Analysis of Machine Learning Techniques for Detection System for Web Applications Using Data Mining

Jugnesh Kumar1, S B Goyal2, Pradeep Bedi3, Sunil Kumar4 and Ashish Shrivastava5

1St. Andrews Institute of Technology and Management, Gurgaon, India
2Faculty of Information Technology, City University, Malaysia
3Department of CSE, Graphic Era Hill University, Dehradun, India
4Department of Mathematics, Graphic Era Hill University, Dehradun, India
5Department of Electrical Engineering, Manipal University, Jaipur, India

E-mail: drsbgoyal@gmail.com

Abstract. Security is a key problem to each computer and computer networks. Intrusion detection System (IDS) is one of the most important research problems in community safety. IDSs are advanced to stumble on each acknowledged and unknown assaults. IDS employs many methods to secure information systems and networks against community-based and host-based threats. IDS utilises different machine learning methods. This thesis analyses IDS machine-learning methods. It also discusses several similar research completed between 2000 and 2012 and specialises in engineering techniques. Linked experiments include used unmarried, hybrid, ensemble, baseline, and datasets.

1. Introduction
Within the data mining step, it is important to choose the correct approach to the task properly. Machine learning approaches are also used [1]. A big change has been that human beings prefer to improve their means of solving issues spontaneously for a long time. People learn from past errors and try to fix them by correcting them or searching for new approaches to the problem. Orthodox programming systems do not take into account the outcomes of their activities and cannot change their behaviour. This exact problem is discussed in the machine learning area and the development of computer programmes that can learn and thereby increase their output by collecting more data and experience [2]. A was the first scientist to establish a self-study programme. In 1952, Samuel created a programme that improved the number of games played when playing the game control. For instance, a computer programme that collects consumer data from an e-Commerce shop and generates better tailored advertising from these pieces of information can gain new knowledge and is nearly artificial intelligence [3]. Furthermore, machine learning systems are normally classified [4].

- The technique used by all conventional computer programmes is defined in rote learning. The application is not able to derive conclusions or transformations from the provided details and all of its information is applied directly by the programmer [5].
- Instruction learning Includes all programmes that turn knowledge from the input language into an internal language. While the programmer already understands how to convert efficiently, it requires
little inference on the computer’s side. This defines therefore a different level of learning compared to repetition learning [6].

- Learning from Training, as opposed to Learning from Training, aims to create new abilities that are nearly like current skills and thus quickly implemented by translating experience into experience. This method requires a dynamic information set to be able to generate mutations and combinations. It creates new features that the original computer programme does not realise and thus needs a lot of inference.

- The most efficient way to learn is nowadays to build computer programmes to totally unknown capacities, or to identify unknown structures and patterns in data. Learning from Examples is one of those learning methods which have become most popular. Learning from examples is a tool commonly used to forecast new data entry class marks utilising a dynamic set of recorded instances in classification and data mining. The suggested research problems will be answered in this paper with methods and algorithms in this category [7]

2. Intrusion Detection System (IDS)

Intrusion Detection System (IDS) are computing systems or equipment that robotize the route to search, dissect and classify the incidents occurring in a PC organisation. As the gravity of threats in the enterprise has significantly increased, intrusion sites are a critical extension of any of the associations’ security foundations. The intrusion position enables association to defend their systems against the dangers associated with improved network availability and data frameworks dependency. In terms of the level and scope of the existing network security threats, the inquiry for safety specialists should not decide whether interruption findings should be used but which interruption positions and capacities should be used. Interruptions are triggered by: attackers reaching the frameworks, allowed users of the frameworks who want to enhance additional benefits, authorised customers who misuse the benefits that they obtain [8].

Intrusion location frames (IDS) are used to interpret and redirect assaults from either an organisation or a host-based approach. In any case, these objects are searching for marks of attack (specific examples), which typically indicate malicious or uncertain intent. When an IDS looks in network traffic for these instances, it is network dependent (figure 1). At the moment that an IDS is looking for threats, they have been identified in log records. Multiple measurements were made in order to differentiate different kinds of interruptions to an organisation; however, the accuracy of their findings [9] was not heuristic. The precise feasibility of an organisation’s ability to differentiate malicious sources cannot be compensated for, until a quick assessment of the execution is available. Paxson's Bro, Leckie et al’s robabilistic approach and the consecutive research principle for control recognition are all three key methods we are considering [10].

![Figure 1. Overview of NIDS](image-url)
3. Literature Review

Machine learning and data mining are research territories of computer science whose fast improvement is because of the advances in data examination research. Creation of database industry and resulting consumer conditions for strategies to isolate significant knowledge from massive data stores. An immense measure of examination work has been done in the mixed media region, focusing on various parts of data investigation, for example, the catch, stockpiling, ordering, mining, and recovery of sight and sound huge data. Nonetheless, not many examination works gives a total review of the entire pine-line of the techniques utilized in machine learning and data mining in the exploration issues. we lead a far-reaching diagram of the best in class techniques, calculations machine learning and data mining in interactive media frameworks [11].

M. Nikhil Kumar, K.V.S. Koushik, K. John Sundar et al. provides a link with the composing audit of machine learning and data review procedures in the area of computerization. Documents covering each method were recognized, screened and compacted in consideration of the volume of references or the congruity of a new procedure. Since data is so important in the field of machine learning and data mining, some striking advanced instructional records used for computerized protection are seen, and a variety of suggestions are given on when a given approach should be used [12].

Bilal Ahmad, Wang Jian, and Zain Anwar Ali et al. in his research highlights advances in internet security analysis on the use of machine learning and data mining. We have context, excitement, challenge discussion and advocate the use of ML / DM in intrusion detection [13].

Anna L. Buczak, , and Erhan Guven, et al. describes a literature survey for cyberanalytics in support of machine learning ( ML) and data mining (DM) methods to detect the intrusion. Short ML / DM method descriptions are provided. Data representing each method have been identified, interpreted and summarised on the basis of the number of citations or the pertinence of an emerging method. Because the data in ML / DM approaches are so significant, some popular cyber data sets that are used in ML / DM are mentioned. The sophistication of ML / DM algorithms is tackled, problems for the use of ML / DM for cyber protection are explored and suggestions are presented on when using a given tool. [14].

JABEZ J Dr. B. Muthukumar et al. suggested a new method called outlier identification in which the neighboring Outlier factor (NOF) tests the anomaly data collection. Here, the trained model consists of large data systems with a distributed storage infrastructure to enhance intrusion detection system performance. The test results have demonstrated that the proposed methodology as easily detects abnormalities like all other strategies [15].

Seema Sharma, Jitendra Agrawal, Shikha Agarwal, Sanjeev Sharma et al. says that Data mining (DM) is the most common tool for knowledge acquisition exploration. Classification is a technique used to map data into the class and category. Healthcare, etc are adapted in many fields. This paper introduces the different grouping methods such as decision tree, vector support machine, nearest neighbour etc. This survey offers a comparative analysis of different algorithms for classification [16].

4. Proposed Methodology

Various baselines are used for confirmation and good for success assessment. It also reveals how much the computer can recognise assaults, and how often incorrect classifications can occur. Figure 2 indicates annual work on IDS baselines from 2000-2012.
5. Data sets
The sets of data used for classification tasks include DARPA1998, DARPA1999 and KDD99. The mainly used data set is KDD99. There is a lot of DARPA backgrounds, such as normal attacks are not realistic, false alarm behaviour. The dataset KDD99 has the same drawbacks and is inherited from DARPA. There are also re-validated. Many individuals worked on multiple classifier data sets. The work on data sets made between 2000 and 2012 as seen in Figure 3 for each year. These datasets are used publicly and are accepted as standard IDS datasets. Data collection from 2000 to 2012 in year-round work is shown in figure 3.

6. Result Analysis
The review of the numerous papers written for IDS about time on machine learning techniques is addressed as follows.
Three classifiers including single, hybrid and ensemble are discussed. Articles on certain classification forms are given in Table 1. These papers shall be distributed annually in Figure 1. In alone, hybrid and ensemble classifications, 70, 62 and 15 papers are written. In 2004, 2011 and 2012, the single classification was extremely concentrated. The papers published are numbered 8, 11 and 18. Many articles are written between 2000 and 2012 on the individual classifier. The hybrid classification
average is three, but the emphasis was on in 2007, and the highest worth in that year was 10 items written. Ensemble classification began in 2003 and for the first time two papers were issued. Classification ensemble was strongly oriented between 2009 and 2012.

Table 1. Articles Written for Types Classifiers

| CLASSIFIERS \ YEAR | 2009 | 2010 | 2011 | 2012 |
|-------------------|------|------|------|------|
| Single Classifier | 7    | 8    | 11   | 18   |
| Hybrid Classifier | 2    | 0    | 3    | 2    |
| Ensemble Classifier | 0  | 10   | 6    | 8    |

6.1. Single Classifiers

There are several individual classifiers, but seven of them have been chosen. The SVM classification is the most common. There are no posts on SVM 31. It represents the largest number of publications in contrast with other publications. In 2009, 20011 and 2012, the highest numbers of articles on SVM as shown in table 2 and figure 4.

Table 2. Articles Written for Types Classifiers.

| CLASSIFIERS \ YEAR | 2009 | 2010 | 2011 | 2012 |
|-------------------|------|------|------|------|
| KNN | 7 | 8 | 11 | 18 |
| GA | 2 | 0 | 3 | 2 |
| SVM | 0 | 10 | 6 | 8 |
| ANN | 2 | 8 | 7 | 4 |
| CNN | 5 | 6 | 4 | 2 |
| AI | 4 | 3 | 5 | 6 |

Figure 4. Year-Wise Work Done for Types of Classifiers.
6.2. Hybrid Classifiers
Important hybrid classification machine teaching strategies are analysed here from 2000 to 2012. The hybrid categories SVM, GA and DT are respectively used for 17, 8 and 7 posts. These are also very common techniques for hybrid classification. Standard use is made with other techniques as shown in figure 5.

![Figure 5. Important Techniques used in Hybrid Classification.](image)

6.3. Ensemble Classifiers
In addition, SVM is a common ensemble classification technique. It is seen for the most part. In 4 articles, SVM is used, whereas in 2 and 2 articles, DT and GA. Only once are used RBF, FL, HMM, Kmeans, SVC and ANN. As shown in figure 6.

![Figure 6. Important Techniques used in Ensemble Classification](image)

For one, hybrid and ensemble classifiers, several machine learning methods are used. In individual, hybrid and ensemble classification systems, SVM is used mostly. Following SVM, GA and DT are the most popular techniques. SVM is used in 31 single grade articles, in 17 hybrid grading articles and in 4 ensemble classifications articles. In combinations and ensemble classifications, SVM also integrates with other strategies.

7. Conclusion
A lot of monitoring and avoidance of intrusions has been achieved. Intrusion detection method, there are several machine learning approaches and they include independent, hybrid and ensemble classifications. Many tools have been used in different techniques of machine learning. These methods function for IDS very well, but it is recognized that not all kinds of attacks are recognizable by a single
technology. Therefore, more efforts are also needed to enhance the productivity of machine learning technology in order to detect all forms of attacks and to reduce false alarms. There are various divisions, but none are total. Classification of hybrids is closer. Taking two or three best individual classifications and improving them a bit, integrating and using them as a single hybrid classification. False alerts can be decreased and the algorithm for choosing features enhanced as well.

8. References

[1]. C. F. Tsai, Y. F. Hsu, C. Y. Lin and W. Y. Lin, “Intrusion Detection by Machine Learning: A Review,” Expert Systems with Application 36, 2009.

[2]. H. Zimmermann, Fuzzy set theory and its applications. Kluwer Academic Publishers. 2001.

[3]. A. A. Aburomman and M. B -I. Reaz, “Evolution of Intrusion Detection System Based on Machine Learning Methods”, Australian Journal of Basic and Applied Sciences, 7(7): 799-813, 2013.

[4]. H. Kaur, G. Singh and J. Minhas, “A Review of Machine Learning Based Anomaly Detection Techniques” International Journal of Computer Applications Technology and Research volume 2 issue 2, 185-187, 2013.

[5]. J. R. Koza, Genetic programming: On the programming of computers by means of natural selection. Massachusetts: MIT, 1992.

[6]. K. Shafi and H. A. Abbass, "An adaptive genetic based signature learning system for intrusion detection." Expert Systems with Applications, 36(10): 12036-12043, 2009.

[7]. Arora, Amandeep Singh, Linesh Raja, and Barkha Bahl. "Data centric security approach: A way to achieve security & privacy in cloud computing." Proceedings of 3rd International Conference on Internet of Things and Connected Technologies (ICIoTCT). 2018.

[8]. Kumar, Ankit, et al. "An improved quantum key distribution protocol for verification." Journal of Discrete Mathematical Sciences and Cryptography 22.4 (2019): 491-498.

[9]. C. M. Bishop, Neural networks for pattern recognition, England, 1995, Oxford University.

[10]. S. Manocha, and M. A. G irolami, “An empirical analysis of the probabilist c K-nearest neighbor classifier,” Pattern Recognition Letters, 28, 18181824.2007.

[11]. HieuTran “A Survey of Machine Learning and Data Mining Techniques Used in Multimedia System” September 23, 2019.

[12]. M. Nikhil Kumar, K.V.S. Koushik, K. John Sundar “Data Mining and Machine Learning Techniques for Cyber Security Intrusion Detection” International Journal of Scientific Research in Computer Science, Engineering and Information Technology © 2018 IJSRCSEIT | Volume 3 | Issue 3 | ISSN: 2456-3307.

[13]. Bilal Ahmad, Wang Jian and Zain Anwar Ali “Role of Machine Learning and Data Mining in Internet Security: Standing State with Future Directions” Journal of Computer Networks and Communications Volume 2018, Article ID 6383145, 10 pages https://doi.org/10.1155/2018/6383145.

[14]. Anna L. Buczak, Member, IEEE, and Erhan Guven, Member, IEEE “A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection” IEEE Communications Surveys & Tutorials, Vol. 18, No. 2, Second Quarter 2016.

[15]. JABEZ J, Dr. B. Muthukumar” Intrusion Detection System (IDS): Anomaly Detection using Outlier Detection Approach” International Conference on Intelligent Computing, Communication & Convergence (ICCC-2015).

[16]. Seema Sharma, Jitendra Agrawal, Shikha Agarwal, Sanjeev Sharma “Machine Learning Techniques for Data Mining: A Survey” 978-1-4799-1597-2/13/$31.00 ©2013 IEEE.