IJDACR
ISSN: 2319-4863

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 7, Issue 11, June 2019)

Intrusion Detection System using Fuzzy C Means and Support Vector Machines

Durgesh Pawar
durgeshpawar1991@gmail.com

Amit Thakur
amitthakurbist@gmail.com

Abstract – **This paper is focused on intrusion detection systems with the development of improvements based on the modeling of the messages exchanged by a communications protocol.**
This research work proposes an intrusion detection framework using Fuzzy C Means (FCM) and Support Vector Machine (SVM) classifier to classify the KDD cup'99 data into five classes; NORMAL, DOS, PROBE, U2R, R2L. The data pre-processing is accomplished by Mapping and Min-Max normalization. FCM is used to select the best features of the dataset. The performance of proposed research work is evaluated using certain evaluation parameters; accuracy, sensitivity, and precision.

Keywords – **FCM, IDS, SVM.**

I. INTRODUCTION

One of the defence mechanisms most used to reduce the risk of companies against attacks directed towards computer goods has been the Intrusion Detection Systems (IDS).
An IDS is an element that listens to and analyses all the information that circulates through a data network and identifies possible attacks. When an attack appears, the system will react informing the administrator and close the doors to the possible intruder reconfiguring elements of the network such as firewalls and routers.
IDS has been widely used over the last few years by many companies because they have provided an additional layer of security. However, it has been found that these elements provide reactive security, that is, for protection there must first be an attack. If an attack is small and effective the IDS will react too late and the attack will achieve its objective.
The first investigations on intrusion detection begin in 1980 in a consultancy work carried out for the North American government by James P. Anderson [1]. Anderson, presented the idea that the normal behaviour of a user could be characterized by analysing their activity in the audit records. Thus, intrusion attempts could be discovered by detecting anomalous activities that deviate markedly from their normal behaviour. From this moment, they begin a great variety of investigations in search of techniques and algorithms for the analysis of the behaviour of a computer system [2] [3].
The IDS allow, not only the detection of attacks covered by other security components, but the detection of intrusions that go unnoticed to other components of the security device. Thus, they perform two fundamental tasks: prevention and reaction [4].
The prevention of intruder activities is done through tools that listen to traffic on the network or on a computer. These programs identify the attack by applying pattern recognition (standards) or intelligent techniques. This work allows to report the attempts of attacks or suspicious activities immediately [5].
There is also the possibility of developing defensive responses before the attack materializes. The reactive method is guaranteed using programs that basically perform the analysis of log files in the systems. It is about detecting patterns of intrusion in the traces of network services or in the behaviour of the system. Modification of common files, system files and others are also considered as signs of intrusion [6].
The main objective of this paper is to implement an intrusion detection framework using Fuzzy C Means clustering and Support Vector Machine classifier to classify the KDD cup'99 data into five classes; NORMAL, DOS, PROBE, U2R, R2L.

II. PROPOSED METHODOLOGY

KDD-Cup99 dataset is used for the proposed work. Initially the data is pre-processed using the certain operations then the feature selection is accomplished using fuzzy c-means algorithm. Selected features are classified by the SVM. Rest of the methodology is explained in the following headings.

A. KDD Cup-99 Dataset

The work described in this paper uses the KDD 1999 database of the KDD 1999 data set [7]. This data set, apart from its obsolete date, has two major flaws:

IJDACR
ISSN: 2319-4863

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 7, Issue 11, June 2019)

1. The attack rate in the KDD dataset is unnatural. About 80% of all cases are attacks, since all the attacks of a packet, for example the smurf attack, are treated as full value connections and are represented as individual cases.
2. The distribution of attacks within the KDD dataset is very unbalanced. It is dominated by denial of service attacks, which cover millions of cases. The most interesting and dangerous attacks, for example web attacks are largely underrepresented.

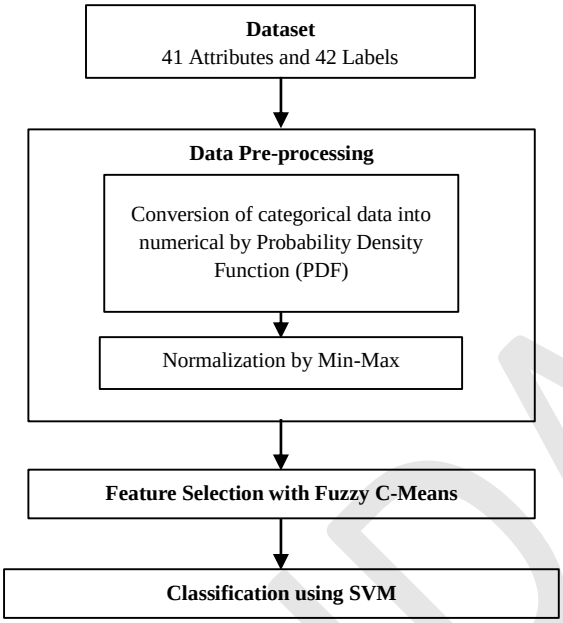

Figure 1: Flow diagram of proposed intrusion detection system

B. Data Pre-processing

1) Mapping

The dataset is labelled according to the following attacks which are fall into one of five categories listed below in Table 1.

Table 1: Class labelling of "10% KDD'99" dataset

Attack Type	Description	Sub Types	Label
(DoS) Denial of Service	Attacker tries to prevent legitimate users from using a service	Smurf	1
		Neptune	
		Back	
		Teardrop	
		Pod	
		Land	
Normal	data with no attack	normal	2
Probe	Attacker tries to prevent legitimate users from using a service	Satan	3
		Ipsweep	
		Portsweep	
		Nmap	
(R2L) Remote to Local	Attacker does not have an account on the victim machine, hence tries to gain access	Warezclient	4
		guess_passwd	
		Warezmaster	
		Imap	
		ftp_write	
		Multihop	
		Phf	
		Spy	
User to Root (U2R)	Attacker has local access to the victim machine and tries to gain super user privileges	buffer_overflow	5
		Rootkit	
		Loadmodule	
		Perl	

2) Normalization by Min-Max

Min-Max Normalization performs a linear transformation on the original data x into the specified interval (New_{min}, New_{max}).

$$x_i = New_{min} + (New_{max} - New_{min}) \times \left(\frac{x_i - x_{min}}{x_{max} - x_{min}}\right) \quad (1)$$

$$x_{max} = \max_{1 \le i \le N} x_i, x_{min} = \min_{1 \le i \le N} x_i \quad (2)$$

This method scales the data from (x_{min}, x_{max}) to (New_{min}, New_{max}) in proportion. The advantage of this method is that it preserves all relationships of the data values exactly. It does not introduce any potential bias into the data.

C. Feature Selection with Fuzzy C-Means

The conventional FCM algorithm is a worldwide accepted image segmentation method. The segmentation of imaging data involves partitioning the image data into different cluster regions with similar intensity image values. Mostly medical images always present overlapping gray-scale intensities for different tissues. Therefore, fuzzy clustering methods are suitable for the segmentation of medical images. FCM can be seen as the fuzzified version of the *K*- means algorithm. It is a method of clustering which allows one piece of data to belong to two or more clusters. FCM is a data clustering technique in which each sample data point belongs to a cluster to some degree that is specified by a membership grade. The FCM clustering algorithm was first introduced by DUNN [8] and later was extended by BEZDEK [9], which has been

IJDACR
ISSN: 2319-4863

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 7, Issue 11, June 2019)

successfully applied to image segmentation by minimizing following objective function with respect to fuzzy membership, and set of cluster centroid V .

$$J_m(U,V)=\sum_{j=1}^{N}\sum_{i=1}^{C}u_{ij}^{m}d^{2}(\mathbf{x}_j,\ \mathbf{v}_i) \tag{3}$$

Where:

The parameter m is weighting exponent on each fuzzy membership and control the degree of "fuzziness" of the resulting classification. In actual application, $m=2$ is frequently adopted. $X=\{\mathbf{x_1},\ \mathbf{x_2},\ldots\mathbf{x}_j,\ldots\mathbf{x}_N\}$ is a $p\times N$ data matrix, where p represents the dimension of each $\mathbf{x}_j$ "feature" vector, and N represents the number of feature vectors (pixel numbers in the image). C is the number of clusters. U is fuzzy matrix. $d\left\|\mathbf{x}_j-\mathbf{v}_i\right\|$ is the Euclidean norm. The membership function is expressed as follows [10]:

$$u_{ij}=\frac{1}{\sum_{k=1}^{C}\left(\frac{d(\mathbf{x}_j,\ \mathbf{v}_i)}{d(\mathbf{x}_j,\ \mathbf{v}_i)}\right)^{2/(m-1)}} \tag{4}$$

$V=\{\mathbf{v_1},\ \mathbf{v_2},\ldots\mathbf{v}_i,\ldots\mathbf{v}_C\}$ Which is a $p\times C$ matrix and denotes the cluster feature center

$$\mathbf{v}_i=\frac{\sum_{j=1}^{N}(u_{ij})^m\,\mathbf{x}_j}{\sum_{k=1}^{C}\left(\sum_{j=1}^{N}u_{ij}\right)^m}\quad(i=1,2,\ldots,C) \tag{5}$$

The Process stops when $\left\|J_{\mathrm{m}}{}^{t+1}-J_{\mathrm{m}}{}^{t}\right\|<\varepsilon$ or predefined number of iterations is reached, whereas t is no. of iteration.

D. Classification using Support Vector Machine (SVM)

Support Vector Machine, or SVMs, are the set of supervised machine learning algorithm designed to solve out regression and discrimination problems. SVMs are generalization of linear classifiers.

The objective is to find a decision boundary (hyperplane separator) that can separate the cloud of points into two regions by making a minimum of errors, i.e., (find the optimal hyperplane) [11].

SVMs can be used to solve binary discrimination problems, that is, decide which class a sample belongs to. The solution to this problem is the construction of a function f which, at an input vector $x\in X$ matches an output $f(x)$: It is then decided that x is of class $+1$ if $f(x)>0$ and of class -1 if $f(x)<0$. It is a linear classifier. The decision boundary $f(x)=0$ is a separating hyperplane.

Let H be a hyperplane, w its normal vector, and b its offset from the origin (see Figure 2). The hyperplane H is then given by [11]:

$$f(x)=w^{T}x+b=0 \tag{6}$$

The goal of the SVM learning algorithm is to find out the parameters w and b of the best hyperplane through a learning set:

$$\mathcal{X}\times\mathcal{Y}=\{(x_1,y_1),\ldots,(x_i,y_i)\}\in\mathbb{R}^{N}\times\{+1,-1\} \tag{7}$$

Where, y_i are the respective labels of x_i, N is a size of learning set.

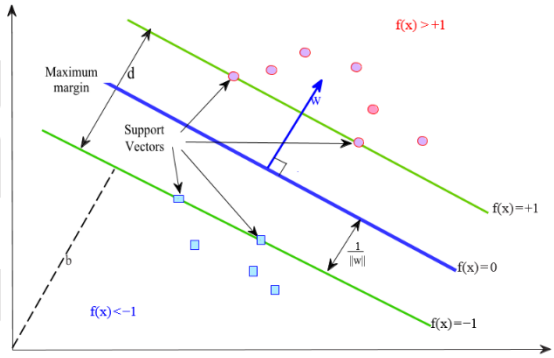

Figure 2: Margin illustration and support vectors [12]

III. Simulation And Results

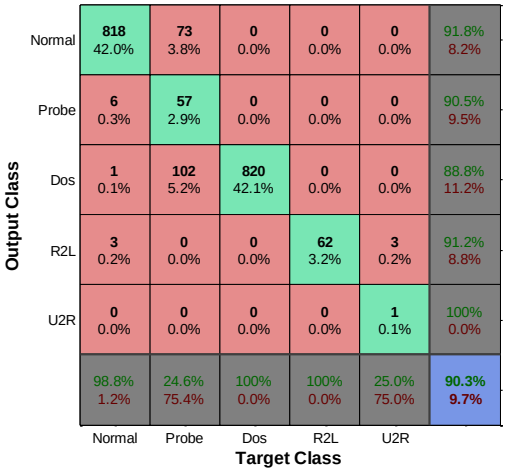

Figure 4: Confusion matrix plot for attack classification using SVM classifier with 60-40 ratio (60 % training and 40 % testing)

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 7, Issue 11, June 2019)

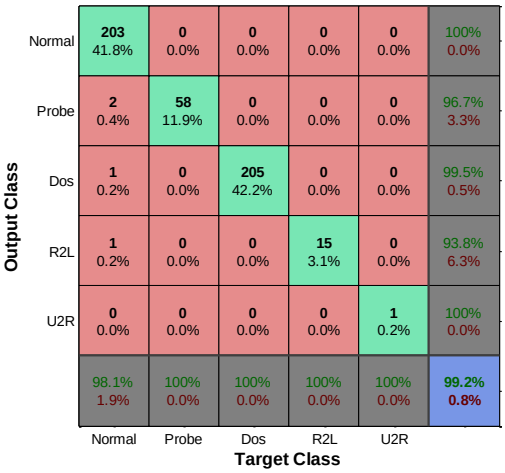

Figure 3: Confusion matrix plot for attack classification using SVM classifier with 90-10 ratio (90 % training and 10 % testing)

IV. CONCLUSION

The degree of connectivity that is experienced today is very high. Who does not have an Internet connection today? Few are those who resist, because today's society requires that degree of connectivity, to precisely be able to access these data, which are no longer on paper, or in libraries, or in office filing cabinets or bank safes, they are in the network. That is why it is essential to keep these data safe, because now the threats are not only thieves with balaclavas, but anyone with a computer, a connection and not very good intentions

In this paper, it has been tried to show the importance and the possibilities that the Intrusion Detection Systems can have on the security of a network. As well as, demonstrate that these systems can still be perfected, helping us with already existing techniques, but not applied in this field. So, by combining aspects of different ways of solving the problem of intrusion detection, a more complete product is obtained that makes our network a safer and more controlled site, not only of known attacks, but also of those still to be known.

This paper presented a Fuzzy C Means (FCM) clustering and Support Vector Machine (SVM) classifier to classify the KDD cup'99 data into five attack classes. After observing the simulation results, it was found that the SVM offers better performance on the basis of accuracy, precision and sensitivity.

REFERENCE

[1] Anderson, James P. *Computer security technology planning study*. Vol. 1. ESD-TR-73-51, 1972.

[2] Alkasassbeh, Mouhammd, Ghazi Al-Naymat, A. B. Hassanat, and Mohammad Almseidin. "Detecting distributed denial of service attacks using data mining techniques." *International Journal of Advanced Computer Science and Applications* 7, no. 1 (2016): 436-445.

[3] Kushwaha, Prashant, Himanshu Buckchash, and Balasubramanian Raman. "Anomaly based intrusion detection using filter based feature selection on KDD-CUP 99." In *TENCON 2017-2017 IEEE Region 10 Conference*, pp. 839-844. IEEE, 2017.

[4] Idhammad, Mohamed, Karim Afdel, and Mustapha Belouch. "Dos detection method based on artificial neural networks." *International Journal of Advanced Computer Science and Applications* 8, no. 4 (2017): 465-471.

[5] Aljawarneh, Shadi, Monther Aldwairi, and Muneer Bani Yassein. "Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model." *Journal of Computational Science* 25 (2018): 152-160.

[6] Ahmad, Iftikhar, Mohammad Basheri, Muhammad Javed Iqbal, and Aneel Rahim. "Performance comparison of support vector machine, random forest, and extreme learning machine for intrusion detection." *IEEE Access* 6 (2018): 33789-33795.

[7] Tavallaee, Mahbod, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani. "A detailed analysis of the KDD CUP 99 data set." In *2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications*, pp. 1-6. IEEE, 2009.

[8] Dunn, Joseph C. "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters." (1973): 32-57.

[9] Bezdek, James C. *Pattern recognition with fuzzy objective function algorithms*. Springer Science & Business Media, 2013.

[10] Nayak, Janmenjoy, Bighnaraj Naik, and H. S. Behera. "Fuzzy C-means (FCM) clustering algorithm: a decade review from 2000 to 2014." In *Computational intelligence in data mining-volume 2*, pp. 133-149. Springer, New Delhi, 2015.

[11] Li, Li. "Support vector machines." In Selected applications of convex optimization, pp. 17-52. Springer, Berlin, Heidelberg, 2015.

[12] Guenther, Nick, and Matthias Schonlau. "Support vector machines." Stata J 16, no. 4 (2016): 917-937.