Volume 4, Issue 1, January 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

A Robust Data Preserving Technique by K-Anonymity and hiding Association Rules

Arvind Batham*
Department of Information Technology
RKDF Institute of Science & Technology
India

Mr.Srikant Lade
HOD of Information Technology
RKDF Institute of Science & Technology
India

Mr. Deepak Patel
Department of Computer Science
Millennium Institute of Technology
India

Abstract— With the development of data mining technology, an increasing number of data can be mined out to reveal some potential information about user. While this will lead to a severe problem, which is users' privacy may be violated easily. The goal of privacy preserving is to mine the potential valuable knowledge without leakage of sensitive records, in other words, use non-sensitive data to infer sensitive data. This work analyze and optimize the technologies of privacy preserving data mining. By the implementation of K-Anonymity and Appear an efficient algorithm has develop which has the property of to extract relevant knowledge from large amount of data and at the same time protect the sensitive information from the data miners. A result shows that without affecting the accuracy of the dataset it successfully preserved sensitive data.

KeyTerms — Data mining, Data Perturbation, Multiparty Privacy Preserving.

I. INTRODUCTION

In recent years, researchers have proposed to publish data in the form of micro data, i.e., data in the original form of individual tipples. Obviously the release of micro data offers significant advantages in terms of information availability, as the original records are kept and people can issue arbitrary queries they are interested in. So it is particularly suitable for ad hoc analysis. However, the release of micro data raises privacy concerns when records containing sensitive attributes (SA) of individuals are published. Existing privacy practice relies on deidentification, i.e., removing explicit identification information (e.g., name, SSN, home address and telephone numbers) from microdata. However, it has been well recognized [2, 3] that simple de-identification is not sufficient to protect an individual's privacy. One's other attributes (so-called quasi-identifiers, or QI for short, such as age, zip code, date of birth and race) are usually needed for data analysis, and thus are kept after de-identification. Individuals' sensitive information may often be revealed when microdata are linked with publicly available information through quasi-identifiers. A famous example is given by Sweeney in [3], where she successfully identified the governor of Massachusetts using only his date of birth, gender, and ZIP code from local hospital records, and then combine this information with the census database.

However, depending on the nature of the sensitive attributes, even these enhanced properties still permit the information to be disclosed or have other limitations. Most of the existing work places more stress on the protection of the specific values, not the sensitive categories that the specific value belongs to. For example, the information of a person who is affected by a Top Confidential disease needs to be protected, no matter whether it is HIV or Cancer.

It will be very useful to propose a privacy model that ensures the protection of not only the specific values, but also the confidential categories they belong to. In the scenarios, the same database is requested for different application purposes by different data requesters. On the one hand, considering the diversity of purposes, the requirements for individual attributes, based on how important they are for requesting purposes, are various. For example, Age and Gender attributes in the census database are essential for demographic purposes, but they are not necessary for some prediction purposes, so a priority weight associated with each attribute is valuable to indicate the importance of the attribute for requesting purposes. While, on the other hand, considering the variety of data requesters, the reliability of data requesters to data providers depends on their trust evaluation. The trust between the data requester and data provider reflects the possibility that the data would be misused by the data requester. The more trustworthy the data requesters are, the less chance they will maliciously use the requested data. Existing work on data anonymisation focuses on developing effective models and efficient algorithms to optimize the trade-off between data privacy and utility. Normally, the same anonymous data are delivered to different requesters regardless of what kind of purposes the data are used for, letting alone the reliability of the data requester. By specifying the requesters' application purpose and their reliability, the result of the data anonymisation will achieve a better trade-off. Recently, a new privacy concern has emerged in privacy preservation research: how to protect individuals' privacy in large survey rating data. For example, movie rating data, which is supposed to be anonymized, is de-identified by linking un-anonymized data from another source [11]. Though several models and algorithms have been proposed to preserve privacy in relational data, most of the existing studies can deal with relational data only.

II. RELATED WORK

The IBM Multinational Consumer Privacy Survey performed in 1999 in Germany, USA and UK illustrates public concern about privacy [2]. Most respondents (80%) feel that \consumers have lost all control over how personal information is collected and used by companies". The majority of respondents (94%) are concerned about the possible misuse of their personal information. This survey also shows that, when it comes to the confidence that their personal information is properly handled, consumers have most trust in health care providers and banks and the least trust in credit card agencies and internet companies.

Privacy-preserving clustering has been previously addressed by Oliviera and Zaıane [5], Vaidya and Clifton [8],. Oliviera and Za¨iane's work [5] use s data transformation in conjunction with partition-based and hierarchical clustering algorithms, while the others use cryptographic techniques to give privacy-preserving versions of the k-means clustering algorithm. Vaidya and Clifton's result [10] addresses privacy-preserving k-means clustering for vertically partitioned data, Jha, Kruger, and McDaniel's [7] addresses horizontally partitioned data, and Bunn and Ostrovsky [6] address arbitrarily-partitioned data.

Tzung Pei et al presented Evolutionary privacy preserving in data mining [9]. Collection of data, dissemination and mining from large datasets introduced threats to the privacy of the data. Some sensitive or private information about the individuals and businesses or organizations had to be masked before it is disclosed to users of data mining. An evolutionary privacy preserving data mining method was proposed to find about what transactions were to be hidden from a database. Based on the preference and sensitivity of the individual's data in the database different weights were assigned to the attributes of the individuals. The concept of prelarge item sets was used to minimize the cost of rescanning the entire database and speed up the evaluation process of chromosomes. The proposed approach was used to make a good tradeoff between privacy preserving and running time of the data mining algorithms.

Han and Keong Ng presented Privacy Preserving Genetic Algorithms for Rule Discovery [4]. Entire data set was partitioned between two parties, and genetic algorithm was used to find the best set of rules without publishing their actual private data. Two parties jointly developed fitness function to evaluate the results using each party's private data but not compromising the privacy of the data by Secure Fitness Evaluation Protocol. To meet the privacy related challenges, results generated by genetic algorithm were not compromising privacy of those two parities having partitioned data. Creation of initial population and ranking the individuals for reproduction were done jointly by both parties. In 2004, the Office of the Federal Privacy Commissioner, Australia, engaged Roy Morgan Research to investigate community attitude towards privacy [2]. According to the

survey, 81% of the respondents believe that \customer details held by commercial organizations are often transferred or sold in mailing lists to other businesses".

III. BACKGROUND

K- Anonymity: In the k-anonymity model, the quasi-identifier feature set consists of features in a table that potentially reveals private information, possibly by joining with other tables. In addition, the sensitive feature is a feature serves as the class label of each record. As shown in table. 1(b), the set of three features {Zip, Gender, Age} is the quasi-identifier feature set, while the feature {Diagnosis} is the sensitive feature. For each record in this table, its feature values in the quasi-identifier feature set are generalized as capsule feature values, while its value of sensitive feature are not generalized. Through generalization, an equivalence class is the set composed of records in the table which has the same values on all features in the quasi-identifier feature set. The 1st, 3rd and 4th records in table. 1(b) are assembled to form one equivalence class, while the 2nd, 5th and 6th records are assembled to form another equivalence class. The number of records in each equivalence class must be not less than k, which is called as the k-anonymity requirement. The value of k is specified by users according to the purpose of their applications. The records in table. 1(b) satisfy 3-anonymity requirement since the numbers of records in its two equivalence classes are both equal to three.

Table. 1. Patient diagnosis records in a hospital

Zip	Gender	Age	Diagnosis
47918	Male	35	Cancer
47906	Male	33	HIV+
47918	Male	36	Flu
47916	Female	39	Obesity
47907	Male	33	Cancer
47906	Female	33	Flu

Table 2. The k-anonymity protected table when k= 3.

Zip	Gender	Age	Diagnosis
4791*	Person	[35-39]	Cancer
4790*	Person	[30-34]	HIV+
4791*	Person	[35-39]	Flu
4791*	Person	[35-39]	Obesity
4790*	Person	[30-34]	Cancer
4790*	Person	[30-34]	Flu

Association Rule Mining

Association Mining is one of the most important data mining's functionalities and it is the most popular technique has been studied by researchers. Extracting association rules is the core of data mining [13]. It is mining for association rules in database of sales transactions between items which is important field of the research in dataset [12]. The benefits of these rules are detecting unknown relationships, producing results which can perform basis for decision making and prediction [14]. The discovery of association rules is divided into two phases [14, 11]: detection the frequent itemsets and generation of association rules.

There are two important basic measures for association rules, Support (s) and confidence (c). Since the database is large and users concern about only those frequently purchased items, usually thresholds of support and confidence are pre defined by users to drop those rules that are not so interesting or useful. The two thresholds are called minimal support and minimal confidence respectively, additional constraints of interesting rules also can be specified by the users. The two basic parameters of Association Rule Mining (ARM) are: support and confidence. Support(s) of an association rule is defined as the percentage/fraction of records that contain X U Y to the total number of records in the database. The count for each item is increased by one every time the item is encountered in different transaction T in database D during the scanning process. It means the support count does not take the quantity of the item into account. For example in a transaction a customer buys three bottles of beers but we only increase the support count number of {beer} by one, in another word if a transaction contains a item then the support count of this item is increased by one. Support(s) is calculated by the following

$$\text{Support } (X \rightarrow Y) = (XUY) / D$$

Confidence: Confidence of an association rule is defined as the percentage/fraction of the number of transactions that contain X U Y to the total number of records that contain X, where if the percentage exceeds the threshold of confidence an interesting association rule $X \rightarrow Y$ can be generated.

$$\text{Confidence } (X \rightarrow Y) = (XUY) / X$$

confidence is a measure of strength of the association rules, suppose the confidence of the association rule $X \rightarrow Y$ is 80%, it means that 80% of the transactions that contain X also contain Y together, similarly to ensure the interestingness of the rules specified minimum confidence is also pre-defined by users.

Aprior Algorithm: Let D the task relevant data, be a set of database transactions where each transaction T is a set of items, called Tid. Let I= {I1, I2,..., Im} be a set of items. An item set contains k items is a k item set. If a k item set satisfies

minimum support (Min_sup) then it is a frequent k item set, denoted by Lk. Firstly Apriori algorithm generated a set of candidates, which is candidate k-item sets, denoted by Ck. If the candidate item sets satisfies minimum support then it is frequent item sets. The description of the algorithm is given below:

1. Suppose a minimum support threshold Min_sup) and a minimum confidence threshold (Min_conf)[8]
2. Scan the dataset, candidate 1-itemsets, C1, and the number of occurrences of each item is determined. The set of frequent 1-itemsets, L1, is then determined, consisting of those candidate 1-itemsets in C1 having minimum support. The algorithm uses $L1 \infty L2$ to generate candidate 2-itemsets, C2.

3. Scan the dataset again, frequent 2-itemsets, L2, is then determined, consisting of those candidate 2-itemsets in C2 having minimum support. Candidate 3-itemsets, C3 is then generated by $L2 \infty L2$.

4. Repeatedly scan the dataset, compare the support count of each candidate in Ck-1 with Min_sup, and then generate Lk-1 , join $Lk-1 \infty Lk-1$ to generate Ck until no more candidate item sets

Hiding Association Rule

Association rule hiding algorithms can be divided into three distinct approaches. They are heuristic approaches, border-revision approaches and exact approaches.

A. Heuristic Approach

Heuristic approaches can be further categorized into distortion based schemes and blocking based schemes. To hide sensitive item sets, distortion based scheme changes certain items in selected transactions from present to absent and vice versa. Blocking based scheme replaces certain items in selected transactions with unknowns. These approaches have been getting focus of attention for majority of the researchers due to their efficiency, scalability and quick responses.

B.Border Revision Approach

Border revision approach modifies borders in the lattice of the frequent and infrequent item sets to hide sensitive association rules. This approach tracks the border of the non sensitive frequent item sets and greedily applies data modification that may have minimal impact on the quality to accommodate the hiding sensitive rules. Researchers proposed many border revision approach algorithms such as BBA (Border Based Approach), Max–Min1 and Max-Min2 to hide sensitive association rules. The algorithms uses different techniques such as deleting specific sensitive items and also attempt to minimize the number of non sensitive item sets that may be lost while sanitization is performed over the original database in order to protect sensitive rules.

C. Exact Approach

Third class of approach is non heuristic algorithm called exact, which conceive hiding process as constraint satisfaction problem. These problems are solved by integer programming. This approach can be concerned as descendant of border based methodology.

IV. PROPOSED APPROACH

In this work privacy of the dataset is maintain by introducing K-Anonymity and reducing the frequent association rule from the dataset which contain sensitive item. So in order to hide an association rule, $X \rightarrow Y$, we can either decrease its support or its confidence to be smaller than user-specified minimum support transaction (MST) and minimum confidence transaction (MCT). To decrease the confidence of a rule, there is two approach:

(1) Increase the support of X, the left hand side of the rule, but not support of $X \rightarrow Y$.

(2) Decrease the support of the item set $X \rightarrow Y$ .For the second case, if we only decrease the support of Y, the right hand side of the rule, it would reduce the confidence faster than simply reducing the support of $X \rightarrow Y$.

Here it only reduce the RHS item of the rule correspondingly. For this algorithm t is a transaction, T is a set of transactions, R is used for rule, RHS (R) is Right Hand Side of rule R, LHS (R) is the left hand side of the rule R, Confidence (R) is the confidence of the rule R, a set of items H to be hidden.

K Anonymity Algorithm:
Input: Dataset D, K value {2, 3, …n}. Sensitive attribute A
OutPut: D with K-anonymity

1. Loop I = T is not empty
2. t ← T[I]
3. Loop J = I+1 is not empty
4. t' ← T[J]
5. If Equals (t[A], t'[A])
6. Count ← count +1
7. Mark(t')
8. EndIf
9. EndLoop
10. While count< K
11. T←Generate_transaction(S, t)
12. EndWhile
13. EndLoop

Hiding Rules:
Input: A source database D, A minimum support in_support (MST), a minimum confidence min_confidence (MCT), a set of hidden items X.

Output: The sanitized database D, where rules containing X on Left Hand Side (LHS) or Right Hand Side (RHS) will be hidden.

Steps of algorithm:
1. R[c,s] ← Aprior(D, X) // c= confidence & s = support

2. Loop I = For each rule R

3. If Intersect(R[I], H) and R[I] > MCT

4. New_transaction ← Find_transaction(R[I], MCT)

5. While (T is not empty OR count = New_treansaction)
6. If t←T have XUY rule then

7. Remove Y from this transaction

8. End While

9. EndIf

10. End Loop

V. EXPERIMENT AND RESULT

Dataset: In order to analyze proposed algorithm, it is in need of the dataset. One grocery shop dataset is use that has following attribute {items, date_of_birth, gender, salary}. Here personal information are from date_of_birth, gender, salary. While sensitive items are important for the Shop owner. So for the privacy preservation both things need to be hide. Evaluation Parameter: In order to compare our work one of the previous algorithm from [10] is utilize in which it increase the support of X and decrease the support of Y, for the confidence=(XUY)/X. In order to evaluate this work following are the few parameters of evaluation:

Lost Rules: Representing the number of non-sensitive patterns (i.e., association, classification rules) which are hidden as side-effect of the hiding process

False Rules: Representing the number of art factual patterns created by the adopted privacy preserving technique.

Perturbation Percentage: This specify the percentage of the dataset perturb by the adopting technique.

Results: This work use Matlab 2012a for the experiment. It was executed under an Intel Core 2 Duo 2.1 Ghz computer, using 3Gb RAM and Windows 7 Professional Edition. Here both the algorithm use same input parameter as minimum confidence, hide to be hide. But proposed algo take one more parameter for K-anonymity.

Table3. Represent Comparison of Lost rules at different confidence values

Confidence	Previous work	Proposed Work
	Lost Rules	
10	100	44.4
12	100	44.4
15	57.14	44.4
17	30.76	11.76
20	0	0
30	0	0

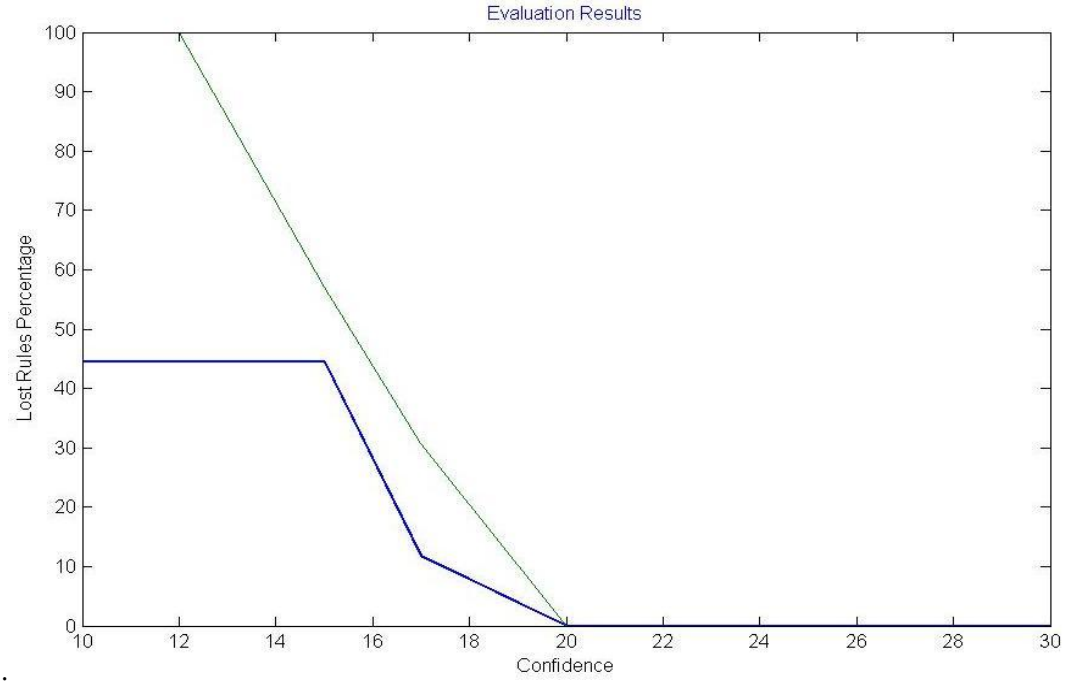

Fig. 1. Represent Graph of different Lost Rules vs Confidence. Dark blue is proposed work

It is observed from the above table 3 and fig 1 that with the decrease of the confidence value the lost rules percentage is increase in both the case. But in Proposed work it become constant when confidence is equal to 15 or less. As seen in the previous work the lost rules percentage in case after 12 is 100%. Means it will lost all rules of the dataset.

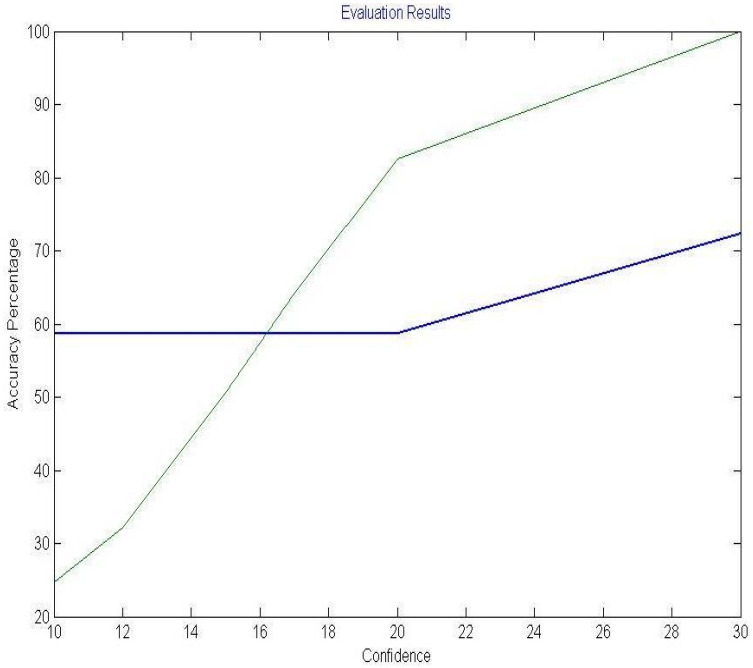

Fig.2. Represent Graph of different Accuracy percentage vs Confidence. Dark blue is proposed work

In Fig. 2. It is shown that Accuracy of the perturb dataset is almost constant for different confidence value case of proposed work. While the accuracy of the previous work get 0 percentage which is shows that dataset get completely corrupt and it can not contain any useful transaction. In both the algorithm False rules percentage is remain zero which means that both can successfully hide the item to be hide.

VI. CONCLUSION

In this paper, a set of algorithms and techniques were proposed to solve privacy-preserving data mining problems. The algorithms were implemented on Matlab. The experiments showed that the algorithms perform well on large databases. It work better as the Maximum lost rule percentage is constant below a certain value of confidence. Then this work shows that false rules value is zero. Comparison with the other algorithm it is better as include th K-Anonymity concet as well which directly hide the sensitive information.

REFERENCES

1. D. FRANKOWSKI, D. COSLEY, S. SEN, L. G. TERVEEN AND J. RIEDL. YOU ARE WHAT YOU SAY: PRIVACY RISKS OF PUBLIC MENTIONS. SIGIR 2006. PP, 565-572.
2. P. SAMARATI AND L. SWEENEY. PROTECTING PRIVACY WHEN DISCLOSING INFORMATION: KANONYMITY AND ITS ENFORCEMENT THROUGH GENERALIZATION AND SUPPRESSION. TECHNICAL REPORT SRI-CSL-98-04, SRI COMPUTER SCIENCE LABORATORY, 1998.
3. L. SWEENEY. K-ANONYMITY: A MODEL FOR PROTECTING PRIVACY. INTERNATIONAL JOURNAL ON UNCERTAINTY FUZZINESS KNOWLEDGE-BASED SYSTEMS, 10(5), PP 557-570, 2002.
4. SHUGUO HAN WEE KEONG NG, "PRIVACY -PRESERVING GENETIC ALGORITHMS FOR RULE DISCOVERY", 2007.
5. S. OLIVEIRA AND O. R. ZAÏANE. PRIVACY PRESERVING CLUSTERING BY DATA TRANSFORMATION. IN PROC. OF THE 18TH BRAZILIAN SYMPOSIUM ON DATABASES, PAGES 304–318, 2003.
6. A. C. YAO. HOW TO GENERATE AND EXCHANGE SECRETS. IN 27TH FOCS, PAGES 162–167, 1986.
7. S. JHA, L. KRUGER, AND P. MCDANIEL. PRIVACY PRESERVING CLUSTERING. IN ESORICS , PAGES 397–417, 2005
8. . J. VAIDYA AND C. CLIFTON. PRIVACY-PRESERVING K-MEANS CLUSTERING OVER VERTICALLY PARTITIONED DATA. IN 9TH KDD , 2003.
9. V. ESTIVILL-CASTRO AND L. BRANKOVIC. DATA SWAPPING: BALANCING PRIVACY AGAINST PRECISION IN MINING FOR LOGIC RULES. IN PROC. OF DATA WAREHOUSING AND KNOWLEDGE DISCOVERY (DAWAK99), 1999.
10. M.MAHENDRAN, 2DR.R.SUGUMAR" AN EFFICIENT ALGORITHM FOR PRIVACY PRESERVING DATA MINING USING HEURISTIC APPROACH". INTERNATIONAL JOURNAL OF ADVANCED RESEARCH IN COMPUTER AND COMMUNICATION ENGINEERINGVOL. 1, ISSUE 9, NOVEMBER 2012
11. R. Srikant, "Fast algorithms for mining association rules and sequential patterns," UNIVERSITY OF WISCONSIN, 1996.
12. J. Han, M. Kamber,"Data Mining : Concepts and Techniques ", Morgan Kaufmann Publishers , Book, 2000.
13. F. H. AL -Zawaidah, Y. H. Jbara, and A. L. Marwan, "An Improved Algorithm for Mining Association Rules in Large Databases," Vol. 1, No. 7, 311-316, 2011
14. T. C. Corporation, "Introduction to Data Miningand Knowledge Discovery", Two Crows Corporation, Book, 1999.

