Data Mining For Web Vulnerability Detection

Vaibhavi Sakurkar[1], Neha Khare[2]
[1, 2] Dept of Computer Science
[1, 2] Takshshila Institute of Engineering & Technology, Jabalpur MP, India

Abstract- *Developing software applications that are free of vulnerabilities is a necessity, especially if these applications are intended to operate over the World Wide Web. Web applications security is a major concern for different organizations and banking areas in today's era. Most organizations and banking areas that use the Internet to provide web-based services protect their sensitive data using firewalls and some access control mechanisms. However, the data of organizations is still revealed by Internet hackers aimed at purposefully designed SQL queries. Therefore, existing security mechanisms are not enough to provide effective security to web databases. In such a scenario, it is necessary to provide additional security mechanisms to safeguard important information received by hackers carefully prepared by SQL queries. This literature review is to analyze the latest development in the field of web security mechanisms to prevent SQL injection and XSS attacks. This analysis has been used to prepare new ways to prevent various types of attacks in web applications.*

Keywords- SQL, Artificial Intelligence, KDD, Database Management System.

I. INTRODUCTION

A wide variation of data mining and machine learning techniques are always trying to improve the ability to predict web application vulnerabilities. For instance, feature extraction and classification are used to predict if SQL (Structured Query Language) injection vulnerability resided in the software or not .Additionally, machine learning methods are used to increase the ability to cover a wide range of malicious web code. It represents a critical review of data mining applications in vulnerability detection. It discusses the techniques were used, the importance of using these techniques and the results of using these techniques. The papers also set the stage for future directions and improvements [15].

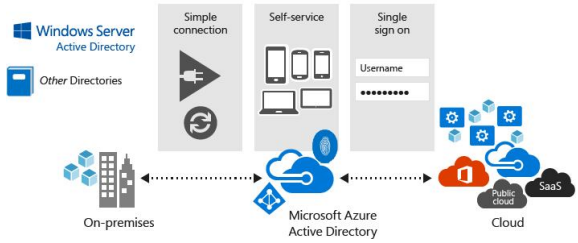


Fig-1: Structured Query Language

Attackers have an over-growing list of vulnerabilities to exploit in order to maliciously gain access to our web applications, networks and servers. Regardless of whether we're a beginner client or utilize a sophisticated hosting service, in the event that really decided, at that point an attacker will discover any vulnerability we've failed to patch and utilize it to their focal points. New vulnerabilities are being found constantly; by security specialists, by attackers, even by clients. Each time changes are made at any level of the infrastructure, there's the potential for new vulnerabilities to be made [12].

II. RULES OF DATA MINING

In data mining, we make some rules that the association rules speak. This rule comes in the work of analyzing data. Data analysis parameters include path analysis (i.e. understanding the path and extracting it about it), classification (splitting it into pieces), clustering (one place to add or fitting), and forecasting (also predicting) in the data parameter are there Path Analysis looks at the pattern so that it can work effectively [14].

III. DATABASE

A database is a composed gathering of information, by and large put away and got electronically from a computer networks. Where databases are increasingly mind boggling they are frequently created utilizing formal plan and demonstrating systems [9]. The Database Management System (DBMS) is the product that associates with end clients, applications, and the database itself to catch and break down the information. The DBMS programming also envelops the center offices gave to direct the database. The entirety of the database, the DBMS and the related applications can be alluded to as a "database networks ". Regularly the expression

"database" is likewise used to freely allude to any of the DBMS, the database networks or an application related with the database.

IV. KNOWLEDGE DISCOVERY IN DATABASES

The knowledge discovery in databases (KDD) process is normally characterized with the stages:

1) Selection
2) Pre-preparing
3) Transformation
4) Data mining
5) Interpretation/assessment

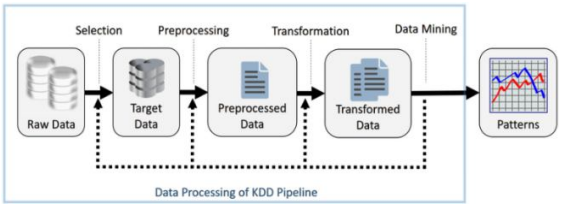


Fig-2: Stages of KDD Process

V. WORK DONE SO FAR

The term Data mining showed up around 1990 in the database network, by and large with positive undertones. For a brief span in 1980s, an expression "database mining" was utilized, however since it was trademarked by HNC, a San Diego-based organization, to pitch their Database Mining Workstation; scientists thus went to information mining [21]. In the scholastic network, the real gatherings for research began in 1995 when the First International Conference on Data Mining and Knowledge Discovery (KDD-95) was begun in Montreal under AAAI sponsorship [20]. After a year, in 1996, Usama Fayyad propelled the diary by Kluwer called Data Mining and Knowledge Discovery as its establishing editorial manager in-boss. The KDD International gathering turned into the essential most astounding quality meeting in Data mining with an acknowledgment rate of research paper entries underneath 18%. The diary Data Mining and Knowledge Discovery is the essential research diary of the field [21]. Sung Soo Kim et al et al [7, 2016] the authors researched and refined the Vulnerability Detection Tools (VDT) based upon open API utilized Open VAS for the prohibition of problems concerning security. Vulnerabilities are determined in advance by these tools. Administrators must safeguard their servers from the various attacks. Since detection methods of distinct tools are different. Due to these factors, it is suggested that instead of acquisition of results from single tool, results are gained from many tools and the installation desires considerable amount of overhead [11].

Lwin Khin Shar et al et al [3, 2018] the author outlines a practical method for the prediction of vulnerable code that prioritize various security balanced efforts. Both supervised and semi-supervised learning are carried out. On the labeled part data, that is fully and partially available. The supervised model attain an average of 5% false alarms and 77% recall rate where the semi-supervised model gain improvement of 27% and recall rate and 3% false alarm rate. To foresee the presence of false positives, we present the clever thought of surveying if the vulnerabilities distinguished are false positives utilizing information mining. To do this evaluation, we measure traits of the code that we saw to be related with the nearness of false positives, and utilize a blend of the three top-positioning classifiers to signal every one of the powerlessness as false positive or not WAP examines and expels input approval vulnerabilities from projects or content. Data mining allows a different approach [8]. Humans label samples of code as vulnerable or not, then machine learning techniques are used to configure the tool with knowledge acquired from the labeled samples. Data mining then uses that data to analyze the code.

VI. MOST COMMON WEB VULNERABILITIES

OWASP or Open Web Security Project is a non-benefit magnanimous association concentrated on improving the security of programming and web applications. The association distributes a rundown of top web security vulnerabilities dependent on the information from different security associations. The web security vulnerabilities are organized relying upon exploitability, perceptibility and effect on programming [22].

Table-1: Top 4 OWASP Security Risks

Security Risk	Exploit	Impact Ability	What the Attacker can do
Injection	Easy	Severe	Executing unintended commands or accessing data without proper authorization
Broken Authentication and Session Management	Average	Severe	Compromise passwords, keys, or session tokens
XSS	Average	Moderate	Hijack user sessions, deface web sites or redirect the user to malicious sites
Insecure Direct Object References	Easy	Moderate	Access of unauthorized data

The main factor of the methods that uses data mining techniques is to reach the highest accuracy and precision of the prediction method with the lowest false positive rate. So it is important to choose the classification method or a combination of classification methods and the appropriate features to be used in the classification process, to techniques, but if the static analysis techniques are combined with other approaches like data mining techniques, this will improve the accuracy and precision and will lower the false positive rates. The main factor of the methods that uses data mining techniques is to reach the highest accuracy and precision of the prediction method with the lowest false positive rate. So it is important to choose the classification method or a combination of classification methods and the appropriate features to be used in the classification process, to fulfill factor [13].

VII. PROPOSED SOLUTION

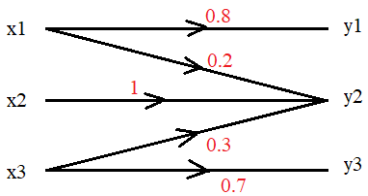
The data mining techniques have improved the quality of vulnerability prediction when combined with static analysis tools. Using data mining techniques decrease the value of false positive, which was weakness point in static analysis tools. Here a different classification algorithm has been used; some of them gave better results compared with others. This indicates that the selection of the classifier plays an important role in improving the results. The word Probability gets from the Latin Probabilitas, which can likewise signify "honor", a proportion of the expert of an observer in a lawful case in Europe, and frequently connected with the observer's respectability. Probability is the proportion of the probability that an occasion will happen. Probability evaluates as a number somewhere in the range of 0 and 1, where, freely, 0 demonstrates difficulty and 1 shows sureness. The higher the likelihood of an occasion, the almost certain it is that the occasion will happen.

In Probability hypothesis, Conditional Probability is a proportion of the Probability of an occasion (some specific circumstance happening) given that another occasion has happened. On the off chance that the occasion of intrigue is An and the occasion B is known or accepted to have happened, "the Conditional Probability of A given B", or "the Probability of An under the condition B", is typically composed as P(A | B), or once in a while P(A/B).
Let,

P (A) = Un- Conditional Probability
P (A/B) = Conditional Probability
P (A/B) might possibly be equivalent to P (A).

A Binary Symmetric Channel has 2 inputs 0 and 1 (x1=0, x2=1) and 2 outputs 0 and 1 (y1=0, y2=1). This channel is symmetric because the Probability of Receiving 1 if 0 is sent is equal to the Probability of Receiving 0 if 1 is sent. This Transmission Probability is denoted by P. Let, there are three web sites URL (Uniform Resource Locator) x1, x2 and x3. Through a single click there may be three outputs on the screen y1, y2 and y3. If y1 and y3 are the two Mirror web sites with two separate links of different URLs. It means y1 and y3 have different Databases but both have the same information. If a Client would like to access y1 then the Probability of accessing y1 is denoted by q and the Probability of intentionally provided the web site with Vulnerability y2 is p. Suppose the Probability of generating the desired web site output y1 from the URL x1 is q=0.8. The Probability of generating the undesired web site output y2 from the URL x1 is p=0.2. x1 is the desired web site URL which has been entered by a client. x2 is the undesired web site URL while x3 is the another Mirror web site URL. Usually by entering URL x1, y1 must be accessed, by entering URL x2, y2 must be accessed while by entering URL x3, y3 must be accessed.

The Probability of generating the second Mirror web site output y3 from the URL x3 is denoted by q and the Probability of intentionally provided the wrong web site y2 due to Vulnerability is p. Suppose the Probability of generating the second Mirror web site output y3 from the URL x3 is q=0.7. The Probability of generating the web site with Vulnerability output y2 from the URL x3 is p=0.3. x3 is the desired web site URL which has been entered by a client.



$$P(Y/X) = \begin{matrix} 0.8 & 0.2 & 0 \\ 0 & 1 & 0 \\ 0 & 0.3 & 0.7 \end{matrix}$$

$$P(X/Y) = \begin{matrix} 1 & 0.1091 & 0 \\ 0 & 0.7273 & 0 \\ 0 & 0.3818 & 0.4286 \end{matrix}$$

Transition or Conditional Probability P(X/Y) is defined as the measurement, what is the Probability through which output Y may be generated from Input X.

P[x1,x2,x3) / y1] = 1+0.1091+0 = 1.1091
P[x1,x2,x3) / y2] = 0+0.7273+0 = 0.7273
P[x1,x2,x3) / y3] = 0+0.3818+0.4286 = 0.8104
Where,

Probability of generating y1 from either x1 or x2 or x3 P[x1,x2,x3) / y1] = 1.1091
Probability of generating y2 from either x1 or x2 or x3 P[x1,x2,x3) / y2] = 0.7273
Probability of generating y3 from either x1 or x2 or x3 P[x1,x2,x3) / y3] = 0.8104
Minimum Value = 0.7273
Maximum Value = 1.1091

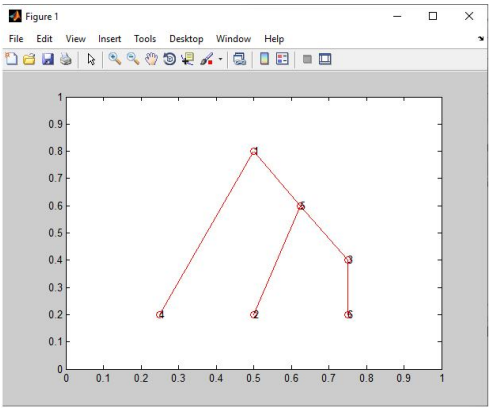


Fig-3: Graphical representation of Conditional Probabilities

VIII. RESULTS AND DISCUSSION

From the above equation, it is clear that the Probability of providing the undesired web site y2 from all the URLs x1, x2 and x3 is very high. However if we calculate the Output Transition or Conditional Probability P(X/Y) then it has been observed that P[x1,x2,x3) / y2] is the lowest value always. Similarly for n number of web sites the lowest value indicates always the web site with Vulnerability. It may be harmful from the point of view of Security or Privacy. It must be blocked.

$$\text{Vulnerability Detection Rate} = \frac{1.1091+0.8104}{1.1091+0.8104+0.7273} * 100\%$$

$$\text{Vulnerability Detection Rate} = \frac{2.403}{3} * 100\%$$

Vulnerability Detection Rate = 80.0995 %

REFERENCES

[1] Danjun Liu and Jingyuan Wang, "Pangr: A Behavior-based Automatic Vulnerability Detection and Exploitation Framework", 2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/ 12th IEEE International Conference On Big Data Science And Engineering 2018.

[2] Nilambari Chhagan Sonawane, "Data Mining Based Web Vulnerability Scanner", International Journal of Innovative Research in Science, Engineering and Technology, Vol. 7, Issue 12, December 2018.

[3] Julian Thome, Lwin Khin Shar and Domenico Bianculli, "An Integrated Approach for Effective Injection Vulnerability Analysis of Web Applications through Security Slicing and Hybrid Constraint Solving", IEEE Transactions on Software Engineering 2018.

[4] Mohamed Saibudeen B and Ganesh Kumar S, "Detecting and Eliminating Web Application Vulnerabilities with Data Mining", International Journal of Pure and Applied Mathematics, Volume 119 No. 7 2018.

[5] Duha A. Al-Darras and Ja'far Alqatawna, "Data Mining for Web Vulnerability Detection: A Critical Review", 2017 8th International Conference on Information Technology (ICIT) 2017.

[6] Anatoliy Gorbenko, Alexander Romanovsky, Olga Tarasyuk and Olga Tarasyuk, "Olga Tarasyuk", 2017 IEEE 28th International Symposium on Software Reliability Engineering 2017.

[7] Sung Soo Kim and Da Eun Lee, "Vulnerability Detection Mechanism based on open API for multi user's convenience", ieee 2016.

[8] Yaohui Wang,Dan Wang and Wenbing Zhao,Yuan Liu, "Detecting SQL Vulnerability Attack based on the Dynamic and Static Analysis Technology", 2015 IEEE.

[9] Julian Thome, Lwin Khin Shar and Domenico Bianculli, "An Integrated Approach for Effective Injection Vulnerability Analysis of Web Applications through Security Slicing and Hybrid Constraint Solving", IEEE Transactions on Software Engineering, 2018.

[10] Ibéria Medeiros, Nuno F. Neves and Miguel Correia, "Automatic Detection and Correction of Web Application Vulnerabilities using Data Mining to Predict False Positives", ieee 2014.

[11] Sanaz Rahimi and Mehdi Zargham, "Vulnerability Scrying Method for Software Vulnerability Discovery Prediction Without a Vulnerability Database", IEEE Transactions On Reliability, Vol. 62, NO. 2, JUNE 2013.

[12] Jingzheng Wu, Yanjun Wu, Zhifei Wu, Mutian Yang and Yongji Wang, "Vulcloud: Scalable and Hybrid Vulnerability Detection in Cloud Computing", 2013 Seventh International Conference on Software Security and Reliability Companion 2013.

[13] Peng Li and Baojiang Cui, "Comparative Study on Software Vulnerability Static Analysis Techniques and Tools", ieee 2010.

[14] Chris Tseng, Mohamed Ali and Rohan Vibhandik, "Common Visual Representation for websites and

smartphones", 2010 IEEE International Conference on Granular Computing 2010.
[15] Fang Yu, Muath Alkhalaf and Tevfik Bultan, "Generating Vulnerability Signatures for String Manipulating Programs Using Automata-based Forward and Backward Symbolic Analyses", 2009 IEEE/ACM International Conference on Automated Software Engineering 2009.
[16] Gary Wassermann and Zhendong Su, "Static Detection of Cross-Site Scripting Vulnerabilities", ICSE'08, May 10–18, 2008.
[17] Xiao-song zhang and lin shao, jiong zheng, "a novel method of software vulnerability detection based On fuzzing technique", ieee 2008.
[18] Hyunha Kim, Tae-Hyoung Choi, Seung-Cheol Jung, Hyoung-Cheol Kim, Oukseh Lee and Kyung-Goo Doh, "Applying Dataflow Analysis to Detecting Software Vulnerability", ICACT 2006.
[19] Nenad Jovanovic, Engin Kirda and Christopher Kruegel, "Preventing Cross Site Request Forgery Attacks", ieee 2006.
[20] Ahmed M. A. Haidar, Azah Mohamed and Aini Hussain, "Vulnerability Assessment of Power System Using Various Vulnerability Indices", 4"h Student Conference on Research and Development (SCOReD 2006), Shah Alam, Selangor, MALAYSIA, 27-28 June, 2006.
[21] https://en.wikipedia.org/wiki/Data_mining
[22] https://www.guru99.com/web-security-vulnerabilities.html
[23] https://www.shabdkosh.com/dictionary/english-hindi/vulnerability/vulnerability-meaning-in-hindi