Application of Data Mining Techniques for Information Security in a Cloud: A Survey

Preeti Aggarwal
CS/IT, KIIT College of Engineering
Gurgaon, India

M. M. Chaturvedi
SET, Ansal University
Sector-55, Gurgaon

ABSTRACT

India is progressively moving ahead in the field of Information technology. The concept of e-commerce is already in place whereas e-governance is also on the same track. Similarly other sectors like health, judiciaries etc. are following the path. With the advent of information technology, malevolent people now have another option to cause damage to people by doing cyber attacks rather than physical damage, wherein the impact of cyber damage is equally devastating. 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 hour 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 task of 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 of information on cloud.

Keywords

Cloud, Data Mining, Intrusion Detection, Information Security

1. INTRODUCTION

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. Cyber security also referred to as information technology security, whose main focus is protection of computers, networks, programs and data from unauthorized access, change or destruction. Since the internet access is getting cheaper people are always connected to the internet via computer or mobile phones. To protect the information exchanged over internet, cyber security standards are required. Cyber security standards are security standards which enable organizations to practice safe security techniques to minimize the number of successful cyber security attacks. In the current scenario cyber attacks and digital spying are identified as the biggest threat to any nation.

The growth of cloud computing as a service on-demand is leading to a new requirement for its sustenance that is, its security. The cloud security has become a key area of research; as a result a new dimension is added to the field of information security. Cyber security plays a major role in cloud security because most of the cloud services are accessed through the cyber interface. Also the security of data while it is being exchanged between the hosts in cloud is an area of concern. Further, the application interface through which cloud services are received has to be robust enough to ensure data security. In this paper in depth analysis of various types of information security requirements, data security attacks, cloud security requirements, types of security vulnerabilities on cloud and comparative study of various data mining techniques that can help fight information security loop holes on a cloud is done. This paper is structures in nine sections; Section 1 covers the basic introduction whereas section 2 presents the review of literature. Section 3 gives characteristics of information security, types of attacks and risks. Section 4 discusses the cloud as a service, its architecture, and applications of cloud, security requirements and vulnerabilities. Section 5 describes the various possible data mining techniques available with their application domains with respect to information security. Section 6 covers the role of data mining techniques and algorithms to provide information security and the conclusion is drawn in the section 7 with the open issues tabulated as well.

2. LITERATURE SURVEY

Security of Information has always been crucial for the sustenance of future development since early days. Earlier information used to be gathered manually and proper means of preserving this information were not available. With the advancement of technology, there has been vast growth in the ways of preserving the information but simultaneously security is becoming a major concern due to the various security threats.

These information security issues may arise in a desktop computer, office environment, on a network or in a cloud. The literature review shows that data mining is key ingredient in the solution to information security problems. The author in [1] discusses the development of data mining and its application areas. Soft computing framework data mining is presented in paper [2] where soft computing approaches like fuzzy logic, neural network are discussed. Data mining provides a number of algorithms that can help detect and avoid security attacks [3].The author in [4] presents a survey on various data mining techniques for intrusion detection wherein the types of intrusion attacks like network and host based are also summarized. One of the intrusion detection technique known as anomaly detection has been discussed in detail [5]. Paper [6] specifies the measurement criterions for intrusion detection. Fraud detection is another area of focus as the number of online transactions is rising exponentially. Various types of frauds like computer fraud are given in [7] with the respective techniques to overcome the situation. A number of methods are proposed for privacy preserving through data mining in [8], for example K-Anonymity. In paper [9], author talks about the sensitivity of data which may risk an individual’s privacy. This data can be general data, user specific or authentication data. PETRE in [10] specifies aspects of cloud computing and the top cloud computing companies with their respective key features. The cloud security issues have been addressed via a trusted third party in [11]. Data mining techniques can also be used for the analysis of various firewall policy rules [12]. Security framework for mobile cloud computing is proposed in [13]. In [14], the authors have identified the following types of attacks which are a major threat to cloud implementation denial of service attack, Cross virtual machine side-channel attack, malicious insiders' attack, Attacks targeting shared memory, and Phishing attack. Table 1 briefs the review of variety of work done in the area cloud computing security with the help of data mining techniques. Paper [15] details the need of mobile cloud computing. As the mobiles are getting cheaper with the availability of internet facility, a mobile can also be considered as an entity in a cloud.

Table 1.Review of Cloud Security Approaches

Area	**Year**	**First Author**	**Work**
Cloud Computing	2012[6]	Dimitrios Zissis	Cloud design principles to control security threats
	2012[21]	Arjun Kumar	encryption and compression of data using secret key at the main server while uploading to the Cloud Storage servers.
	2012 [14]	Md. Tanzim Khorshed	Types of attacks on cloud, SVM
	2011[17]	Pardeep Kumar	Security in cloud using HMM model
	2010[22]	Zhidong Shen	the function of trusted computing platform in cloud computing.
	2010[26]	Aman Bakshi	Eradication of the DOS attacks using IDS over the cloud
	2010[18]	Qian Tao	trustworthy management approach for cloud to get rid of the influence of malicious attacks
Mobile cloud computing	2013[15]	Niroshinie Fernando	Motivation for Mobile cloud computing
	2012[13]	Abdul Nasir Khana	Security framework for Mobile cloud computing
Privacy Preserving	2013[19]	Mohammad Farhatullah	Integrate privacy preservation with pattern recognition approaches to privacy leak detection in the context of text mining
	2009[25]	Jian Wang	k-anonymity -allows attacks due to lack of diversity and l-diversity, a framework that gives stronger privacy guarantees
	2008[24]	Saeed Samet	protocols for perceptron learning algorithm in multi-party environment to preserve the privacy of the output model as well as the input data

Fraud Detection	2013[16]	Saman Zonouz	Cloud based service for intrusion tolerance
	2010[27]	Shiguo wang	categorizes, compares, and summarizes the algorithm and performance for fraud detection
	2008[23]	P Jayashree	double-filtering mechanism with efficient usage of Leaky Buckets for accurate detection of attack packets

3. 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.

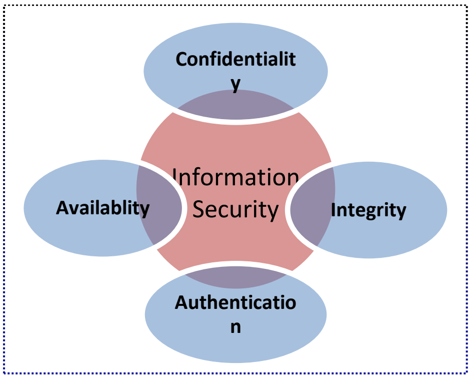

Fig 1: Information Security Attributes

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:

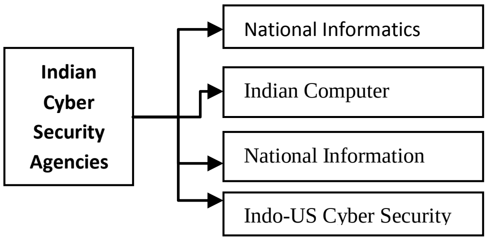

Fig 2: Indian Cyber Security Agencies

3.1 Types of Attacks

Cyber crime is spread over the complete cyber space which is defined as a network that includes the Internet as a major component. 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 of information from a system but does not attempt to alter the system, its resources, its data or its operations.

3.2 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:

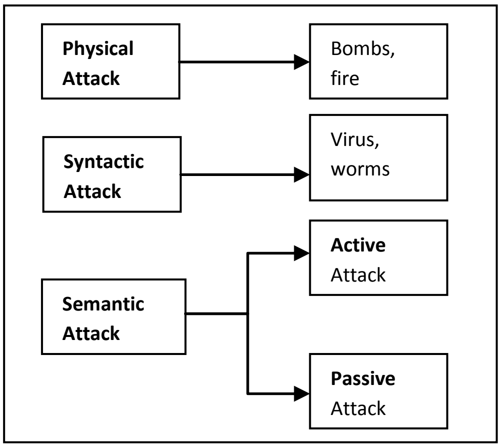

Fig 3: Types of Attacks

3.3 Cyber Security

One of the essential requirements of cyber security is to provide information security whose key attributes are confidentiality, data integrity, authentication and data availability. Cryptography is one of the most common techniques used to provide security services. The first step in the direction of protecting a computer or network is to recognize the risks and become familiar with the terminologies associated with them. For example, a list of entities or hosts that are blocked or denied privileges or access (Blacklisted) can be identified. Similarly, lists of entities that are considered trustworthy and are granted access or privileges are called White-listed. There are basic utilities available with the help of which the cyber attacks can be detected:

Cryptography-The information is protected by converting it into an unreadable format (cipher text). This message can be deciphered by only those who possess the secret key.

Intrusion Detection −The method of analyzing information from networks and information systems to determine if a security breach or security violation has occurred.

Penetration Testing −An evaluation methodology whereby assessors search for vulnerabilities of a network or information system.

4. 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.
- Community cloud: The cloud infrastructure is shared by several organizations and supports a specific community that has communal concerns (e.g., security requirements, policy, and compliance considerations). It is again managed by a third party or an organization 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.

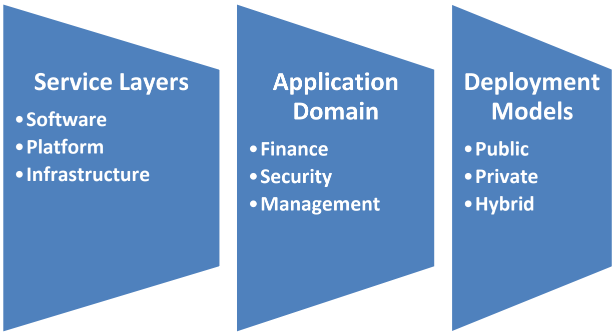

Fig 4: Cloud Architecture

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.

4.1 Security of Cloud

The various security issues with respect to cloud are [14]:

- 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 [17]. Apart from information assurance, it is aimed that a malicious user should be blocked from entering the 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 [17] are, on demand self-service, pervasive network access, location independent resource pooling, rapid elasticity, measured service.

5. DATA MINING

Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD) [3], 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 [2] are statistics, Artificial Intelligence & Machine Learning, Databases, Pattern discovery, visualization, business Intelligence etc. The various Data mining techniques are listed in figure 5:-

- Clustering – It is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.
- Classification – It is the task of generalizing known structure which can be applied to new data. For example, an email program might attempt to classify an email as genuine or spam. Regular algorithms are decision tree learning, Naive Bayesian classification, neural networks (soft computing) and support vector machines.
- Regression - Attempts to find a function which models the data with the least error.
- Association Rule Learning - Searches for relationships between variables.

6. ROLE OF DATA MINING IN INFORMATION SECURITY

Data mining is extraction of hidden, useful and precious information from large databases [1]. 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.

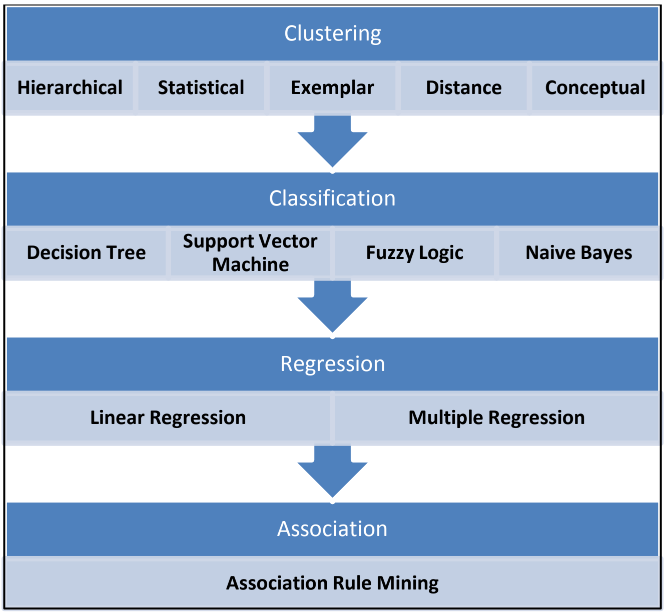

Fig 5: Data Mining Techniques

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 and prevention 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 [4]. Application of these data mining methods have resulted in stimulating results that has concerned many researchers in both data mining and information security areas. Table 2 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.

Table 2. Data Mining Techniques for Information Security

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

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.

7. 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 below in table 3. Some issues are related to development of new algorithm or approach to solve the security problem whereas others involve enhancement of a method to overcome certain limitations.

Table 3.Open Issues

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

8. 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.

[9] Tharam Dillon, “Cloud Computing: Issues and Challenges Digital Ecosystems and Business Intelligence”, Institute Curtin University of Technology Perth, Australia, 2010 24th IEEE International Conference on Advanced Information Networking and Applications.

[10] Piatetsky-Shapiro and Gregory, “The Data-Mining Industry Coming of Age”, in IEEE Intelligent Systems, vol. 14, issue 6, Nov 1999.

[11] Hsu J., “Data Mining Trends and Developments: The Key Data Mining Technologies and Applications for the 21st Century”, in the Proceedings of the 19th Annual Conference for Information Systems Educators.

[12] Korosh Golnabi, Richard K. Min, Latifur Khan, and Ehab Al-Shaer, “Analysis of Firewall Policy Rules Using Data Mining Techniques”, IEEE, 2006.

[13] Abdul Nasir Khan, M.L. Mat Kiah, Samee U. Khan and Sajjad A. Madani, “Towards secure mobile cloud computing: A survey”, Elsevier B.V, 2012.

[14] Md. Tanzim Khorshed, A.B.M. Shawkat Al, and Saleh A. Wasimi, “A survey on gaps, threat remediation challenges and some thoughts for proactive attack detection in cloud computing”, Elsevier B.V, 2012.

[15] Niroshinie Fernando, Seng W. Loke, and Wenny Rahayu, “Mobile cloud computing: A survey”, Elsevier B.V, 2013.

[16] Saman Zonouz, Amir Houmansadra, Robin Berthiera, Nikita Borisov, and William Sanders, “Secloud: A cloud-based comprehensive and lightweight security solution for smartphones”, Computers & Security, 2013.

[17] Pardeep Kumar, Nitin, Vivek Sehgal, Kinjal Shah, Shiv Shankar Prasad Shukla, and Durg Singh Chauhan. “A Novel Approach for Security in Cloud Computing using Hidden Markov Model and Clustering”, 2011 World Congress on Information and Communication Technologies, IEEE, 2011.

[18] Qian Tao, Huiyou Chang, Yang Yi, and Chunqin Gu, “A Trustworthy Management Approach For Cloud Services QOS Data”, Proceedings of the Ninth International Conference on Machine Learning and Cybernetics, Qingdao, IEEE, 2010.

[19] Mohammad Farhatullah, “ALP: An Authentication and Leak Prediction Model for Cloud Computing Privacy”, 3rd IEEE International Advance Computing Conference (IACC), 2013.

[20] P. Mell and T. Grance, “Draft nist working definition of cloud computing - v15”, 21. Aug 2009.

[21] Arjun Kumar, HoonJae Lee, and Rajeev Pratap Singh, “Efficient and Secure Cloud Storage for Handling Big”, Data, Information Science and Service Science and Data Mining (ISSDM), 2012.

[22] Zhidong Shen and Qiang Tong, “The Security of Cloud Computing System enabled by Trusted Computing Technology”, 2010 2nd International Conference on Signal Processing Systems (ICSPS), IEEE, 2010

[23] P Jayashree, K.S.Easwarakumar, Anandharaman V, Aswin K, and Raja Vijay S, “A Proactive Statistical Defense Solution for DDOS Attacks in Active Networks”, First International Conference on Emerging Trends in Engineering and Technology, IEEE, 2008.

[24] Saeed Samet and Ali Miri, “Privacy-Preserving Protocols for Perceptron Learning Algorithm in Neural Networks”, 2008 4th International IEEE Conference "Intelligent Systems", IEEE, 2008.

[25] Jian Wang, Yongcheng Luo, Shuo Jiang, and Jiajin Le, “A Survey on Anonymity-based Privacy Preserving”, IEEE, 2009.

[26] Aman Bakshi and Yogesh B, “Securing cloud from DDOS Attacks using Intrusion Detection System in Virtual Machine”, 2010 Second International Conference on Communication Software and Networks, IEEE, 2010.

[27] Shiguo wang, “A Comprehensive Survey of Data Mining-based Accounting-Fraud Detection Research”, 2010 International Conference on Intelligent Computation Technology and Automation, IEEE, 2010.

9. AUTHOR’S PROFILE

Preeti Aggarwal holds M.Tech in IT from GGSIPU, New Delhi, M.Sc in Informatics and B.Sc (H) in Electronics from University Of Delhi and is currently pursuing Ph.D. in the area of Information Security in cloud through Data Mining techniques from Ansal university, Gurgaon. She is working as an Assistant Professor in department of Computer Science Engineering in KIIT College of Engineering, Gurgaon. She is also a life time member of Computer Society of India.

Manmohan Chaturvedi is a retired Air Commodore from Indian Air Force with PhD in Information Security domain from IIT Delhi. He has about 35 years of experience in managing technology for IAF. An alumnus of National Defense College, New Delhi, he has held various appointments dealing with operational and policy dimensions of Information and Communication Technology. He graduated from Delhi College of Engineering and completed post graduation from IIT Delhi. Currently he is a Professor at School of Engineering and Technology, Ansal University with research interests in vulnerability of evolving ICT infrastructure and protection of Critical Information Infrastructure.