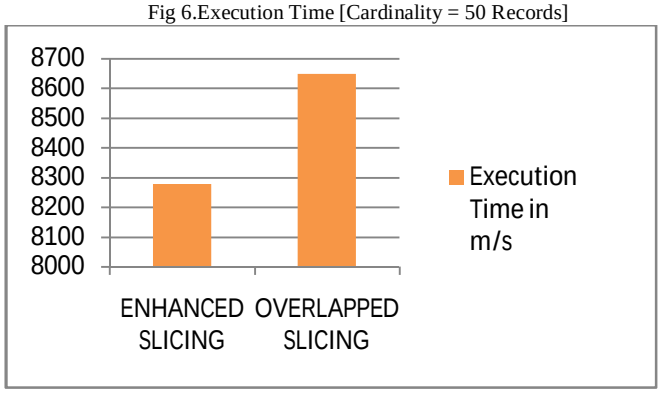


Fig 7.Execution time [Cardinality = 100 Records]

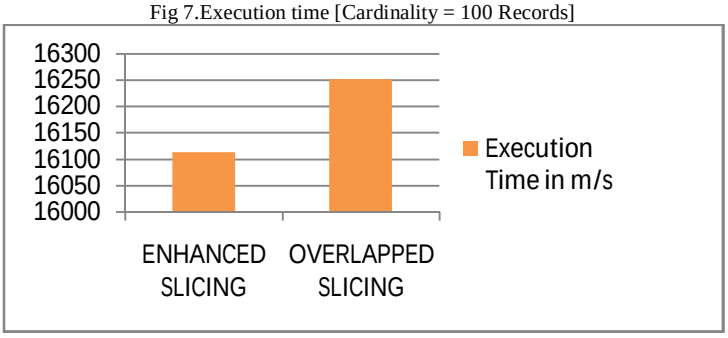
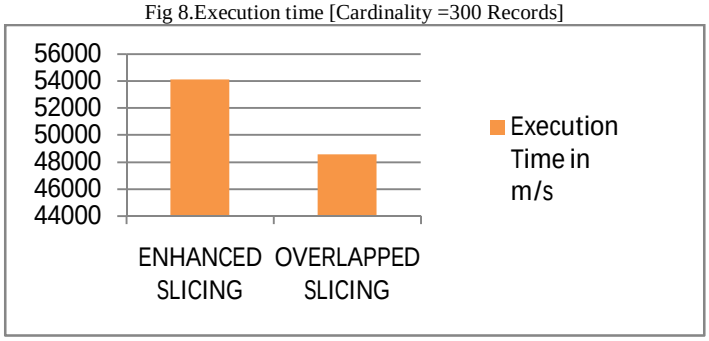


Fig 8.Execution time [Cardinality =300 Records]



ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2016

VI. CONCLUSION AND FUTURE WORK

Enhanced slicing overcomes the limitation of the existing technique i.e. Overlap slicing as by protecting identity disclosure. But Overlap slicing has a better performance measurement and better accuracy rate compared to Enhanced slicing. Enhanced slicing has less execution time compared to Overlap slicing on cardinality=50 records and 100 records. Overlap slicing has less execution time compared to Enhanced slicing on cardinality=600 records
The aim of designing an effective privacy algorithm with better performance and better accuracy rate is left to the future work.

Table 4: Conclusion Table

PARAMETERS	ENHANCED SLICING	OVERLAP SLICING
Classification Accuracy	Less performance measurement	Better performance measurement
RMSE	Less accuracy rate	Better accuracy rate
Execution Time	Takes less execution time	Takes more execution time
Identity Disclosure	Doesn't protect identity disclosure	Protects identity disclosure

REFERENCES

1. Latanya Sweeney, "Achieving k-anonymity privacy protection using generalization and suppression", 2002, volume 10, issue 5.
2. G.Kesavaraj and S.Sukumaran, "Bucketization based Flow Classification Algorithm for Data Stream Privacy Mining", IJCA, Volume 81, Issue 12, 2013.
3. Dr. S. Govinda Rao , D. Siva Prasad and M. Eswara Rao, "A New Approach Slicing for Micro Data Publishing", IJCSIT, Volume 4, Issue 5, 2013,
4. Suman S. Giri and Mr.Nilav Mukhopadhyay, "Overlapping Slicing with New Privacy Model", International Journal of Scientific and Research Publications, Volume 4, Issue 6, 2014.
5. http://stackoverflow.com/questions/12252254/meaning-of-correctly-classified-instances-weka.
6. http://www.eumetcal.org/resources/ukmeteocal/verification/www/english/msg/ver_cont_var/uos3/uos3_ko1.html.