Australian Journal of Basic and Applied Sciences, 7(8): 497-504, 2013
ISSN 1991-8178

An Approach for Security in Mining Using Crossover Genetic Algorithm

[1]G. Kirubhakar and [2]Dr.C.Venkatesh

[1]Department of Computer Science and Engineering, Surya Engineering College, Erode, Tamilnadu, India.
[2]Dean, Faculty of Engineering, EBET Group of Institutions, Kangayam, Tamilnadu, India.

Abstract: Problem statement: The current trend is that organizations need to share data for mutual benefit. As a side effect of the mining algorithm some sensitive information is also revealed. This frequently results in individuals either refusing to share their data or providing incorrect data. In turn, such problems in data collection can affect the success of data mining, which relies on sufficient amounts of accurate data in order to produce meaningful results. **Approach:** Based on the analysis of shortcomings of earlier technologies this paper proposes a new method for securing data using crossover genetic algorithm. **Results:** An average of misclassification error was around 0 to 10% of the dataset for 95 % to 99% security level. **Conclusion:** The results obtained prove that the basic objectives of this approach is met with balanced security and utility compared to traditional methods.

Key words: security, privacy, genetic algorithm, clustering, Quantification.

INTRODUCTION

Data mining tools are increasingly being used to infer trends and patterns. In many scenarios, access to large amounts of personal data is essential in order to draw accurate inferences. However, publishing of data containing personal information has to be restricted so that individual privacy is not hampered. One likely answer is that rather than of issuing the entire database, only a part of it is released which can answer the ample queries and does not reveal sensitive information. Only those queries are answered which do not reveal sensitive information. But this can be considered to be loss of data (P.Samarati, 2001). The solution which we have proposed is an enhancement of the genetic algorithm based anonymization model. The difference being that the model proposed not only conserves the privacy of the dataset to a large extent, but also the data utility. And most importantly it preserves the distribution of data.

The suggested solution guarantees privacy against most of the attacks renowned to be likely to get private data of individuals. It furthermore provides the essential patterns to researchers and data miners without deviating from the original data values. Most importantly the solution does not disturb the distribution of the dataset.

2. Problem Definition:

Most of these secured mining algorithms in order to conserve the privacy and enhance the security end up losing essential data to a large span. This information loss does not solve the purpose of security maintenance because it renders the data useless (D.Aruna, 2011).

Thus there is a need to design a privacy preserving algorithm which not only preserves the privacy of the dataset but also does not lead to information loss. The main objective of the approach is to design a secured data mining system which transforms a dataset while preserving the privacy and distribution using genetic algorithm model.

The goal of this paper is to present technologies to solve security related data mining problems over large data sets with reasonable efficiency.

3. Literature Survey:

K Anonymity:

A number of papers have also appeared on the *k*-anonymity model recently. Other related works discuss the method of top-down specialization for privacy preservation, and workload-aware methods for anonymization (W.Du, 2004).

The Perturbation Approach:

Agrawal (2000) develops a new distribution-based data mining algorithm for the classification problem, whereas the techniques in Vaidya and Clifton (2002) and Rizvi and Haritsa (2002) develop methods for privacy-preserving association rule mining.

Corresponding Author: G. Kirubhakar, Department of Computer Science and Engineering, Surya Engineering College, Erode, Tamilnadu, India.
E-mail: ergalaxy81@gmail.com, kirubhakarg@gmail.com, Mob: 91-8438313186

Cryptographic Techniques:

Another branch of privacy preserving data mining which uses cryptographic techniques was developed. This branch became hugely popular (S.Laur, 2006).

Randomized Response Techniques:

Randomized Response technique was first introduced by Warner as a technique to solve a survey problem (H.Polat, 2005).

The Condensation Approach:

In condensation approach, data modification is used in order to modify the original values of a database that needs to be released to the public and in this way ensure high privacy protection (V.S. Verykios, 2004).

Proposed Genetic Algorithm Approach:

The goal of this paper is to present technologies to solve security related data mining problems over large data sets with reasonable efficiency based on genetic algorithm based anonymization. This method has a number of advantages over the above models in terms of disclosing the sensitive attributes in an effective way.

MATERIALS AND METHODS

4.1. Cluster Analysis:

Clustering is an important data mining problem. The goal of clustering, in general, is to discover dense and sparse regions in a dataset. Most previous work in clustering focused on numerical data whose inherent geometric properties (V.S. Verykios, 2004) can be exploited to naturally define distance functions between points. However, many datasets also consist of categorical attributes on which distance functions are not naturally defined. Recently, the problem of clustering categorical data started receiving interest.

4.2. Traditional k-Anonymity Security Method:

K-anonymity alone does not provide full privacy. Suppose attacker knows the non-sensitive attributes (zip, age and nationality) of Chen sui and Jason.

Zip	Age	Nativity
36900	23	Chinese
52013	32	American

Chen Sui

Jason

Fig. 1: Quasi data of individuals.

And the fact that Chinese have very low incidence of heart disease, then Homogeneity and Background knowledge attacks are possible as shown in Fig. 1.

Table 1: Original Data Set.

Id	Zipcode	Age	Nativity	Diagnostic
1	42302	25	Indian	Flu
2	52020	22	American	Flu
3	36900	23	Chinese	Heart disease
4	52013	29	American	Cancer
5	42025	31	Indian	HIV
6	13025	38	German	HIV
7	13022	36	German	HIV

Table 2: Anonymous data set (with k=4).

1	423**	<30	*	Flu	Chen Sui matches here (Background knowledge attack)
2	520**	<30	*	Flu	
3	369**	<30	*	Heart disease	
4	520**	<30	*	Cancer	
5	42***	3*	*	HIV	Jason matches here (Homogenity attack)
6	13***	3*	*	HIV	
7	13***	3*	*	HIV	
8	52***	3*	*	HIV	

Table 1 is anonymized by taking k value as 4, and then the resulting anonymous table looks as in table 2.

A privacy threat occurs either when an identity is linked to a record or when an identity is linked to a value on some sensitive attribute. These threats are respectively called record linkage and attribute linkage.

4.2.1. Attacks on k-Anonymous Table:

4.2.1.1. Record Linkage:

The record linkage occurs when some values q of quasi identifiers Q identifies a smaller number of records in the released dataset D. In this case, the record holder having the value q is vulnerable to being linked to a small number of records in D. However the main drawback in k-anonymity is its vulnerability to attribute linkage.

4.2.1.2. Attribute Linkage:

If some sensitive values are predominate in a group, an attacker has no difficulty to infer such sensitive values for a record holder belonging to this group. Such attacks are called attribute linkage. Particularly, k-anonymity suffers from two types of attribute linkage:

4.2.1.3. Homogeneity Attacks:

K-anonymity protection model can create groups that leak information due to lack of diversity in the sensitive attribute. In fact, k anonymization process is based on generalizing the quasi-identifiers but does not address the sensitive attributes which can reveal information to an attacker.

Table 2 shows the homogeneity attack. It shows how Jason's information is obtained by the homogeneity attack

4.2.1.4. Background Knowledge Attack:

Beside to homogeneity attacks, the background knowledge attacks can compromise privacy in k-anonymous database. In fact, an interloper can have knowledge that a priori enables him to guess sensitive data with high confidence.

Table 2 shows the Background knowledge attack. It shows how Jason's information is obtained by the background knowledge attack. This kind of attacks depends on other information available to an attacker.

5. Proposed Work:

The proposed work consists of three modules as shown in Figure 2.

Identification of attributes

Vertical partitioning

Sensitive based Anonymity

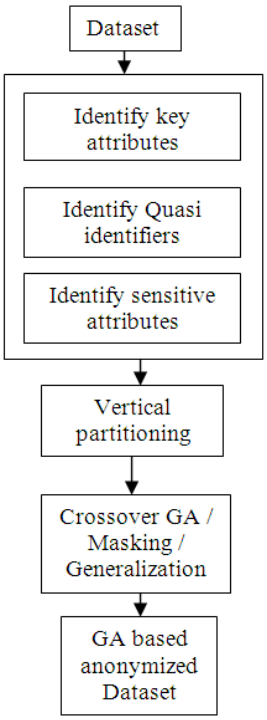

Fig. 2: Proposed GA based work.

5.1. Identification of Attributes:

The attributes of tables are classified into three classes. Identify the unique attribute such as id, as a key value. Identify the common attributes that are publicly available in all records as quasi attributes. Sensitive attributes are the attributes which are need to be protected. The selection of sensitive attributes is important because we need to anonymize only the most sensitive data to avoid the overhead and to increase the data utility. After identifying all of the attributes, suppression is applied only to the key attribute.

In our method the key/Identifier attributes have been removed and the quasi-identifier and sensitive attributes are usually kept in the released and initial data set.

Basically the values for the sensitive attributes are not available from any external source. This guarantees that an interloper cannot use the sensitive attributes to increase the chances of exposure. Unfortunately, an interloper may use record linkage techniques between quasi-identifier attributes and external available information to gain the identity of individuals from the modified data set. To avoid this possibility of privacy exposure, sensitive attribute based anonymity is anticipated.

5.2. Vertical Partitioning:

Vertical partitioning divides a table into multiple tables that contain fewer columns. The two types of vertical partitioning are normalization and row splitting. In this step we use the normalization method. Normalization is the standard database process of removing redundant columns from a table and putting them in secondary tables that are linked to the primary table by primary key and foreign key relationships**.** Assign unique class id to the table. After assigning class id divide the table into two. One table containing sensitive data along with quasi identifiers and the other table with non sensitive data.

Vertical partitioning query scans less data. This increases query performance. For example, a table that contains seven columns of which only the first four are usually referred may help to split last three columns into a separate table. Vertical partitioning must be carefully considered, because analyzing data from various partitions requires query that link the tables. Vertical partitioning also changes the performance if partitions are very large.

5.3. Sensitive Based Anonymity:

Anonymize only the most Sensitive attributes and quasi attributes to increase the data utility. Masking is applied only to those quasi and sensitive attributes. The sensitive attributes identified in the previous step is anonymized using different masking techniques. Zip code and credit card number are anonymized using masking technique called shuffling and then age and income are Generalized using recoding to produce the anonymized results.

Initially each field is anonymized using different anonymization techniques. Firstly genetic algorithm is applied to one field and then recoding techniques like normalization is applied to another field. Other techniques include generalization, masking , swapping which are applied to other respective fields.

Algorithm 1: Algorithm for Sensitivity Based Anonymity Method
Input: A dataset D, quasi-identifier attributes Q,
Sensitive values A, Anonymity parameter k
Output: Releasing Table D*
Step 1: Select Data set D from a Database
Step 2: Select Key attribute, Quazi-identifier attribute and Sensitive Attribute from give n attribute list.
Step 3: Select the set of most sensitive values A from list of all sensitive values that is to be preserved.
Step 4: For each tuple whose sensitive value belongs to set A If t[S] belongs to A then move all these tuples to Table T1 and rest to table T2.
Step 5: Find the statistics of quasi attributes of table T1 i.e. distinct values for that attribute and total no of rows having that value.
Step 6: Apply generalization/masking/crossover GA on quasi identifiers of table T1 to make it k- anonymized.
Step 7: Join both tables T1 and T2.
T*=T1+T2 which is table ready to release.

5.4. Generalization Using Recoding:

Generalization is an important technique for protecting privacy in data distribution. In the framework of generalization, k-anonymity is a strong notion of privacy. However, since existing k-anonymity measures are defined in terms of the most specific sensitive attribute (SA) values, algorithms based on these measures can have narrow eligible ranges for data that has a heavily skewed distribution of Sensitive attribute values and produce anonymous data that has a low utility.

Algorithm 2: Algorithm for Generalization using recoding
Input
Dataset with attribute of string and integer data type.

Output
Dataset with normalized attributes.
input data values from field f1;
store in array sa1;
set range from lower limit to upper limit;
if sa1[i] lies in range ri
assign respective value to the array ;
repeat test in every range
repeat till last
update array sal
end

5.5. Genetic Algorithm Based Anonymization:

Figure 3 shows the flow of control in the implementation of crossover genetic algorithm Randomly one position in the chromosomes is chosen

Child 1 is the head of chromosome of parent 1 with the tail of chromosome of parent 2

Child 2 is the head of chromosome of parent 2 with the tail of chromosome of parent 1

Algorithm 3:Crossover Genetic Algorithm
Input
A dataset with an attribute of data type integer.
Output
An anonymized dataset.
Steps
for each input values from field1
store in array of string s1;
convert each value to binary;
for each binary number less than length 12
add zeros to the most significant position;
invert first four digits;
store remaining digits in temp1;
store same positions of following value in temp2;
swap temp1 and temp2;
repeat till last
convert to decimal value;
display;
end

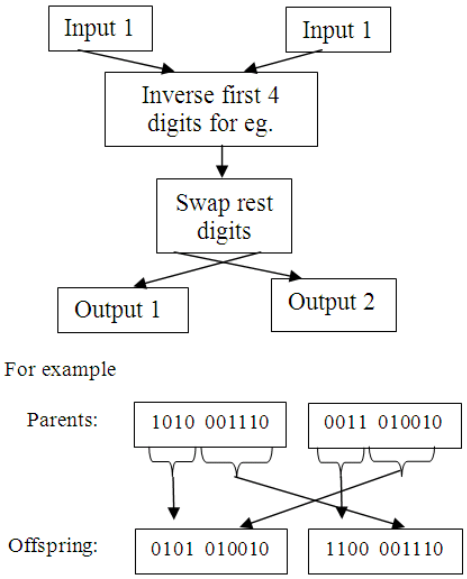

Fig. 3: Implementation of Crossover Genetic Algorithm.

Explanation:

The genetic algorithm is used to enhance the privacy and security of an attribute. The mechanism behind this algorithm is to first take a decimal number as input. This decimal digit is converted into a 12 digit binary equivalent and then two-point crossover is applied to the 12 digit number. That is, the first four digits are inverted left to right and the remaining 8 digits are swapped with the corresponding 8 digits of the following number. This renders the number completely impossible to track down.

The reason why genetic algorithm is preferred over other algorithms is that most of the algorithms use a mathematical equation to transform or anonymize the number. So the adversary can easily track down the original data if the mathematical equation used by the algorithm implemented is known. But, in case of a genetic algorithm, there is no particular mathematical equation used. So, there are n possible outcomes for a single encoded value. Thus, the difficulty in choosing between the n possibilities makes genetic algorithm the strongest to break through.

5.6. Cross Join:

In figure 4, we assign an identifier to the raw table and form a new table with an added attribute to each tuple using which forms the basis for performing the lossy cross join. After assigning class ID we divide the table into one containing sensitive data and the other with non sensitive data with class ID assigned to both the tables. Keeping the class ID as the foreign key we perform the join function to get the final anonymized dataset.

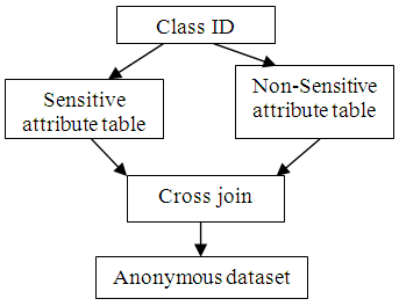

Fig. 4: Crossjoin.

By using the class id as the foreign key join both the sensitive and non-sensitive attribute tables using the sql cross join function to get the anonymized dataset. The result of this cross join enables the dataset to get multiple k-Anonymous records where the impostor finds it difficult to find the exact data of a person.

6. Experimental Results:

6.1. Detailed results:

Data set taken: Bank Dataset
Sensitive item: 4 attributes
Mining task: Clustering-Performance evaluation
Method adopted for sensitive disclosure: Crossover genetic algorithm based anonymization

Table 3 provides a brief description of the data including the attributes used in method, the number of distinct values for each attribute, the generalization that was used for Quazi identifier attributes and the height of the generalization hierarchy for each attribute.

Table 3: Description of Bank Data Set.

Attribute	Distinct	Generalization	Height
Age	74	10-,20-,30-	4
Marital status	7	Taxonomy Tree	3
Race	5	Taxonomy Tree	2
Sex	2	Person	1
Annual Income	20	Sensitive Attribute	
Occupation	14	Sensitive Attribute	
Medical Status	8	Sensitive Attribute	

Figure 5 shows privacy vs. utility graph for the sensitive attribute based anonymity. From figure it is clear that the traditional k-anonymity algorithm yields more information loss because all attributes in the data set are generalized. Only sensitive attributes are generalized in our approach.

Further, the runtime of the algorithm is reduced, due to the anonymization of selective attributes. Experiments show that this method can greatly improve the privacy quality without sacrificing accuracy.

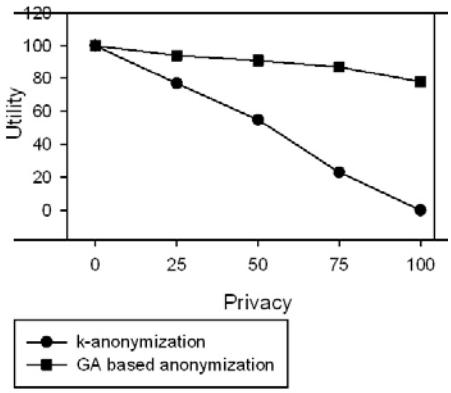

Fig. 5: Privacy vs. Utility.

7. Discussion:

Various techniques were analyzed and assessed based on the certain criteria, as shown in Table 4.

Table 4: Assessment using Evaluation Framework.

Techniques	Merits	Demerits
Anonymization	Identity disclosure security	Attribute disclosure, homogeneity attack and the background knowledge attack
Perturbation	Independent treatment of attributes	Difficulty in reconstruction
Randomized response	Easy implementation	High information loss.
Condensation	Works with pseudo-data	Algorithm is not redesigned
Cryptographic	Numerous algorithms – and quantification methods	Relies on encryption technique used
Multiparty techniques	Suitable for multiple attributes	Difficult to scale-
Crossover Genetic Algorithm based Anonymization	Identity disclosure Security Attribute disclosure Security	Need to be extended for complete GA based SA encryption

9. Conclusion:

It appears that complete privacy is impossible to maintain while allowing useful data-mining. To best knowledge this is the first effort toward a building block solution for the problem of securing data collection. This work can be extended in two directions: (a) combining cryptography and crossover GA to increase both accuracy and privacy; (b) designing new methods for security and measuring efficiency.

REFERENCES

Agrawal, R. and R. Srikant, 2000. Privacy-preserving data mining. Proceedings of SIGMOD00. DOI: 10.1145/342009.335438. pp: 439-450.

Aruna Kumari, D., Dr.K. Rajasekhar Rao, M. Suman, 2011. Privacy preserving clustering in data mining using vector quantization. Research journal of Computer science and engineering (RJCSE). ISSN: 2230-8563.

Du, W., Y. Han and S. Chen, 2004. Privacy preserving multivariate statistical analysis: Linear regression and classification. Proceedings of the Fourth SIAM International Conference on Data Mining. DOI: 10.1.1.38.595. pp: 222-233.

Laur, S., H. Lipmaa and T. Mielikainen, 2006. Cryptographically private support vector machines. Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. DOI: 10.1145/1150402.1150477. pp: 618-624.

Polat, H. and W. Du, 2005. SVD based collaborative filtering with privacy. The 20th ACM Symposium on Applied Computing, Track on Ecommerce Technologies, Santa Fe, New Mexico. DOI: 10.1145/1066677.1066860. pp: 791-795.

Rizvi, S.J. and J.R. Haritsa, 2002. Maintaining data privacy in association rule mining. Proceedings of the 28th VLDB Conference, Hong Kong, China. DOI: 10.1.1.19.9310. pp: 682-693.

Samarati, P., 2001. Protecting respondent's privacy in micro data release. IEEE Transaction on Knowledge and Data Engineering. DOI: 10.1109/69.971193. pp: 1010-1027.

Vaidya, J. and C. Clifton, 2002. Privacy preserving association rule mining in vertically partitioned data. Eighth ACM SIGKDD International conference on Knowledge Discovery and Data Mining, Edmonton, Alberta, CA. DOI: 10.1145/775047.775142. pp: 639-644.

Verykios, V.S., E. Bertino, I.N. Fovino, L.P. Provenza, Y. Saygin and Y. Theodoridis, 2004. State of the art in privacy preserving data mining. Proceedings of ACM SIGMOD. DOI: 10.1145/974121.974131. pp: 50-57.