Privacy Preserving of Data Mining Based on Enumeration and Concatenation of Attributes using K-Anonymity

Rajesh.N*
Department of MCA, Sir MVIT, Bangalore.
&School of Computer Science and Engg.,
Bharathiyar University,India
rajkansrk@gmail.com

Dr. A. Arul Lawrence Selvakumar
Department of Computer Science & Engineering
Rajiv Gandhi Institute of Technology
Bangalore, India.
dr.arullawrence.cse@gmail.com

***Abstract:* Privacy preserving data mining and security of data has been developed for privacy protection and protection of data during the data mining process and discovery of knowledge. The individual personal data can be used for data mining process without disclosing the individual identity or any information regarding the individual identity should be removed. Privacy preserving data mining is concerned with the valid data mining results. For example, data mining applications like banking, credit card, hospitals which is concerned with the individual personal sensitive information for those applications that is concerned with personal data, the individual sensitive data can be removed from the database or the personal information can be encrypted and stored at different locations in order to protect the privacy of the individuals. This paper includes a formal protection model of enumeration and concatenating attribute using k-anonymity. The methods used in this paper1.Enumeration and concatenation of age, sex attribute 2. Concatenating and encoding city and zipcode attribute.**

Keywords: Enumeration, Anonymity, Concatenation, Encryption

I. INTRODUCTION

Data mining is one of the most growing research areas these days with the huge information available from the web and the powerful data mining tools that have been developed and being used. There is increasing need for data mining can pose a threat to the privacy and security of the individual data and there is a need for protecting the sensitive information of the individuals. For example, in a hospital when the patients diagnosis reports are being released, so that the researchers can study the characteristics of the various disorders. The original data contain the identities of the individuals, which must not be released in order to protect the privacy of sensitive information [8]. There may be other attributes that can be consolidated with other databases in order to recover the personal information of the individuals. Consider, that the hospital publishes the data in **Table 1** the micro data in which the names of the patients are not released. But if an opponent has access to the voter database[1] in **Table 2** he can easily diagnosis the identity of all patents by combining the two tables on the attributes are age, city, sex and zipcode. There is a ways a risk of publishing the original data. Therefore the personal information needs to be encrypted and stored at different location [5] in a database by protecting the sensitive of the data and by decrypting the data we can get the original data by combining with a voter database [10].

Even if some of attributes have been removed, the opponent can use some of the cryptographic techniques. In order to overcome the [4] re-identification, the mechanism of formal protection model of enumeration and concatenating attributes using k-anonymity [7] [3] was proposed. There are numerous papers which describes the k-anonymity models. Privacy preserving data mining can be described in the form of a data can be segmented vertically or horizontally [9]. Privacy preserving of data mining based on enumeration and concatenation of attributes using k-anonymity which allows the consolidation of various databases and sensitive information is encrypted and stored at different location.

II.PROBLEM DEFINITIONS

We consider **Table 1** as the beginning of micro data table and **Table 5** is the end of released micro data table. Both tables are having set of tuples over an attribute set. Attributes are classified into following categories [3].

Identifier attributes: Identifier attributes used to identify the complete record of particular person name or entity name.

TABLE 1. MICRODATA TABLE

Rid	Age	Sex	City	Zipcode	Disease
1	36	M	Bangalore	560056	HIV
2	74	F	Hyderabad	500063	CANCER
3	42	M	Cochin	682023	HIV
4	32	M	Chennai	600012	FLU

TABLE 2. VOTER DATABASE

Rid	Name	Age	Sex	City	Zipcode	Disease
1	Anand	36	M	Bangalore	560056	HIV
2	Angelin Sofi	74	F	Hyderabad	500063	CANCER
3	Robert	42	M	Cochin	682023	HIV
4	Rakesh	32	M	Chennai	600012	FLU

TABLE 3. ANONYMIZATION OF AGE ATTRIBUTE

Rid	Age	Sex	City	Zipcode	Disease
1	[20-39]	M	Bangalore	560056	HIV
2	[60-89]	F	Hyderabad	500063	CANCER
3	[40-59]	M	Cochin	682023	HIV
4	[20-39]	M	Chennai	600012	FLU

Quasi-identifier attributes: Quasi-identifier attributes may be known by violator such attributes are age, sex, city and zipcode. This QI attributes are presented in to the all tables and using this attributes can identify the identifier attribute.

Sensitive attributes: Sensitive attributes may not know to the violator and need to be securing this attribute such as disease of all patients. This attribute is presented to the all tables.
In this problem we consider only quasi-identifier attribute and sensitive attributes are used in all tables and we have been removed identifier attributes for secure the patient name. Sometimes violator may use record linkage techniques [11] to extracting form other source and available data can find the individual name of particular patient. To avoid this problem we introduce the method called enumeration and concatenation of attributes using k-Anonymity.

K-anonymity: K-anonymity [2] is an approach where we combine various database and limits access to personal information to those who requires knowing and the personal information can be encrypted and stored at various databases at different locations. In order to protect the sensitive information and to maintain the privacy of data and we can decrypt the information when there is a need. Consider **Table 5**, where data has been encrypted and combining external data source it's very difficult to identify the identifier attribute.

III.CONCATENATION OF ATTRIBUTES USING K-ANONYMITY

Consider the micro data **Table 1** in the QI Attributes are {age, sex, city, zipcode}, use to identify the Identifier attribute (patient name) in **Table 2**. To avoid identifying the person name using QI attributes are introducing the enumeration and concatenation of attributes using k-anonymity method.

Figure 1 age is domain attribute that classified in four categories Age = {child teenage, youth, middle age, senior} each categories assigned range of interval values are child teenage[1-19], youth[20-39], middle age[40-59], senior[60-99] in **Table 3**. Some time intruder using this interval values of age attribute can identify the person categories like child to senior, example {Select count(*) from microdata age between 20 and 39;} after executing the query, 2 tuples are selected and category of those tuple is youth.

Here introducing method called enumeration to set value of each categories [6] of age {child = 1, youth = 2, middle age =3, senior = 4} in **Table 4**. The idea of enumeration method using next QI is sex, sex is discrete domain attribute here assume only two values M/F and set the values of sex {F = 1, M = 2} in **Table 4**.

TABLE 4. ENUMERATION OF AGE, SEX ATTRIBUTE

Rid	Age	Sex	City	Disease
1	2	2	Bangalore-56	HIV
2	4	1	Hyderabad-63	CANCER
3	3	2	Cochin-23	HIV
4	2	2	Chennai-12	FLU

TABLE 5. CONCATENATION OF ATTRIBUTE USING ANONYMITY

Rid	Age	City	Disease
1	22	Bangalore-06	HIV
2	41	Hyderabad-13	CANCER
3	32	Cochin-73	HIV
4	22	Chennai-62	FLU

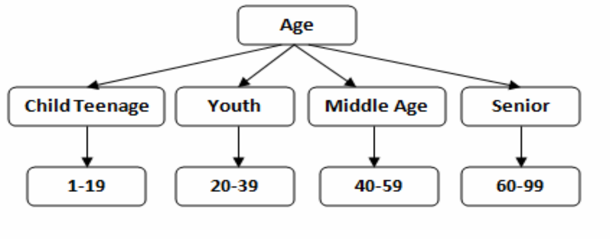

Figure 1. (Classification of Age)

In releasing **Table 5** age attribute concatenate the two attributes are A1 and A2 from **Table 4** will get duplicate value of age attribute for each person, example {select age||sex as Age into **Table 5** from **Table 4**;}, the value of age similarly to original and giving only child to middle age category level for after concatenation of two attributes and will not be get age category of senior level after encrypted of age attribute, we can reconstruct the original data for age category and sex using attribute A1 in **Table 5**. Another concatenate quasi identifier attributes are city (A3) and zipcode (A4) in **Table 4**. In city contains many areas each area contains one unique zipcode, example Rid1 in microtable city is Bangalore and zipcode is 560056, this code referring area is Viswavidalaya. Bangalore city all area zipcodes first four common number is 5600, last two digit is represent unique code for each area in Bangalore, When combining two attribute it gives Bangalore-56 refers same Viswavidalaya place. Example {select zipcode mod 100 from **Table 3**} will get last two digit of zipcode from **Table 3** and joining with city attribute from **Table 3**, example {select city+'-'+zipcode as city into **Table 4** from **Table 3**;} it gives same area code for the particular city. Before concatenate the attribute to show duplicate value of zipcode using private data encrypting method, example {update **Table 5** set zipcode = case when zipcode between 00 and 49 then zipcode+50 else when zipcode between 50 and 99 then zipcode-50 end ;} in **Table 5**.

IV. EXPERIMENT AND RESULT

In **Figure 2** and **Figure 3** we assume that Y-axis range values from 00 to 99, those values consider for two attributes (age, zipcode), here age range is 00 to 99 and zipcode of last two digits value also 00 to 99. **Figure 2** Represent the original age value and last two digit of zipcode for all four records from **Table 1**. In experiment taken four records data for five attributes, here concatenate attribute1 is age and sex, attribute2 is city and zipcode in the result of releasing table. The real age for all four records in mirodata (36,74,42,32) and category for those age is(youth, senior, middle-age, youth).The age of all records after the encrypted (22,41,32,22) and category for those age are (youth, middle-age, youth, youth), here first and fourth record category is youth, its original data category value in microdata and other two records showing(**Figure 3**) different category, in this experiment 50% age category is secured. The next concatenate attributes are city and zipcode, before encoding the values for city in **Table 4** (Bangalore-56, Hyderabad-63, Cochin-23, and Chennai-12) this data refers current areas are (Viswavidalaya, Adarshnagar, Vaduthala, Perambur).

After the encoding values for city, in **Table 5** (Bangalore-06, Hyderabad-13, Cochin-73, and Chennai-62) this data showing some other places (J C Nagar, Amberpet, Puthencruz, and Sathyamurthinagar). In releasing table its look like original data, intruder get confusion to identify the person name in particular place using voter database.

1. ORIGINAL DATA

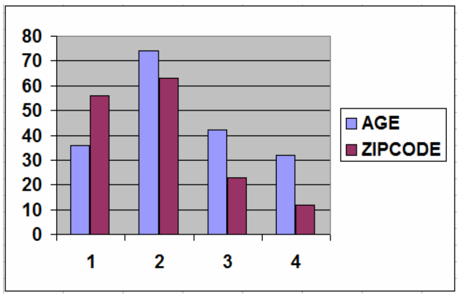

Figure 2

2. ENCRYPTED DATA

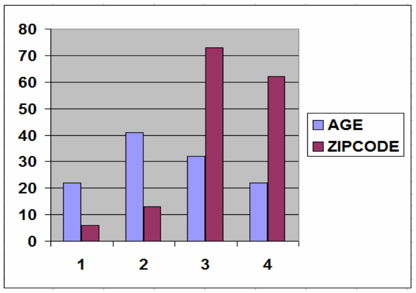

Figure 3

V. CONCLUSION

Privacy becomes an important factor in data mining, so here a new method called enumeration and concatenation of attributes using K-anonymity is proposed to secure individual data. Group based anonymization technique called suppression used to remove complete quasi identifier attributes to secure the sensitive data. Most of the previous research for providing the privacy data they are not used combined attributes techniques. In this method, original data is obtained by combining two attributes. Which we can reconstruct the original data and after combining attributes. In this method if using limited records data, intruder cannot know encryption method to reconstruct the original. In the future research work there is a plan to propose PPDM to combine more than two attributes to show one quasi-identifier like original attribute to protect the personal data and using unlimited records.

REFERENCES

[1].L.Sweeney.K-Anonymity: “A model for Protecting Privacy”. International Journal on Uncertainty Fuzziness Knowledge based System, 10(5), pp 557-570, 2002.

[2].Nivetha.P R and Thamarai selvi.K, “A Survey on Privacy Preserving Data Mining Techniques”, IJCSMC, vol.2, Issue 10 October2013, pg.166-170.

[3].Xiaoxun Sun and Hua Wang, “Achiving P-sensitive K-Anonymity via Anatomy”, 2009 IEEE International Conference on e-Business Engineering.

[4].Pingshui WANG, “Personalized Anonymity Algorithm using clustering Techniques” Journal of Computational Information System 7:3(2011) 924-931.

[5].Xiaoxun sun, Hua Wang, Jiuyong Li and Traian merius truta, “Enhanced P-sensitive K-Anonymity models for privacy preserving data publishing” transaction on Data Privacy1(2008) 53-66.

[6]. Benjamin C.M. Fung, Ke Wang, and Philip S. Yu,”Anonymizing Classification Data for Privacy Preservation” IEEE transactions on knowledge and data engineering, vol. 19, no. 5, may 2007.

[7].Bo peng, Xingyu geng and Jun Zhang, “Combined data distortion strategies for privacy-prserving data mining”. 2010 3rd International conference on Advanced computer theory and engineering (ICACTE).

[8].W.Du, Y Han, and S.Chen, “Privacy-preserving multivariate statistical analysis: Linear regression and classification”, In proceedings of the fourth SIAM International conference on Data mining, 2004, pp 222-233.

[9].K.Chen and L.Liu, “Privacy preserving data Classification with rotation perturbation”. In the fifth International conference of Data mining (ICDM’05), 2005, pp 589-592.

[10].K.Liu, H.Kargupta, and J.Ryan, “Random Projection-Based Multiplication Data Perturbation for privacy preserving Distributed Data mining”, IEEE transaction on knowledge and Data Engineering (TKDE), January 2006, pp 92-106.

[11]. Li Liu, Murat Kantarcioglu and Bhavani Thuraisingham “The Applicability of the Perturbation Model-based Privacy Preserving Data Mining for Real-world Data”, Sixth IEEE International Conference on Data Mining - Workshops (ICDMW'06) 0-7695-2702-7/06.