International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 7 Issue IV, Apr 2019- Available at www.ijraset.com

Enhanced Database Security using Genetic Algorithm

Mrs. K. Pushpalatha[1], Dr. S. Gunasekaran[2], Ms. P. Glaris Dyana[3], Ms. R. Manjuladevi[4], Ms. R. Padma Priya[5], Ms. B. Pavithra[6]
[1]*Assistant Professor*, [2]*Professor*, [3, 4, 5, 6]*Department of CSE, Coimbatore Institute Engineering and Technology, Coimbatore,India*

Abstract: In today's era, databases tend to be an important element in many organizations for storing their confidential information. Because of rapid growth in telecommunication industry, these relational databases are openly available in the collaborative environments for information gain. This has eventually led to data thefts which degrades the performance of the database. Hence it is very essential to protect the database from various security attacks like insertion, deletion or alteration of data. Over these years, many encryption techniques have been proposed to protect the database from unauthorized database. Traditional database security techniques make changes in the database, which greatly compromise the quality of data in the data base. This paper discusses the performance of genetic algorithm compared with other data encryption techniques. We have investigated the system under different malicious environments. The Experimental result shows the effectiveness and performance of the system for both numeric as well as non- numeric relational database.
Keywords: Database, Security, Genetic algorithm, Robustness, Data recovery

I. INTRODUCTION

The tremendous growth in the telecommunication industry has motivated many organizations in making their relational databases openly available over the Internet. Hence, the security is very important to protect the database from unauthorized access. Database and database technology have a major impact on the growing use of computers. It is fair to say that database play a critical role in almost all areas where computers are used, including business, electronic commerce, engineering, medicine, law, education and library science. A database is a collection of data. For example: Consider the age, phone number and residential of the people you know. These data are accessed via queries, written in Structured Query Language (SQL). The architecture of database is illustrated in Fig 1.1

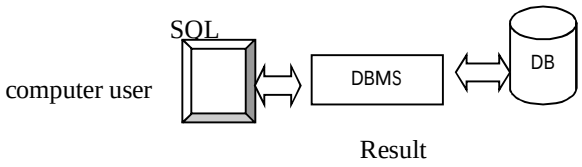

Figure 1.1 Database system

Many algorithms have been discussed in the technical literature for providing the security of the database. Data security refers to the process of securing the data from the database from many types of attacks. It is important to protect the data in order to protect it from data loss and modification of original data. The property of database security includes confidentiality, integrity and availability. Confidentiality is limiting the data access and Integrity is used to ensure that the data is accurate. Availability is to make sure the data is readily available to those who need it.

From past few decades, Security techniques are being used for ownership protection of database. With the sharing of databases across the Internet, the same requirement has evolved for relational databases. Data owners allow their data to be accessed and used remotely; therefore, may become a victim of data theft. Although, these technology helps them to prove their ownership through identifying data piracy, yet introduces permanent modifications into the data which are irreversible and the data is different from the original content. Consequently, data analysis and decision making on distorted version of data is not acceptable.

Relational database is a relatively emerging area. It is a candidate solution to ensure data ownership protection as well as data integrity. Relational data has different format as compared to other digital data such as audio, video, software, and images data. However a major drawback of these techniques is that they modify the data to a very large extent which often results in the loss of

data quality. There is a strong need to preserve the data quality in data so that it is of sufficiently high quality and fit for use in decision making as well as in planning processes in different application domains. Data quality can be defined as the appropriateness of data for its intended applications. There is no work has been conducted on overcoming the problems of data quality in the presence of malicious attacks. Not robust against heavy attacks (attacks that may target large number of tuples).Achieving robustness (attack resilience) in the presence of reversibility (ability to recover the watermark and the original data) is a challenging task.

This paper mainly focuses on threats in database security and measures taken to solve these threats. Also discussed the performance evaluation of genetic algorithm under malicious environments. The rest of the paper is organized as follows. Section 2 describes related works, Section 3 describes the performance evaluation of genetic algorithm under malicious environments. Section 4 concludes the paper.

II. RELATED WORKS

Sesay et al. proposed a database encryption scheme using cryptography. Here , the users are divided into two levels; Level1(L1) users and Level2(L2) users. The private data and unclassified public data are accessed by L1 users, whereas L2 users have access to their own private data and encrypted public data.[1]

Liu et al. proposed a novel database encryption mechanism. The recommended mechanism performs column-wise encoding that allows the users to classify the data into sensitive data and public data. This classification helps in selecting to encrypt only that data which is critical and leaves the public data untouched thereby reducing the burden of encrypting and decrypting the whole database, as result of which the presentation is not devalued.[2]

Kuo et al. presented a different approach to conceal data. In this scheme the picture is divided into static number of blocks. Each block is calculated along with the maxima and minima to mask of data. This mechanism increases the encrypting capacity of the data[3].

Kadhem, H.et al.,proposed Mixed Cryptography Database (MCDB) to encrypt databases over un trusted networks in a mixed form using many keys owned by different parties. This model is very useful in strengthening the security of sensitive data even if the database server is attacked at multiple points from the outside or inside. This work is used to give confidentiality, privacy and integrity of data in the database. The framework is explained in four steps: data is classified into groups, data in encrypted in database, Query Management Agent (QMA) and Result Analysis (RA) is used for query execution and finally, security of data storage and data transmission is verified. The results show the probability of outside attacker getting the encoding and decoding value it is measured and found to be very less.[4]

Gang Chen et al proposed a Database Encoding Scheme for enhanced sharing of data inside a database along with preserving data privacy. It is combined with conventional encryption and public key encryption along with using the fast speed of conventional encryption and convenience of public key encryption. A model is given which mainly gives the threats faced by the database. A user can encode the private data with a randomly generated working key with conventional encrypted algorithm and if a user wants to see the encrypt data by first decoding the private key with the passphrase and with this private key the working can be decrypted to access the key. A security index technology is used where it has strict access control and cannot be updated by even administrator. Future work will address two issues: research how to improve the security ad performance of database in terms of encryption algorithm and devise some self tuning mechanisms to manage keys. [5]

Isalm et al presented the Database Encryption scheme that provides large security without decreasing the performance of the database system while limiting the added time and cost of encryption. This method basically divides the data into sensitive data and insensitive data where the insensitive data is stored in the clear for fast retrieval and sensitive data is stored in encrypted for to conceal the data from the intruders. The classified sensitive dataset (classified and private) are encoded/decoded using Data Encryption standard technique. And their decryption is very fast as only one key is needed to decrypt a whole column of encrypted data. The examined encrypted private data need to decrypted separately using their own unique keys but the requests of private data are very rare [6].

Patel et al proposed a work to make the E-government Procurement Secure by protecting the data in the database in which encryption based Private Information Retrieval is used along with compression. This allows to save, proceed and return data in secure fashion. To secure the data transaction a Secure Box is introduced between the customer and central database where the important information of the user is stored in the encoded form and decrypted the data when requested by customer in a secure way. For retrieving the PIN of a vendor a recovering scheme is introduced. Contraction is used along with encryption so as to save the cost of storage and computation expansion. Huffman Algorithm is used to contract the bit file [7].

Pan et al this paper manipulates the original data and stores it in a database this framework mainly consists of four modules: Database catching, Virtual database encoding and Database encoding algorithm, Negative Database conversion which is applied to actual data. The original data passes through first three modules to generate the data required for fourth module to generate false data which stored along with positive data named as negative database. Returns invalid results for malicious users and retrieval of original data for legitimate queries performance of security work is O(n) which is very high can be compensated to the low-security high risk of data for other applications. This works only with INSERT and SELECT query and future development is to work with UPDATE query.[8]

Son et al proposed an Adaptive Policy called secure two-phase locking loop to address the requirement of multilevel security in transaction scheduling and concurrency control. If two conflicting transactions arise, i.e. one is blocked and waiting for other transaction to release the lock, then balance between security and priority is given by looking up the past history. The two methods by which the adaptive policy works is: the security factor and the factor resembling the deadline-miss ratio. This system ensures only partial security. [9]

III. PERFORMANCE ANALYSIS OF GENETIC ALGORITHM

Ownership rights of the databases need to protect from malicious recipients. Recent research studies enunciate that computational intelligence techniques, such as Genetic algorithm (GA) is a promising branch of evolutionary computation that model hard constrained optimization problems using biological inspired computing algorithms. GA can also be modeled as an optimization problem as demonstrated by some recent research works and that use different data formats and the results are quite encouraging. An optimal duplicate value is created through the GA and inserted into the selected feature of the relational database in such a way that the data quality remains intact.

A. Genetic Algorithm

A genetic algorithm is an optimization technique inspired by Charles Darwin's theory of natural evolution. This algorithm uses a collection of datasets, called chromosomes which provide a feasible solution for the problem. It is initialized with the features of randomness chromosomes. The feature size of GA is determined by the number of attributes and obtaining the optimal information through GA. Here we have investigated the efficiency of genetic algorithm for improving the robustness of data in the database. [10]

1) *Selection:* Selection is the stage of genetic algorithm which is used to select the fittest and let them pass their genes to the next generation. Individual are selected based on their fitness scores, which are then normalized. Normalization means dividing the values of each individual
2) *Cross Over:* Crossover is the most significant phase in genetic algorithm. Each genetic representation can be recombining with different crossover operators. For each pair of data to be mated in a database. A random crossover attributes are selected and it is swapped to get new values. There are two types of crossover mentioned below:

a) ***One Point Crossover:*** A random crossover point is selected and the tails of the data values are swapped, to get new values. The fig 3.1 shows the one point crossover.

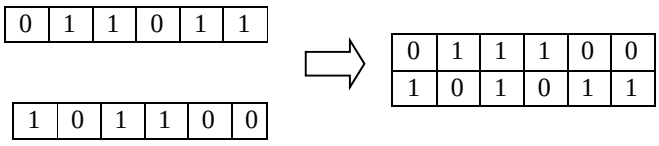

Fig 3.1 one point crossover

b) *Multi Point Crossover:* Multi point cross over is same as one point but in this alternating segments are swapped to get a new values. Fig 3.2 shows the value of multiple crossovers.

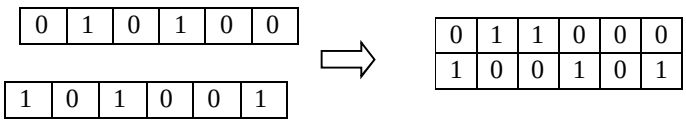

Fig 3.2 Multipoint crossover

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 7 Issue IV, Apr 2019- Available at www.ijraset.com

3) Mutation: In certain new values formed, some of their values can be subjected to a mutation with a low random probability which alters one or more mean values in a chromosome from its initial state. The purpose of mutation in GA is to preserve the original data in the database. Fig 3.3 shows the value of mutation.

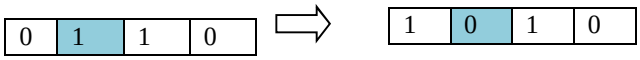

Fig 3.3 Mutation

IV. RESULT AND DISCUSSION

Data protection is the process of securing important information from corruption, compromise or loss. The importance of data protection increases as the amount of data created and stored continues to grow at high efficiency. Consequently, a large part of a data protection strategy is ensuring that data can be restored quickly after any loss of data. Using genetic algorithm the highly confidential data's are encoded using mutual information and secured. So nobody can hack the data from the database. This would be an efficient method to secure the data. The attributes and dataset used in this paper are mentioned below:

Table 1- Attributes and data types

ATTRIBUTES	DATA TYPE
Age	Int
Sex	Char
Cp	Float
Trestbps	Float
Chol	Float
Fbs	Float
Restecg	Float
Thalach	Float
Exang	Float
Oldpeak	Float
Slope	Float
Ca	Float
Thal	Float

A. Preprocessing

Datasets involving large number of features or large number of tuples. Those data's are contains some noises and symbols. The data set is a comma separated value (CSV) file. The mutual information and the further process cannot be done without finishing the pre process step. In this, the comma delimiter is used to split every feature values and it stored in the database. Table 2 represents the values of preprocessing.

Table 2- preprocessing

Age	Sex	Cp	Trestbps	Chol	Fbs	Restecg
63	1	1	145	233	1	2
67	1	4	160	286	0	2
37	1	3	130	250	0	0
41	0	2	130	204	0	2
56	1	2	140	230	0	0
62	0	4	146	268	0	2
57	0	4	178	354	0	2
63	1	4	130	254	0	0
53	1	4	165	203	0	2

Here **0** is mentioned as **female**, **1** is mentioned as **male**.

Mutual Information is a well known information theory (concept), statistically measures the amount of information that one feature contains about the other features in a database. In mutual information, it is used to select a suitable (candidate) feature from the database for encoding. The mutual information measure for determining relative importance of features. Table 3 represents the value of mutual information.

Table 3- Mutual information

Serial no	Features	MI
1	Age	0.124396
2	Sex	0.125381
3	Cp	0.5824781
4	Trestbps	2.2482361
5	Chol	4.965495
6	Fbs	0.027389
7	Restecg	0.182595
8	Thalach	2.759071
9	Exang	0.06025
10	Oldpeak	0.19172

B. Feature Selection

The value of MI of each feature is then used to rank the features. The attacker can try and predict the feature with the lowest MI in an attempt to guess which feature has been encoded. To deceive the attacker for this particular scenario, a secret threshold can be used for selecting the feature for data encoding. In this context, the data owner can define a secret threshold based on MI of all the features in the database. Selected features are: Fbs, Exang, Cp, Age, Sex.

C. Database Encoding

GA is a population-based computational model; basically inspired from genetic evolution GA evolves a potential solution to an optimization problem by searching the possible solution space. In the search of optimal solution, the GA follows an iterative mechanism to evolve a population of chromosomes. The GA preserves essential information through the application of basic genetic operations to these chromosomes that include: selection, crossover, mutation and replacement. The GA evaluates the quality of each candidate chromosome by employing a fitness function. The evolution mechanism the GA continues through a number of generations, until some termination criteria is met. Table 4 represents the value of encoding process.

Table 4- Encoding

Age	Sex	Cp	Trestbps	Chol	Fbs
61.63	-0.37	1	145	233	-0.37
65.63	-0.37	4	160	286	-0.37
35.63	-1.37	3	130	250	-1.37
39.63	-0.37	2	130	204	-0.37
54.63	-1.37	2	120	230	-1.37
60.63	-0.37	4	140	268	-0.37
55.63	-0.37	4	120	354	-0.37
61.63	-0.37	4	130	254	-0.37
51.63	-0.37	4	140	203	-0.37
55.63	-0.37	4	140	233	-0.37

D. Data Decoding

In the decoding process, the first step is to locate the features which have been encoded. The process of optimization through GA is not required during this phase. We use a decoder , which calculates the amount of change in the value of a feature that does not affect its data quality. This decoder decodes the value by working with one bit at a time. After detecting the data string, some post processing steps are carried out for error correction and data recovery. The optimized value of b computed through the GA is used for regeneration of original data. Table 5 represents the value of decoding process.

Table 5- Decoding

Age	Sex	Cp	Trestbps	Chol	Fbs	Restecg
63	1	1	145	233	1	2
67	1	4	160	286	0	2
37	1	3	130	250	0	0
41	0	2	130	204	0	2
56	1	2	140	230	0	0
62	0	4	146	268	0	2
57	0	4	178	354	0	2
63	1	4	130	254	0	0
53	1	4	165	203	0	2

E. Attacker Channel

In this process if the attacker not able to find the original data for a duplicate dataset there is a chance of deleting the existing dataset .so, to control the data loss from malicious attack, attacker channel is used. If attacker try to delete the dataset it will show that one row is affected to them. But the dataset will not get deleted in original database. Table 6 represents the value of attacker channel.

Table 6- Attacker channel

Age	**Sex**	**Cp**	**Trestbps**	**Chol**	**Fbs**
61.63	-0.37	1	145	233	-0.37
65.63	-0.37	4	160	286	-0.37
35.63	-1.37	3	130	250	-1.37
39.63	-0.37	2	130	204	-0.37
54.63	-1.37	2	120	230	-1.37
60.63	-0.37	4	140	268	-0.37
55.63	-0.37	4	120	354	-0.37
61.63	-0.37	4	130	254	-0.37
51.63	-0.37	4	140	203	-0.37
55.63	-0.37	4	140	233	-0.37

V. CONCLUSION

Database security has become a hot research as the increasing demand of ownership protection when sharing database information. Traditional database security techniques make changes in the database, which greatly compromise the data quality. Genetic algorithms used to solve this problem because they can recover original data from the database and ensure data quality. Many security techniques have been proposed, but those techniques are not robust enough against malicious attacks. In this paper, a novel reversible and robust technique for securing numerical data of relational databases has been proposed. We unite GA with a new proposed method to minimize distortion and improve robustness for database security. The characteristics of GA are embedding a data bit repeatedly in one group optimally by GA and ensuring the quality of database. Since the grouping and the majority voting mechanism are used, the proposed method allows extracting most of the information and recovering a large portion of the data even after being subjected to malicious attacks. A number of experiments have been conducted with different attack scenarios. Our future work is to develop reversible and robust security for non-numeric data and propose schemes for the shared databases in distributed environments.

REFERENCE

[1] Sesay, S.; Zongkai Yang; Jingwen Chen; Du Xu; A secure database encryption scheme; Consumer Communications and Networking Conference, 2005. CCNC. 2005 Second IEEE ; Publication Year: 2005 , Page(s): 49 – 53
[2] Lianzhong Liu and JingfenGai, "A New Lightweight Database Encryption Scheme Transparent to Applications", 6th IEEE International Conference on Industrial Informatics, 13-16 July 2008, pp.135-140
[3] Wen-Chung Kuo, Dong-Jin Jiang, Yu-Chih Huang, "A Reversible Data Hiding Scheme Based on Block Division", Congress on Image and Signal Processing, Vol. 1, 27-30 May 2008, pp. 365-369
[4] Kadhem, H.; Amagasa, T.; Kitagawa, H.;A Novel Framework for Database Security based on Mixed Cryptography; Internet and Web Applications and Services, 2009. ICIW '09. Fourth International Conference on ; Publication Year: 2009 , Page(s): 163 – 170
[5] Gang Chen; Ke Chen; JinxiangDong;A Database Encryption Scheme for Enhanced Security and Easy Sharing; Computer Supported Cooperative Work in Design, 2006. CSCWD '06. 10th International Conference on ; Publication Year: 2006 , Page(s): 1 – 6
[6] Islam, M.S.; Dey, S.; Kundu, G.; Hoque, A.S.M.; A Solution to the Security Issues of an EGovernment Procurement System; Electrical and Computer Engineering, 2008. ICECE 2008. International Conference on ; Publication Year: 2008 , Page(s): 659 – 664
[7] Patel, A.; Sharma, N.; Eirinaki, M.; Negative Database for Data Security; Computing, Engineering and Information, 2009. ICC '09. International Conference on ; Publication Year: 2009 , Page(s): 67 – 70
[8] Pan, L.; Using Criterion-based access control for multilevel database security; Electronic Commerce and Security, 2008 International Symposium on; Publication Year: 2008 , Page(s):518 – 522
[9] Son, S.H.;Supporting Timeliness and Security in Real-Time Database Systems; RealTime Systems, 1997. Proceedings, Ninth Euromicro Workshop on; Publication Year: 1997, Page(s): 266 – 273
[10] A study on genetic algorithm and its application- L.Haldurai, T.Madhubala, R.Rajalakshmi