International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Analysis of NSL-KDD Dataset for Fuzzy Based Intrusion Detection System

Macdonald Mukosera[1], Thabiso Peter Mpofu[2], Budwell Masaiti[3]

[1, 2, 3]Jawaharlal Nehru Technological University Hyderabad, School of Information Technology, Kukatpally, Hyderabad 500085, India

Abstract: ***In a bid to provide useful information for intrusion detection, we focused on analyzing the NSL-KDD dataset. In this analysis, we seek to simplify the process of mining fuzzy rules by reducing the features and categorizing the dataset into various smaller clusters as smaller units of the dataset are easier to work with than the whole single large dataset. It is less complex to observe and discover sound fuzzy rules from a smaller dataset and this work serves as a foundation to a fuzzy logic based intrusion detection system. This paper presents a methodology for data preprocessing towards an intrusion detection system and Microsoft excel was used in the process.***

Keywords: Fuzzy rules, fuzzy logic, intrusion, NSL-KDD dataset, mining

1. Introduction

As the world of today heavily relies on computer based systems there is much more need to continue researching and increasing the knowledge of protecting these systems from various kinds of threats being faced by the data housed in these computer systems. [3] Intrusion detection is a field that forms another line of defense in computer security that is it takes over where preventative security measures fail. An intrusion can be defined as any intentional event performed on a computer system where an intruder gains access that compromises the confidentiality, integrity, or availability of computers, networks, or the data residing on them.

Intrusion detection is the resultant work of intrusion detection software systems and these systems mainly fall into two categories that is misuse detection and anomaly detection. [4] Misuse detection approach is to first build a pattern of malicious behavior and then uses that pattern to detect malicious intrusions from normal intrusions which makes it weak in situations of unknown attacks, whereas anomaly detection focuses on defining the expected normal behavior of a network in advance such that in future any access that fails to conform with the defined normal behavior will be classified as an attack. This makes anomaly based systems able to detect unknown attacks whilst they suffer from high false alarm rate than misuse based systems.

In this paper we are focusing on the preprocessing performed on intrusion detection data, in this case we are specifically working on NSL-KDD dataset. The target system which our work is builds up to is a fuzzy logic based anomaly intrusion detection system. There are many approaches to build anomaly detection systems but in this paper we focus on the fuzzy logic approach. The rest of the paper is organized as follows: section 2 describes the KDD 99 dataset and explain the improvements done to KDD 99 to come up with NSL-KDD dataset, section 3 is a description of the fuzzy logic concepts, section 4 discusses about the work related to this paper, section 5 is about our proposed methodology to the data preprocessing, section 6 is about experimental work and in that section we describe the work we did to the NSL-KDD dataset and some experimental analysis to our work, , in section 7 we concluded our work, section 8`` points to the future work in section 9 we list the references to this paper and lastly section 10 is about the authors profiles.

2. KDD and NSL-KDD Dataset Description

NSL-KDD is a name given to a dataset which was produced as a result of research efforts and improvements on the KDD'99 dataset by [1] Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani. We discuss more about this dataset in the related work section. KDD'99 dataset [6] is a very popular dataset that has been the most widely used for the evaluation of intrusion detection systems. This intrusion data set was prepared by Stolfo et al. [7] and was built based on the data captured in DARPA'98 IDS evaluation program. The test data have around 2 million connection records. KDD training dataset consists of approximately 4,900,000 single connection vectors each of which contains 41 attributes and is labeled as either normal or an attack, with exactly one specific attack type. The simulated attack belongs to one of the following classes:

- Denial of Service Attack (DoS): This is a type attack where the attacker makes some computing or memory resource inaccessible to legitimate users. The attacker will open too many connections directly or indirectly to the computer system making it too busy and denying new legitimate connections
- User to Root Attack (U2R): This is a class of attack in which the attacking party initially starts with access to a normal user account on the system gained legally or illegally, and through exploiting some system weakness to gain root access to the system, the attacker ends up with more rights to manipulate the system than the previous limited ones they had.
- Remote to Local Attack (R2L): This situation happens when an Attacker does not have an account to a computer system but is able to send packets from a remote location, and then the attacker exploits some system vulnerability to gain local access as a legitimate

user of that machine.

- Probing Attack: In this kind of attack, the attacker attempts to gather information about a network of computer systems for the purpose of investigating or finding a way to gain access to the system

The test data is not from the same probability distribution as the training data, and it includes specific attack types not in the training data which make the task more realistic. The datasets contain a total number of 24 training attack types, with an additional 14 types in the test data only.

3. Fuzzy Logic

Fuzzy Logic is a mathematical tool that can be used for dealing with uncertainty and it allows us to develop systems that mimic human brain ability to reach a conclusion given vague or imprecise information. [8] With fuzzy logic we are able to represent linguistic constructs such as low, medium, high or very high and more as per situation requirements. The diagram below shows an abstract view of a fuzzy system.

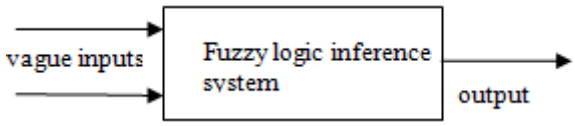

Figure 1: Fuzzy system

Fuzzy logic theory heavily relies on the notion of fuzzy sets and membership functions for its reasoning process. [9] The concept of fuzzy set extends a typical classical set. A fuzzy set includes the possibility of partial membership of its elements. The space from where inputs are drawn is called universe of discourse (UOD). If X is the UOD and its elements are represented by x, then fuzzy set F in X is defined as a set of ordered pairs.

$$F = \{x, u_F(x) \mid x \varepsilon X\} \quad (1)$$

Where $u_F(x)$ is called the membership function of x in set F. the membership function associated with a given fuzzy set F, maps each element of X that is the UOD element to a membership value between 0 and 1. Simply put a membership function is a curve that defines how each point in the universe of discourse is mapped to a membership value between 0 and 1. This membership value represents the degree of membership of an element to a fuzzy set. Examples of common membership function curves are the triangular and Gaussian membership functions.

The fuzzy operators are a superset of standard Boolean logic operators and the truth of any statement in fuzzy logic is a matter of the degree to which an element is a member of a given fuzzy set and the value can be anywhere between 0 and 1, whereas in standard logic operators 1 represent truth and 0 represent false and there are no other values a proposition can take except 0 or 1.

All we have discussed above are the factors that the process of fuzzy inference involves. Fuzzy inference allows us to define IF–Then rules that describe the behavior of the system, thus forming the knowledge of the system. Our main focus in this paper is to structure the intrusion detection dataset in way that will make it easy to visualize if then rules from the dataset. The Fuzzy inference system will map a given set of inputs into an output using the fuzzy logic defined by the if then rules in the knowledge base.

4. Related Work

Much work has been done on analysis of the intrusion detection dataset in different ways using different techniques. All these works have got a goal in common that is to improve the performance of the intrusion detection system and to reduce the false alarm rate of the intrusion detection system. In this subsection we look at some of the related work

Mohammad Khubeb Siddiqui and Shams Naahid [1] analyzed 10% 0f the KDD 99 dataset using data mining methods and they used the k means algorithm which is implemented in the Oracle Data miner tool. In their work they formed 1000 clusters and they managed to establish relationships between the attack types and the protocol used by hackers. They worked with 10% of the data only.

Mahbod Tavallaee and others [2] performed a statistical analysis on the KDDCUP 99 dataset. Since this dataset is the most used in evaluation of intrusion detection systems they found issues that affects the systems evaluated using this dataset. From their work they came up with a new version of KDD 99 dataset which is called NSL-KDD which is the dataset we used in this paper. This dataset leaves out some problems faced by KDD 99 like redundancy, duplication and the number of records is reasonable such that the whole dataset can be used in an experiment rather than selecting a percentage of records only.

For the objective of providing helpful information for intrusion detection systems, [10]Anish Das ,S and Siva Sathya analysed the KDDCUP 99 dataset and they proposed an algorithm for selecting the most relevant features of the dataset that can be best used for classification avoiding misclassification of the records. They came up with a fuzzy approach to feature reduction and their approach minimized the misclassification rates as compared to other feature selection algorithms.

5. Proposed Methodology

1. First step is to acquire the NSL-KDD dataset. The text form of this dataset is available for researchers through the website http://nsl.cs.unb.ca/NSL-KDD/
2. Select all records of normal intrusions to the system and leave out the attack records using data processing software..
3. Select important features to consider in generating the fuzzy rules. The fuzzy approach to feature reduction technique [10] can be used here to minimize the rate of misclassification
4. The resultant from step 3 above is a subset of the original dataset which contains normal records with selected features. Further categorize the dataset into smaller categories such that from those categories it will require little effort to observe patterns that enables easy mining of fuzzy rules

5. Using smaller categories obtained from step 4 above, observe the data patens and mine as much fuzzy rules as possible to create knowledge of the system.

The result will be fuzzy rules for categories of the datasets such that the fuzzy rules can be combined to form one big fuzzy inference system or be treated categorically to form multiple fuzzy inference systems that can be later integrated into a single fuzzy system

6. Experimental Analysis

We downloaded the training dataset with all records that is the attacks and normal intrusions. Initially the records were 125 973 and each record had 41 attributes. After selecting normal intrusion records only without attribute reduction we remained with 67 341 records. We then performed attribute reduction using the fuzzy attribute reduction approach and we reduced the number of the record attributes from 41 to 8. The attributes retained were namely protocol, service, src_bytes, dst_bytes, count, srv_count, dst_host_count, and dst_host_srv_count. This reduction reduces the overhead of dealing with too many attributes in generating fuzzy rules thus making the fuzzy rules less complex and also reducing the misclassification rate but achieving the same task at the end of the day. After this still we were left with 119 000 records which makes it more difficult to exhaust all fuzzy rules hence The dataset was further divided into small categories for easy observation and mining of the fuzzy rules. To achieve these small categories whilst working with the attributes remaining we did the following:

We considered separating the remaining 67 341 records based on the protocol and service attributes. We selected a particular protocol and then separated the data according to each and every service on that protocol thus remaining with a smaller dataset which is easy to process for generating fuzzy logic rules. The process of selecting the records was achieved through the use of Microsoft excel data filtering

The NSL-KDD dataset [11] records are from three protocols namely tcp, udp and icmp, and the total number of services there is 66. So under the tcp protocol we had 62 services hence we ended up with 62 smaller categories on tcp protocol. For the protocol udp we observed that there are 5 services only hence we had 5 smaller categories. For the icmp protocol the total number of services under this protocol was 6 hence we made 6 smaller categories. The normal records of the NSL-KDD dataset we managed to subdivide it into a total of 73 categories. The table below summarizes the protocols and number of services per each protocol

Table 1: Number of protocols and services

Protocol	*Number of services*
tcp	62
udp	5
icmp	6

It is from these 73 categories where we propose to observe and mine the fuzzy rules to build a fuzzy inference system of the fuzzy logic based intrusion detection system. Below is a snippet of the ftp_data service on tcp protocol.

A	B	C	D	E	F	G	H	I
protocol	service	src bytes	dst_bytes	count	srv_count	dst_host_count	dst_host_srv_count	
1	1	641	0	2	2	175	48	
1	1	12	0	1	1	44	47	
1	1	748	0	1	1	6	92	
1	1	9	0	2	2	185	119	
1	1	1766	0	2	2	255	6	
1	1	245	0	3	2	99	68	
1	1	19	0	1	1	201	5	
1	1	14416	0	1	1	205	50	
1	1	7422	0	8	8	255	44	
1	1	12	0	2	2	11	88	
1	1	567	0	1	1	53	38	
1	1	10389	0	1	1	255	81	
1	1	319	0	2	2	6	2	
1	1	276	0	1	1	76	24	
1	1	192	0	2	3	255	115	
1	1	852	0	6	6	219	39	
1	1	12	0	1	2	75	65	
1	1	13239	0	4	4	100	63	
1	1	383	0	3	2	37	29	
1	1	4491	0	1	1	63	79	
1	1	641	0	2	2	62	55	
1	1	59	0	16	16	187	52	
1	1	1874	0	9	9	151	30	
1	1	12	0	1	1	137	44	

Figure 2: ftp_data service on tcp protocol

In the diagram above the value one (1) in the protocol column stands for the protocol tcp and the value one (1) in the service columns represents the ftp_data service. We converted protocol and service to numeric for easy of processing the data. Taking a closer look at the snippet above, we can observe that the value of the dst_bytes is always zero in case of protocol tcp and ftp_data service. Also the values of count and srv_count are not exceeding thirty (30) in this situation. So after proper definition of the ranges of low, medium and high values, we can easily come up with a fuzzy rule in the case of ftp_data on tcp_service as follows:

If dst_bytes is low and if count is low and if srv_count is low then intrusion is normal

The above fuzzy rule was just an example where dst_bytes, count, srv_count and intrusion are fuzzy variables whilst low, medium, high and normal are fuzzy sets. It can be seen that more rules can be mined and the process of mining the rules is less complex than trying to observe 41 attributes without any strategic division to form smaller categories that are easy to work with.

7. Conclusion

The objective of this research was to analyze and process the intrusion detection dataset and leave the data in a way that we can easily mine fuzzy rules of the system. This work contributes to the development of an improved performance fuzzy based intrusion detection as working with the smaller data units not only reduces complexity but also enhances performance by reducing the overhead of processing large dataset with many attributes at once.

8. Future Work

The proposed methodology of NSL-KDD dataset preprocessing provides a simple and effective way to process the whole intrusion dataset as smaller components, making it easier to observe hidden patterns in the dataset.

The Future scope of this study is to further mine and exhaust all the possible fuzzy rules from these smaller datasets and build an improved anomaly fuzzy logic based intrusion detection system.

References

[1] Mohammad Khubeb Siddiqui and Shams Naahid, " Analysis of KDD CUP 99 Dataset using Clustering based Data Mining," International Journal of Database Theory and Application Vol.6, No.5 (2013), pp.23-34
[2] Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani," A Detailed Analysis of the KDD CUP 99 Data Set," Proceedings of the 2009 IEEE Symposium on Computational Intelligence in Security and Defense Applications (CISDA 2009)
[3] George Mohay, Alison Anderson, Byron Collie, Olivier de Vel, and Rodney McKemmish , "Computer and Intrusion Forensics," Artech House 2003
[4]Kapil Wankhade, Sadia Patka, and Ravindra Thool, "An Efficient Approach for Intrusion Detection Using Data Mining Methods," IEEE 2013
[5] Yingbing Yu, Han Wu, "Anomaly Intrusion Detection Based Upon Data Mining Techniques and Fuzzy Logic," 2012 IEEE International Conference on Systems, Man, and Cybernetics October 14-17, 2012, COEX, Seoul, Korea
[6] KDDCup1999,Online. Available: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html, Ocotber 2007. [Accessed: June. 11, 2014].
[7]S. J. Stolfo, W. Fan, W. Lee, A. Prodromidis, and P. K. Chan, "Costbased modeling for fraud and intrusion detection: Results from the jam project," discex, vol. 02, p. 1130, 2000.
[8] S. N. Sivanandam, S. Sumathi and S. N. Deepa, "Introduction to Fuzzy Logic using MATLAB," Springer-Verlag Berlin Heidelberg 2007
[9] Raj Kumar Bansal, Ashok Kumar Goel and Manoj Kumar Sharma, "Matlab and its applications in engineering" Pearson, 2009
[10] Anish Das ,S and Siva Sathya, "A Fuzzy Approach to Feature Reduction in KDD Intrusion Detection Dataset," ICCCNT 12 July 26 - 28, 2012 - Coimbatore, India

Author Profile

Macdonald Mukosera received the B.Tech Hons degree in Computer Science from Harare Institute of Technology in 2010. During 2011-2012, he worked as a Teaching assistant at Harare Institute of Technology in the Software Engineering Department. He is now studying M Tech Computer Science at JNTUH SIT India.

Thabiso Peter Mpofu received B. Tech degree in Computer Science at Harare Institute of Technology (HIT), Zimbabwe in 2010. He is currently pursuing M. Tech Computer Science at JNTUH, India. He is a HIT staff development research fellow. His research interests are in the area of Data Mining, Network Security and Mobile Computing.

Budwell T Masaiti received the B.Tech Hons degree in Computer Science from Harare Institute of Technology (Zimbabwe) and Daejeon University (South Korea) in 2011. During 2011-2012, he worked as a Teaching assistant at Harare Institute of Technology in the Computer Science Department. He is currently pursuing M Tech Computer Science at JNTUH College of Engineering. (India)