Abstract

Web-based attacks pose a serious threat to the users of Web applications and sites. Attacks on Web applications and sites are devastating because the Web is an integral part of our lives as they provide several services to the community which are critical for their functioning. Virtuous users of an application are always of the notion that these services provided are reliable and safe to use. However, with rapid development in technology the threats that these applications face are also increasing immensely. Attackers use sophisticated techniques to find vulnerabilities in the applications and if successful they can gain unauthorized gains into the system, or steal valuable personal information of the users, or may even disrupt the services of the application. Taking into account the impacts that an attack may have on the functioning of a Web application, three kind of attacks are considered namely Cross-site Scripting attacks, HTTP Flooding attacks and Attacks in Critical Infrastructure. In Cross-site Scripting attacks, malicious content of the attacker is injected into a vulnerable Web application which is then unknowingly executed by the end-user. In HTTP Flooding attacks, the attacker tries to take down the services provided by an application by overwhelming the server's resources with legitimate HTTP requests. As a result, legitimate users of the application are denied the service. On the other hand, attacks on the Critical Infrastructure disrupt normal functioning of Cyber Physical Systems for a particular geographical area. Example of such systems are Water treatment plants, Gas storage plants, and Nuclear plants to name a few. Defense mechanisms to detect such Web-based attacks are often evaded by attackers by engineering their malicious payload very similar to the normal payload. The main goal of any detection mechanism is to differentiate the behavior between normal and malicious operations. Researchers use different approaches for effective detection of Web-based attacks, approaches such as data mining techniques, statistical techniques, machine learning approaches to name a few. In a detection system to counter the aforementioned attacks processing takes a significant amount of time as it has to analyze a huge amount of data. In recent times, machine learning techniques have proven to be efficacious in this regard and in such systems, preprocessing is a key factor. Using preprocessing steps such as feature selection the desired features can be selected to differentiate between normal and malicious instances.

In this thesis, firstly a dataset preparation pipeline for Cross-site Scripting attacks is introduced which employs a feature extraction method to extract features

from raw scripts and URLs. The pipeline consists of different stages, each consisting of modules performing a designated task. The output of the proposed pipeline is the dataset XSSD comprising of 6695 instances, 21 features and 2 class labels. Next, three feature selection methods are proposed namely, MICC-UD, INFS-MICC and FSRA for the detection of the selected Web-based attacks. These methods select a subset of features which help discriminate malicious and normal instances in case of each attack. Performance of each method is evaluated using several benchmark datasets.

For the detection of Cross-site Scripting attacks, MICC-UD is introduced which is a mutual information and correlation based method. MICC-UD, a traditional feature selection method selects a subset of highly relevant and irredundant ranked feature subset. A good feature subset is one in which the features are not only relevant to the target class but at the same time the features amongst themselves should be irredundant. Redundant features if present only leads to overfitting the learning model. In MICC-UD, the relevant features are selected based on feature-class mutual information and for irredundancy, feature-feature correlation is used. The main highlight of the proposed method is that it can handle multi-class data as well. MICC-UD is compared with three benchmark feature selection methods on 16 datasets (comprising both XSS attack and other security datasets) to demonstrate its effectiveness in detecting attacks.

Based on the foundations of MICC-UD, an incremental feature selection method named INFS-MICC is proposed for the detection of HTPP Flooding attacks in the application layer. Identifying a flooding attack is difficult because the characteristics are similar to normal traffic. At the same time, detecting an attack in near real time using low computational resources is of the essence. Considering these issues, the proposed method utilizes an incremental approach. It is incremental in the sense that it can handle added-in data without having to start the processing from scratch. To demonstrate the effectiveness of INFS-MICC it is evaluated with three well-known HTTP-Flooding datasets and results show that it can discriminate normal and malicious traffic efficiently.

Finally, an ensemble feature selection methods named FSRA is proposed to detect attacks in Critical Infrastructure. Ensemble learning relies on the fact that decision given by a group is better than the decision given by an individual. In the same manner, FSRA combines the ranks given by three individual feature selection methods to obtain a final list of ranked methods. FSRA ensures that the rank

given to a feature is preserved in the aggregation process. The method is evaluated with three critical infrastructure facility datasets. Performance comparison shows that FSRA performs well compared to the base feature selection methods.

Keywords — Web-based attacks, Cross-site Scripting, HTTP Flooding attacks, Critical Infrastructure, Feature selection, Ensemble, Relevance, Irredundant, Machine Learning.