INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS.....

Disquisition of a Novel Approach to Enhance Security in Data Mining

Gurpreet Kaundal[1], Sheveta Vashisht[2]

[1]Student
Lovely Professional University,
Phagwara, Pin no. 144402
gurpreetkaundal03@gmail.com

[2]Asst. Professor,
Lovely Professional University
Phagwara, Pin no. 144402
sheveta.16856@lpu.co.in

Abstract: *Data Mining is a process of knowledge discovery means its use to discover important information from a large database. Generally it's used to mine data from. Data mining techniques are the result of a long process of research and product development. Many data mining tools currently operate outside of the warehouse, requiring extra steps for extracting, importing, and analyzing the data. Here security is a challenge issue in data mining because number of attacks is possible when a normal user accessing information from database through data mining so now day's security became a major issue. Here we are going to purpose a novel approach which is based hill cipher, iterative data mining, etc. this technique will provide the security of data during its stored in database as well as when user is accessing it.*

Keywords: *Data mining, Hill cipher, iterative data mining, genetic algorithm, security.*

1. INTRODUCTION

Data mining is the extraction of hidden predictive information from large databases. It is a new technology, which helps the companies to focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer the questions that were too time consuming to resolve.[1] Data mining techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resources. It can be integrated with new products and systems as they are brought online. Data mining tools can analyze massive databases to deliver answers to various questions.

The Foundations of Data Mining: Data mining techniques are the result of a long process of research and product development. This evolution began when business data was first stored on computers, continued with improvements in data access. Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery.

The Scope of Data Mining: Data mining technology can generate new business opportunities by providing these capabilities:

- **Automated prediction of trends and behaviors**. Data mining automates the process of finding predictive information in large databases. Data mining uses data on past promotional mailings to identify the targets most likely to maximize return on investment in future mailings. Other predictive problems include forecasting bankruptcy and other forms of default, and identifying segments of a population likely to respond similarly to given events.

- **Automated discovery of previously unknown patterns**. Data mining tools sweep through databases and identify previously hidden patterns in one step. An example of pattern discovery is the analysis of retail sales data to identify seemingly

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS.....

unrelated products that are often purchased together. Other pattern discovery problems include detecting fraudulent credit card transactions and identifying anomalous data that could represent data entry keying errors.

Techniques Used in Data Mining: The most commonly used techniques in data mining are:

- **Artificial neural networks**: Non-linear predictive models that learn through training and resemble biological neural networks in structure.
- **Decision trees**: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID) .
- **Genetic algorithms**: Optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of evolution.
- **Nearest neighbor method**: A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k ³ 1). Sometimes called the k-nearest neighbor technique.
- **Rule induction**: The extraction of useful if-then rules from data based on statistical significance.

Architecture for Data Mining: Many data mining tools currently operate outside of the warehouse, requiring extra steps for extracting, importing, and analyzing the data. Furthermore, when new insights require operational implementation, integration with the warehouse simplifies the application of results from data mining. The resulting analytic data warehouse can be applied to improve business processes throughout the organization, in areas such as promotional campaign management, fraud detection, new product rollout, and so on. Figure 1 illustrates architecture for advanced analysis in a large data warehouse.

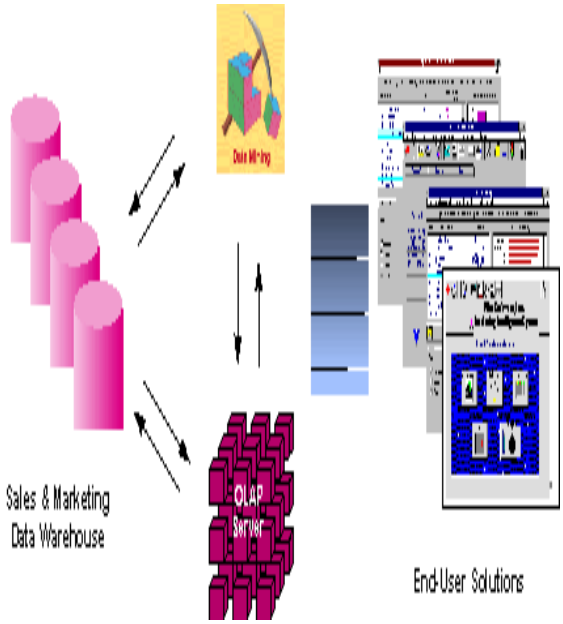


Figure 1: Integrated Data Mining Architecture

An OLAP (On-Line Analytical Processing) server enables a more sophisticated end-user business model to be applied when navigating the data warehouse.[2] The multidimensional structures allow the user to analyze the data as they want to view their business – summarizing by product line, region, and other key perspectives of their business. The Data Mining Server must be integrated with the data warehouse and the OLAP server to embed ROI-focused business analysis directly into this infrastructure. An advanced, process-centric metadata template defines the data mining objectives for specific business issues like campaign management, prospecting, and promotion optimization. Integration with the data warehouse enables operational decisions to be directly implemented and tracked. As the warehouse grows with new decisions and results, the organization can continually mine the best practices and apply them to future decisions.

Knowledge Discovery in data mining: Data Mining is also popular for Knowledge Discovery in Databases (KDD). [2] It generally refers to the nontrivial extraction of implicit, unknown and potentially useful information from data in databases. Data mining and knowledge discovery in databases are frequently treated as synonyms.

The figure 2 shows the data mining as a step in an iterative knowledge discovery process.

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS.....

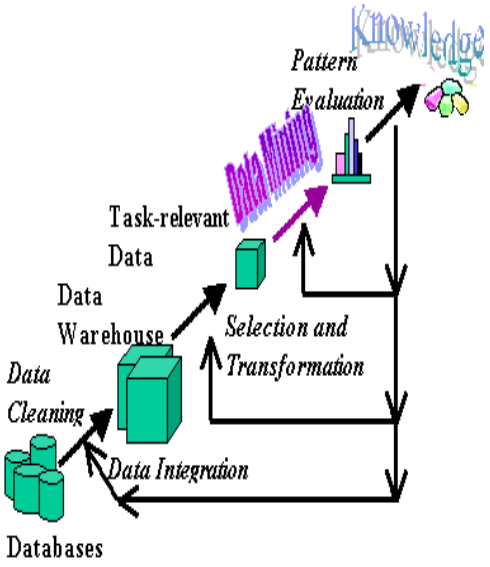


Figure 2: Knowledge discovery process

The Knowledge Discovery in Databases process comprises of a few steps leading from raw data collections to some form of new knowledge. The iterative process consists of the following steps:

- **Data cleaning**: also known as data cleansing, it is a phase in which noise data and irrelevant data are removed from the collection.
- **Data integration**: at this stage, multiple data sources, often heterogeneous, may be combined in a common source.
- **Data selection**: at this step, the data relevant to the analysis is decided on and retrieved from the data collection.
- **Data transformation**: also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.
- **Data mining**: it is the crucial step in which clever techniques are applied to extract patterns potentially useful.
- **Pattern evaluation**: in this step, strictly interesting patterns representing knowledge are identified based on given measures.
- **Knowledge representation**: is the final phase in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results.

2. LITERATURE REVIEW

Taeshik Shon et al (2007) proposed an enhanced SVM approach framework for detecting & classifying the novel attacks in network traffic. [3] The overall framework consist of an enhanced SVM- based anomaly detection engine & its supplement components such as packet profiling using SOFM, packet filtering using PTF, field selection using Genetic Algorithm & packet flow-based data preprocessing. SOFM clustering was used for normal profiling. The SVM approach provides false positive rate similar to that of real NIDSs.

Snehal A. Mulay et al (2010) propose algorithm for IDS that gives better performance results for multiclass classification. In this paper the decision tree model consists of two class SVMs. The structure of the tree is determined by distance between two class patterns and the number of each class pattern. [4] It is the integration of decision tree model and SVM model.

Shailendra Kumar Shrivastava et al. (2011) focuses on the dimensionality reduction using feature selection. The Rough set support vector machine (RSSVM) approach deploy Johnson's & genetic algorithm of rough set theory to [5] find the reduce sets & sent to SVM to identify any type of new behavior either normal or attack one.

Sadiq Ali Khan (2011) proposes a genetic algorithm can be effectively used for formulation of decision rules in intrusion detection through the attacks which are more common can be detected more accurately. The proposed genetic algorithm is used to tune [6] the membership function which has been used by IDS. A survey was performed using approaches based on IDS, and on implementing of Gas on IDS.

Norouzian M.R et al (2011) defined Multi-Layer Perceptron (MLP) for implementing & designing the system to detect the attacks & classifying them in six groups with two hidden layers of neurons in the neural networks. [7] Host based intrusion detection is used totrace system calls. This system does not exactly need to know the program codes of each process. Normal & intrusive behavior are collected through system call & analysis is done through data mining & fuzzy technique.

3. PRESENT WORK

The present work uses hybrid approach advance hill-cipher algorithm generates cipher and DES algorithm generates key. After Key generation, a secret data can be embedded to original image. One additional feature is also added that is to add password authentication up to 24 characters & it. Secret

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS.....

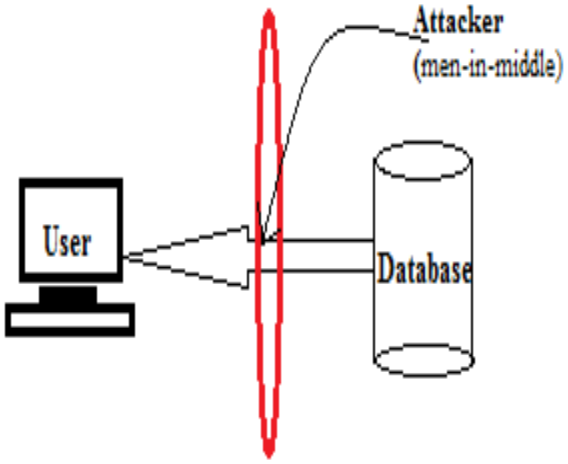
text hiding basically deals with hiding the text in the digital representation of the image. This work uses to enhanced approach using Advance Hill-cipher & DES techniques is discussed encryption add one more feature that is password authentication to enhance the security. Add the password up to 24 characters and it will be used for further query when are going to retrieve back our secret data. The requirements of a secret text hiding system when used for cryptographic purposes are of high hiding capacity and imperceptibility. Keeping in view some conflicting features a reasonable amount of text data has been taken to be embedded in the cover medium. The strength of the algorithm is discussed by explaining the complexities in encryption and decryption. The concept can be further enhanced by adding digital signatures to the cipher and the key.

Network security covers a variety of computer networks, both public and private that are used in everyday jobs conducting, transactions communications in businesses, government agencies and individuals. Network security is involved in organizations, enterprises and other types of institutions. The requirements of a secret text hiding system when used for cryptography purposes of high hiding capacity and imperceptibility. Keeping in view some conflicting features a reasonable amount of text data has been taken to be embedded in the cover medium .In proposed scheme any image is taken as input that calculate the elapsed time for text hide and image process which gives better results as compared to other techniques like Hill Cipher and advanced Hill Cipher. The proposed work is a secret text hiding approach, which is the combination of Advance Hill cipher & DES techniques for securing confidential data from unauthorized access. DES is used for data hiding processing in little time.

4. PURPOSED WORK

Security in data mining: Now security is a major issue in data mining. If we talk about IT sector there information is very import data, that information can be there client contacts or it may be the information about their accounts, in short that information is very confidential information that they never want to disclose. But because of security attacks an attacker can access this information.

For example: let's take example of men-in-middle attack, suppose you are working in a XYZ Pvt. Ltd. And you are accessing their accounts information by using data mining from their database, suddenly an attacker perform men-in-middle attack and he disclose your all important information.



Here in this figure its shown that user is discovering some information and an attacker is also accessing same information from path by using men-in-middle attack.

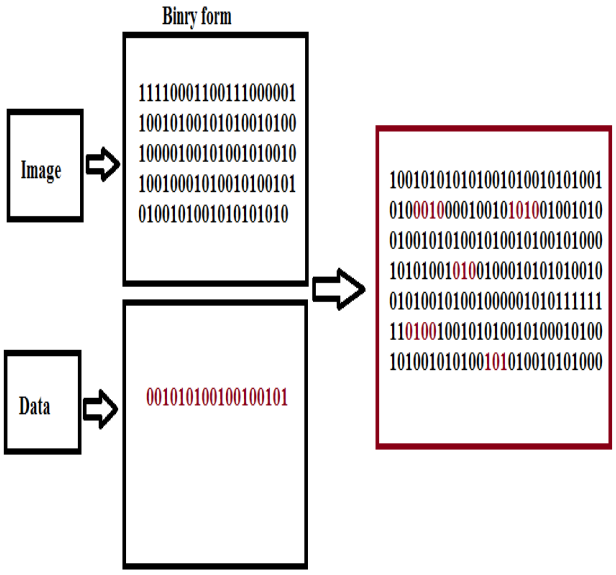
So here to prevent these type of attacks we proposed a methodology which will works as following:

To protect our important information first of all we need to hide our information so that an attacker can't watch it. So in our work we will store our information by hiding it into images. We will store our information inside a image using Hill cipher algorithm and then we will store it into database. After that we will create a data mining system with anti- hill cipher algorithm.

By this scenario when we will access our information we retrieve it in form of image, now in this case if an attacker perform any attack and he perform attack successfully then he will take only image because our information is hidden inside the image, only user can discover that information from image. So it will make our information secure.

To store our information into image first of all we convert our image and information into binary and then by hill cipher algorithm we will mix all data as shown in figure:

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS.....



Algorithm: Hill cipher

$$\begin{pmatrix} c1 \\ c2 \\ c3 \end{pmatrix} = \begin{pmatrix} k11 & k12 & k13 \\ k21 & k22 & k23 \\ k31 & k32 & k33 \end{pmatrix} \begin{pmatrix} p1 \\ p2 \\ p3 \end{pmatrix} \pmod{26}$$

Here 'p' is a plain text and 'c' is cipher text and 'k' is a secret key

Genetic algorithm for data mining: genetic algorithm (GA) is a search heuristic that mimics the process of natural selection. it generate solutions to optimization problems using techniques inspired by natural evolution

Iterative Data Mining: it will discover information on the basis of iterations.

- Preserving Row and Column Margins
- Preserving Clustering Structure
- Preserving Itemset Frequencies

REFERENCES

[1] http://www.thearling.com/text/dmwhite/dmwhite.htm

[2] Suhail Ansari, Ron Kohavi, Llew Mason, and Zijian Zheng ,Integrating E-Commerce and Data Mining: Architecture and Challenges, WEBKDD' 2000 workshop: Web Mining for E-Commerce -- Challenges and Opportunities.

[3] Taeshik Shon, Jong Sub Moon, (2007) "A Hybrid Machine Learning Approach to Network Anomaly Detection", Information Sciences 2007, Vol: 177, Issue: 18, Publisher: USENIX Association, pp- 3799-3821, ISSN: 00200255, DOI:10.1016/j.ins-2007.03.025.

[4] Snehal A. Mulay, P.R. Devale and G.V. Garje, (2010) "Intrusion Detection System using Support Vector Machine and Decision Tree", International Journal of Computer Applications (0975 – 8887) Volume 3 – No.3.

[5] Snehal A. Mulay, P.R. Devale and G.V. Garje, (2010) "Intrusion Detection System using Support Vector Machine and Decision Tree", International Journal of Computer Applications (0975 – 8887) Volume 3 – No.3.

[6] Sadiq Ali Khan, (2011) "Rule-Based Network Intrusion Detection Using Genetic Algorithm", International Journal of Computer Applications, No: 8, Article: 6, DOI: 10.5120/2303-2914.

[7] Norouzian.M.R, Merati.S, (2011) "Classifying Attacks in a Network Intrusion Detection System Based on Artificial Neural Networks", in the Proceedings of 13th International Conference on Advanced Communication Technology (ICACT), ISBN:978-1-4244-8830-8,pp-868-873.