# Application of Data Mining Techniques for Information Security in A Cloud

D . Bhanu Sravanthi,
Department of computer science and engineering,
SV College of Engineering,
Tirupati,India

A . Nageswara rao,
Head of computer science department,
SV College of Engineering,
Tirupati,India

**Abstract--- Data Mining is a process of extracting potentially useful information from raw Data, so as to improve the quality of the information service. With the rapid development of the Internet, the size of the data has increased from KB level to TB even PB level;Cloud computing can provide infrastructure to massive and complex data of data mining, as well as new challenging issues for data mining of cloud computing research are emerged.Data mining techniques are very important in the cloud computing paradigm.The integration of data mining techniques with Cloud computing allows the users to extract useful information from a data warehouse that reduces the costs of infrastructure and storage. As people are launching themselves into the e-world completely, the Cloud as a service is now shaping up the future.Since the cloud services are available through internet, it is the need of our to prevent cyber attacks and at the same time trace the ill-willed persons for the sake of securing business,personal information and nation.Data Mining techniques and algorithms contribute tremendously to this taskof assuring security of information on the cloud. In this paper, review of various data mining techniques and algorithms is presented which can help achieve security and privacy of information on cloud.**

***IndexTerms:Datamining,cloudcomputing,intrusion detection,privacypreserving***

## 1.INTRODUCTION:

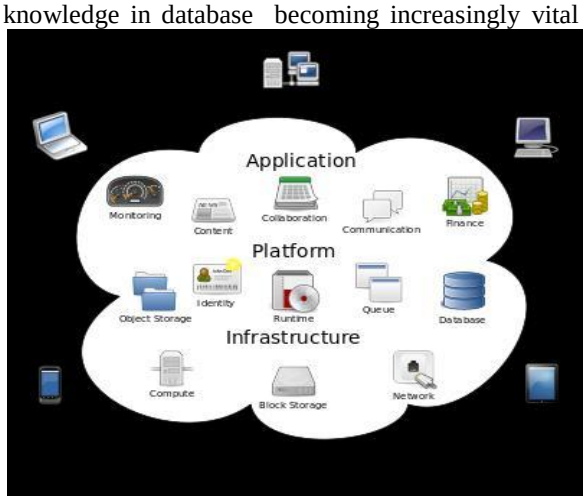
Data mining has been an effective tool to analyse data from different angles and getting useful information from data. Classification of data,categorization of data, and to find correlation of data patterns from the dataset.On the other hand, challenges as data storage and transfer approaches need to deal with prohibitive amount of data. The management of data resource and dataflow is becoming the main bottleneck. Large data set has

become a major challengeand dataintensive computing is now considered as the "fourth paradigm" in scientific discovery a fter theoretical, experimental,and computationalscience.

The internet is becoming an increasingly vital tool in everybody's life, both professional and personal, as its user and becoming more numerous. The most revolutionary concept of recent year is Cloud Computing. Many companies are choosing as an alternative to building their own IT infrastructure to host database or software, having a third party to host them on its large servers, so company's would have access to its data and sof tware over the Internet. The cloud services are accessible to the user through internet hence security of cloud projects cyber security as the prime concern.Cyber security involves protecting information by preventing, detecting, and responding to attacks.

The use of cloud computing is gaining popularity due to its mobility, huge availability and low cost. On the other hand it brings more threats to the security of the company's data and information. In recent years,data mining techniques have evolved and become more used, discovering

knowledge in database becoming increasingly vital




*Transferring data from one server to another server through the data mining*

In business,medicine,science,engineering and spatialdata

## 2. INFORMATION SECURITY

Information security (sometimes shortened as Info-Sec) is the practice of protecting information from unauthorized user, disclosure, disruption, modification or destruction. Computer and communication systems repeatedly suffer security and privacy attacks. Nowadays, most of the companies spend good amount of money on their network security and privacy requirements. Four key features of information security are mentioned in figure 1.

Information security technology is an essential component for protecting public and private computing infrastructures.Advancement in technology is making people more oriented towards frequent use of information technology resulting in more usage of online resources which in turn is giving rise to a large number of security threats to these resources.

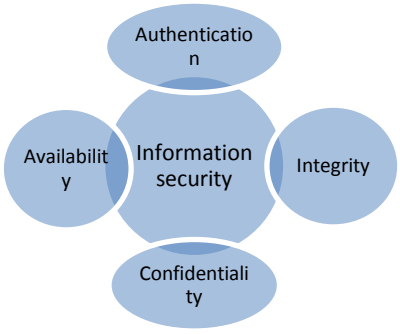



*Fig 1: Information Security Attributes*

The increasing number of security breaches is requiring some security agencies to deploy security policies and mechanisms to limit or wipeout these threats. Some of the Indian cyber security agencies are mentioned in the figure 2 below:

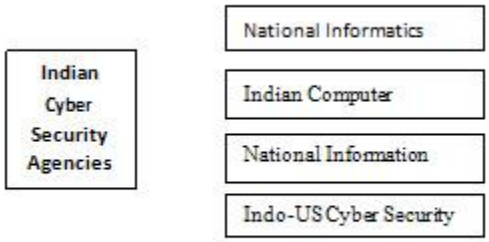



**Fig2 Indian Cyber Security Agencies**

*2.1Types of Attacks*

One of the common ingredients of cyber crime is the malicious code such as viruses, worms, and Trojan horses.*Active Attack* is an intentional threat that attempts to modify a system, its resources, its data or its operations whereas *passive attack* is also a threat that attempts to learn or make use ofinformation from a system but does not attempt to alter the system, its resources, its data or its operations.

*2.2 Types of Risks*

*Viruses* - This is a malicious code that requires the end user to perform some action before it infects the computer like opening an email attachment or going to a particular web page.

*Worms* - Worms propagate without user intervention and start by exploiting software vulnerability. Similar to viruses, worms can spread through email,web sites,or network-based software.The key characteristic of worm is that it propagates automatically.

*Trojan horses* - A Trojan horse program is software that does not let the user know its actual consequences. For example, a program which claims that it will speed up your computer may actually be sending confidential information to a remote intruder.

*Hacker,Attacker,Intruder or Denial of Service* - These terms are applied to the people who seek to exploit weaknesses in software and computer systems for their own gain. Although it is difficult to

comment on one's intention for doing this because they may or may not cause direct harm to the end user but denial of service definitely deprives the end user to be properly served. The various types of attacks can be broadly classified as
shown in the figure 3 below:

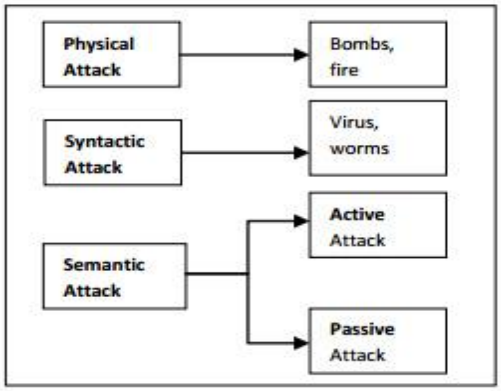



**Fig 3: Types of Attacks**

## 3.CLOUD ARCHITECTURE

Cloud computing is not a technology but a service which can be made available on demand through internet. In today's world where people are looking for services like infrastructure, software, platform etc. conveniently, fast and at low cost, a CLOUD provides the best solution. Hence, user pays only for the amount of service used and the duration for which the service is used thereby reducing the usage, installation and maintenance cost. The National Institute of Standards and Technology (NIST) [20] mentions the essential characteristics of cloud computing as resource pooling, on-demand service, broad network access, measured service, and rapid elasticity. Four deployment models for cloud architecture are described below:

- *Private cloud*: The cloud infrastructure is operated for a private organization. It is generally managed by an organization or a third party.
- *Communitycloud*:Thecloud nfrastructure is shared by several organizations and supports a specific community that has communal concerns(e.g.,security requirements,policy,andcompliance considerations). It is again managed by a third party or an organisation and may exist inside or outside the premises.
- *Public cloud*: The type of cloud infrastructure is made available to the general public or a large industry group and is owned by an organization selling cloud services.
- *Hybrid cloud*: The cloud infrastructure is a composition of two or more clouds (private,community, or public) that remain unique and independent entities, but are bound together by some standardized or proprietary technology, which can enable portability of application and data.

In cloud computing, the available service models are:
*Infrastructure as a Service (IaaS):* It provides the consumer with the potential to stipulate processing, storage, and other fundamental computing resources, and allows the consumer to deploy and run software, which may include operating systems and other applications. The architecture of cloud is shown in figure 4.

*Platform as a Service (PaaS):* It provides the consumer with the capability to deploy onto the cloud infrastructure; consumer created or acquired applications, produced using programming languages and tools supported by the provider. The consumer has organize the deployed applications only does not supervise or run the underlying infrastructure like servers, network, operating systems, or storage, etc.

*Software as a Service (SaaS):* It provides the consumer with the capability to use the provider's applications running on a cloud infrastructure. These applications are available from different client devices, through interface, like web browser. Similar to PaaS, the customer has no right to manage or structure the basic cloud infrastructure.

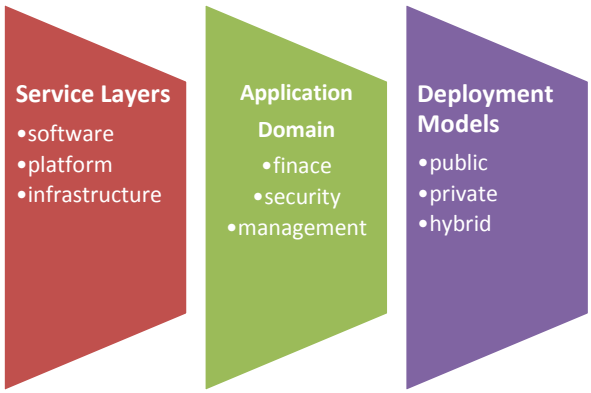



**Fig 4: Cloud Architecture**

*3.1 Security of Cloud*

The various security issues with respect to cloud are :

- Storage Security
- Middleware security
- Data security
- Network security

- Application security

Another aspect of security focuses on virtualization. Due to the complex nature of cloud, it is very difficult to achieve end-to-end security in a cloud also the boundary in a cloud is identified to be fuzzy in nature.Apart from information assurance, it is aimed that a malicious user should be blocked from enteringthe system or if entered, should be immediately identified and countermeasure is taken against them.

A Cloud is an application platform that uses internet-based services to support business process or in other words, it provides a framework which can be used to rent IT-services on a utility-like basis. The key attributes of a cloud which makes it so popular are: the low startup costs, fast deployment, costs based on usage, and multi-tenant sharing of services. The essential characteristics of cloud are, on demand self-service, pervasive network access, location independent resource pooling, rapid elasticity, measured service.

## 4.DATA MINING

Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD) is a field of computer science, which involves discovering patterns from large data sets through methods of artificial intelligence, machine learning,statistics, and database systems. The main aim of the data mining process is to extract information from a data set and transform it into an understandable format for future use. Apart from basic analysis, the data mining process covers database and data management aspects, data preprocessing, inference considerations, complexity considerations, post-processing of discovered structures,and online updating.Roots of Data Mining are statistics,Artificial Intelligence & Machine Learning,Databases,Pattern discovery, visualization, business Intelligence etc. The various Data mining techniques are listed below in table 5

| Technique | Key Feature |
|---|---|
| Regression | Technique for predicting a continuous numerical outcome such a customer lifetime value, house value, process yield rates. |
| Attribute Importance | Ranks attributes according to strength of relationship with target attribute. Use cases include finding factors most associated with customers who respond to an offer, factors most associated with healthy patients. |
| Anomaly Detection | Identifies unusual or suspicious cases based on deviation from the norm. Common examples include health care fraud, expense report fraud, and tax compliance. |
| Feature Extraction | Produces new attributes as linear combination of existing attributes. Applicable for text data, latent semantic analysis, data compression, data decomposition and projection, and pattern recognition. |

| Technique | Key Feature |
|---|---|
| Clustering | Useful for exploring data and finding natural groupings. Members of a cluster are more like each other than they are like members of a different cluster. Common examples include finding new customer segments and life sciences discovery. |
| Classification | Most commonly used technique for predicting a specific outcome such as response / no-response, high / medium /low value customer, likely to buy / not buy. |
| Association | Find rules associated with frequently co-occurring items, used for market basket analysis, cross-sell, root cause analysis. Useful for product bundling, in-store placement, and defect analysis. |

Fig 5: Data Mining Techniques

*4.1ROLE OF DATA MINING IN INFORMATION SECURITY*

Data mining is extraction of hidden, useful and precious information from large databases Data mining came into being with an objective to support large databases that are used in various business applications for predicting future trends,
analyzing data and making proactive decisions. Data mining has emerged as a tool that provides its users to identify the vulnerabilities and helps in providing a defensive mechanism against a number of threats to the information systems. There are various applications of data mining in the area of information security.

Commonly discussed domain in the field of information security is intrusion detection where the threats to the system are identified and prevented. Good amount of work has been done in this area by the researchers and various data mining techniques have been applied for detection and prevention of security attacks on the system.With the advancements in the area of information security, the applications of data mining has also increased immensely to various other areas of information security and are not restricted to just intrusion detection andprevention systems.Network intrusion detection is another area which requires immediate attentions,as the number of intrusion attacks are increasing.It is a unique form of computer-generated threat analysis to identify nasty actions that could compromise the integrity, confidentiality,and availability of information resources.Intrusion detection mechanisms based on data mining are extremely useful in discovering security breaches.In literature, a number of data mining based algorithms have been proposed to deal with the information security and privacy problems, by using approaches like classification,frequent pattern mining, and clustering methods to do intrusion detection, anomaly detection, and privacy preserving.Application of these data mining methods haveresulted in stimulating results that has concerned many researchers in both data mining and information security areas.

Table2 lists the various data mining algorithms that have been used for detection and avoidance of different information security attacks like intrusion detection,fraud detection, etc.

As mentioned in table 2, the intrusion can be identified as host based or network based. Some of ways to detect an intrusion on a computer, network, or a cloud is detecting an anomaly or finding misuse of the services or resources. Similarly frauds can be detected by outliers and self organizing maps which involves unsupervised learning. One of the ways to detect loopholes in privacy preserving is K-Anonymity method wherein identity disclosure is detected. Buffer overflow can result in information leakage whereas denial of service attacks can result due inability to differentiate the valid user request from the multiple invalid ones.

| Area | Types | Detection |
|---|---|---|
| Intrusion Detection | • Network Based<br>• Host Based | • Anomaly ID<br>• Misuse ID<br>• Data mining Based **Avoidance**<br>• Data fusion based<br>• Immunological Approach based |
| Fraud Detection | • Management Fraud<br>• Customer Fraud<br>• Network Fraud<br>• Computer Based Fraud | • Outlier detection<br>• Self Organizing Maps |
| Privacy Preserving | • Data Privacy<br>• User Privacy | • K-Anonymity (Identity disclosure)<br>• Perturbation Approach<br>• Cryptography<br>• Randomized Response<br>• Condensation Approach |
| Detecting Information Leakage | • Buffer Overflow attack<br>• Data Mining | • Brute Force method<br>• Exploratory data analysis **Avoidance**<br>• Legitimacy tags<br>• External Leakage |
| Firewall | • Basic<br>• Distributed Network | • Anomaly Detection<br>• Generalization<br>• Association rule mining<br>• Frequency based technique |
| Data Security Enhancement | • Multi-level Security model<br>• Encryption-Blind signatures<br>• Biometric encryption<br>• Anonymous databases | NA |

**Table 2.Data Mining Techniques for Information Security**

*4.2 PRIVACY PRESERVING THROUGH DATA MININIG:*

Though, data mining also poses a risk to privacy and information protection if not done or used properly.For example, association rule analysis is an accepted tool for discovering useful associations from huge amount of data and some valuable hidden information could be simply discovered using this sort of tool. Hence, the security of sensitive hidden information has become a significant issue to be resolved. The aim of privacy preserving data mining is to hide certain information so that they cannot be exposed through data mining techniques such as association rule analysis. There have been two significant approaches
for privacy preserving data mining are: output and input privacy.
The output privacy approach is to modify the data before delivery to the data miner so that real data is hidden and mining result will not reveal certain privacy. For example, blocking, merging, swapping and sampling are some methods that have been proposed for this type of output privacy . The input privacy approach, on the other hand, is to change the data using data distribution methods. In this approach, mining result is not affected or minimally affected.For example,reconstruction based and cryptography based are some techniques that have been proposed for this type of input privacy.

Data mining has also emerged as a way for identifying patterns and trends from large quantities of data.For example, shopping centres found out that male customers who buy diaper usually buy beers by analyzing consuming lists. This forms the relation between diaper and beer through rearranging these goods.This improvement of goods arrangement after analysis yields more sale. This kind of analysis can be used in many fields such as Credit Cards, Banking sectors, etc. Hence, techniques of data mining without leaking the private information are needed. Research on privacy preserving data mining is developed for
this purpose.The privacy preserving data mining and knowledge discovery should be developed aiming at these problems. In order to secure an openly availablesystem , it must be ensured that not only that private
sensitive data should be trimmed out, but also to make sure that certain inference channels should also be blocked as well. Under privacy constraints, the association rule mining problem was extensively researched.Many efficient methods for privacy preserving association rule mining were found. However,most of these methods resulted in information loss and side-effects to some extent, such as non-sensitive rules falsely hidden and spurious rules falsely generated, may be formed in the sensitive rule hiding process.

Sequential pattern mining can be defined as finding the complete set of frequent subsequences in a set of sequences.Sequential pattern mining can be used for discovering meaningful sequential patterns among a large quantity of data. For example, let us see the sales database of a bookstore. The revealed sequential pattern could be “70% of people who bought Twilight also bought Harry Potter at a later time”.The bookstore can make use of this information for shelf placement, promotions, etc.

## 5. CONCLUSION & OPEN ISSUES

This paper provides the review of literature on how data mining techniques and related algorithms can play a vital role in ensuring information security in a cloud. With the growing dependence of humans on machines, it is required to create a better framework to provide a secure electronic-infrastructure to work upon and ensure information security.Cloud proposes services on demand at a much affordable rate with minimum overheads thereby increasing the popularity of cloud. At the same time issues of information security becomes critical like only an authorized user should be allowed to use the services of a cloud. Therefore, need of the hour is to implement information security in such a manner that the valid users get the maximum availability of services and the invalid ones be identified, and stopped from misusing and disrupting the services. Data mining algorithms provide a solution to this challenge of detecting and avoiding the information security attacks like intrusion, fraud, information leakage, etc. This paper gives a review of various data mining approaches which can protect a cloud from different information security attacks.

With the help of literature review, a number of open issues have been identified and listed in table 3.

**Table3.Open issues**

| Cloud Security Area | Future Challenge |
|---|---|
| Intrusion Detection | □ Reduce number of false negatives<br>□ Anomaly detection (Malicious user/code) |
| Privacy Preserving | □ Homogeneity attack<br>□ Background knowledge attack<br>□ Personalized privacy preserving (ARM) |
| Mobile Security | □ Biometrics<br>□ Authentication |
| Firewall | □ Multiple firewalls for distributed networks<br>□ Application layer feedback based approach for spam detection<br>□ Handling massive log data<br>□ Analysis of Network traffic<br>□ Detecting faulty and leaky network |
| General | □ Authentication<br>□ DOS Attacks |

## 6.REFERENCES

[1] Dharminder Kumar and Deepak Bhardwaj,"Rise of Data Mining: Current and Future Application Areas", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 5, No 1, September 2011.

[2] S. Mitra, S. K. Pal, and P. Mitra, "Data mining in soft Computing framework: A survey", IEEE Trans. Neural Networks, vol. 13, pp. 3 - 14, 2006.

[3] Han, J. and Kamber, M., "Data mining: Concepts and Techniques", Morgan-Kaufman Series of Data Management Systems. San Diego: Academic Press, 2011.

[4] Amanpreet Chauhan, Gaurav Mishra, and Gulshan Kumar, "Survey on Data Mining Techniques in Intrusion Detection", International Journal of Scientific & Engineering Research Volume 2, Issue 7, July-2011.

[5] Jose F. Nieves, "Data Clustering for Anomaly Detection in Network Intrusion Detection", 2009.

[6] Dimitrios Zissis and Dimitrios Lekkas, "Addressing cloud computing security issues", Department of Product and Systems Design Engineering, University of the Aegean, Syros 84100, Greece, Future Generation Computer Systems 28 (2012) 583–592.

[7] Mohamed Hamdi, "Security of Cloud Computing, Storage, and Networking", School of Communication Engineering, Technopark El Ghazala, 2083 Tunisia, IEEE, 2012.

[8] Albert Greenberg, James Hamilton, David A. Maltz and Parveen Pate, "The Cost of a Cloud: Research Problems in Data Center Networks", Microsoft Research, Redmond,WA, USA.