ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 5, May 2015

Intrusion detection using classification via clustering

Divya D. Nimbalkar[1], Shubha Puthran[2]

Student, Computer Engineering, MPSTME, Mumbai, India[1]

Assistant Prof.ShubhaPuthran, Computer Engineering, MPSTME, Mumbai, India[2]

Abstract: In today's world there is widespread use of internet. It hence becomes a necessity for securing this access to the data that is stored on theworld wide web . Intrusion detection system is one such mechanism for detecting the intrusive patterns from the traffic patterns on the network . Datamining and statistical data analysis are some ways to detect these attacks. In this paper, we have presented a novel technique of intrusion detection where is classification is done on the results one gets after clustering the data set KDD '99 . The results obtained here are better than directly performing classification or clustering.

Keywords: Intrusion detection , data mining , statistical analysis , KDD '99

I. INTRODUCTION

Intrusion detection systemhelps monitoring the traffic over the internet so that if any malicious activity comes in it can be alarmed to the users. These intrusive patterns are stored in the database for future reference as attack pattern.

There are two different detection techniques employed in IDS to search for attack patterns Misuse and Anomaly[1][2]. In Misuse detection systems the known attack signatures are looked for in the monitored resources while in Anomaly detection systems attacks are found by detecting changes in the pattern of behavior of the system.

We find many data mining techniques like classification , clustering , association rule mining to be used for this detection of intrusive patterns . Also statistical analysis like chi square analysis can be used for detection of intrusive patterns .In case of classification the known patterns of attacks are only classified and clustering helps identifying unknown attack patterns as well . Classification done alone is therefore of no use because if any new pattern arrives in it would fail to detect the attack. However , the results when compared to classification done after clustering proves better because it initially clusters all unknown patterns of data that fall in one category and then classifies them accordingly later.

Classification technique being a supervised method of learning only the ones that are previously classified to one of the classes of intrusion or as normal pattern will be classified further and no new attack pattern will beidentified[1][2] . Like if there are two attack patterns smurf and portsweep only attack patterns that have similar values to the recordspreviously classified as smurfor portsweep will be identified further and if anynewattackpattern like sync flood or buffer overflow comes in than the attack wont be identified .

Clustering on the other hand being unsupervised form of learning will cluster all the attack patterns with similar values and form a cluster depicting one attack pattern [2] . Like stated in the previous case even though the data base has values stored only for smurf and portsweep and new attack say buffer overflow or sync flood comes in then this new attack pattern will be clustered in the new cluster showing the attack is of new category other then the ones already present .

Statistical analysis technique like chi square analysis set up a threshold value for checking if a particular pattern is intrusive or not [7] . However , detection of individual record as intrusive or non intrusive here highly depends on the significance level chosen and also the threshold set is very general as to bifurcate records as intrusive or not but it cannot identify individual attack patterns .

Classification done after clustering helps overcome this problem since the clustered results are more accurate with newly identified attack patterns and classification performed on these results will give better classified records than directly performing classification whichonly identifies known attack patterns

II. EXPERIMENTAL SETUP

An easy way to comply with the conference paper formatting requirements is to use this document as a template and simply type your text into it.

A. Database Pre-processing :

The KDD 99 dataset is used for our analysis purpose . It has around 41 different attributes and around 500000 records .
We have preprocessed it to select the most relevant attributes based on information gain and haveselected these 15attributes:

1. Duration
2. Protocol_type
3. Service
4. Flag
5. Src_bytes
6. Dst_bytes

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 5, May 2015

7. Wrong fragment
8. Failed login
9. Logged_in
10. Num of roots
11. Count
12. Srvc_count
13. Serviceerror_rate
14. Dst host count
15. Dst host srvc count

We have generalized the dataset so as to include all different attack types in the four categories of DOS , probe , r2l and u2r which are further given numerical values 2 , 3 , 4 and 5 respectively (1 being for normal records)

Also the other attributes having textual values have been converted to numerical values so that it could be used for clustering analysis . Like the protocol_type attribute has values tcp ,udp and icmp which are given values as 201 , 202 and 203 respectively .Similar is the case with the other attributes.

Also we are working on 50000 randomly selected records having a combination of all attack types and normal records where in we have –

TABLE I
ATTACK TYPES AND THE NUMBER OF SUCH RECORDS IN DATASET

Attack type	No. Of Records
Normal (class 1 in dataset)	37866
DOS	11624
Probe	391
R2L	113
U2R	5

B. Classification of the dataset using decision tree :

The preprocessed dataset is used for analysis using decision tree in R . The decision tree algorithm starts with the attribute having highest information gain and then splits the attribute values in ranges and continues the process till the time you reach at leaf nodes where no further splitting is possible. The classification of the 30 percent testing set as per the decision tree algorithm is as follows

TABLE II
CLASSIFICATION RESULT USING DECISION TREE ALGORITHM IN R

testPred	DOS	Normal	Probe	R2L	U2R
DOS	3517	6	1	0	0
normal	1	11316	3	3	0
probe	1	4	108	0	0
r2l	0	1	0	27	0
u2r	0	4	0	0	0

The above Table II depicts out of 30 percent testing data in every class how many are correctly classified and how many are incorrectly classified. Like in case of DOS attack 3524 records forms to be 30 percent of 11624 DOS attack records actually present in dataset of which 3517 are correctly classified and 7 are incorrectly classified. Similarly , we can find the results for other classification

C. Clustering using Kmeans :

In case of clustering using Kmeans means or clusters are selected as per the number of classes . Like in case if we have 4 different classes say Normal , DOS , Probe and R2l so we can set the number of clusters and then checking results of clustering we find the results as –

TABLE III
CLUSTERING RESULT IN R USING KMEANS ALGORITHM

Attack Type	1	2	3	4
DOS	10630	994	0	0
normal	37773	80	13	0
probe	390	0	0	1
r2l	113	0	0	0

Here in above Table III we find cluster 1 has most records of DOS , normal , probe and R2l while cluster 2 , 3 and 4 have the ones that are misclustered and not with the group

Similarly if we add a new attack type to the group and specify the number of clusters to be 5 we get following result-

TABLE IV
CLUSTERING RESULT WITH NEW RECORDS OF NEWLY INTRODUCED ATTACK

Attack type	1	2	3	4	5
DOS	0	0	994	10630	0
Normal	0	31	47	37776	12
Probe	1	0	0	390	0
R2L	0	0	0	113	0
U2R	0	0	0	5	0

The Table IV above shows how a newly introduced attack u2r gets merged into cluster 4 which proves clustering helps identifying new attacks .

D. Classification via clustering :

In classification via clustering , the results of clustering are used for classification . The clusters are formed initially and their cluster values are appended to the dataset as to which cluster every record is put into then classification is performed on this dataset. The results are as follows :

TABLE V
CLASSIFICATION VIA CLUSTERING RESULT IN R

testPred1	DOS	Normal	Probe	R2l	U2R
DOS	3422	3	0	0	0
Normal	0	11485	8	3	2
Probe	0	5	108	0	0
R2l	0	0	0	28	0
U2r	0	0	0	0	0

In the Table 5 above for 30 percent test data we find only 3 records to be misclassified which was 7 in case of directly performing classification 994 on using only clustering. Similarly we can check for other records also.

DOI 10.17148/IJARCCE.2015.4554

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 5, May 2015

III. RESULT ANALYSIS

A. *Classification :*

TABLE VI
ANALYSIS OF RESULT FROM CLASSIFICATION

Testing Set			
Attack type	**Correctly classified**	**Incorrectly classified**	**Percentage incorrect**
DOS	3517	7	0.198
Probe	108	5	4.424
R2L	27	1	3.57
U2R	0	4	infinite
Normal	11316	7	0.061

B. *Clustering:*

TABLE VII
ANALYSIS OF RESULT FROM CLUSTERING

Attack type	**Correctly classified**	**Incorrectly classified**	Percentage incorrect
DOS	10630	994	8.55
Probe	390	1	0.255
R2L	113	0	0
U2R	5	0	0
Normal	37776	90	0.237

C. *Classification via clustering :*

TABLE VIII
ANALYSIS OF RESULT FROM CLASSIFICATION VIA CLUSTERING

Testing Set			
Attack type	**Correctly classified**	**Incorrectly classified**	**Percentage incorrect**
DOS	3422	3	0.087
Probe	108	5	4.424
R2L	28	0	0
U2R	0	0	0
Normal	11485	11	0.095

The above tables shows the percentage of records that are correctly identified to different categories when you find the percentage of this we find that classification via clustering gives out the best result when compared to the other data mining techniques.

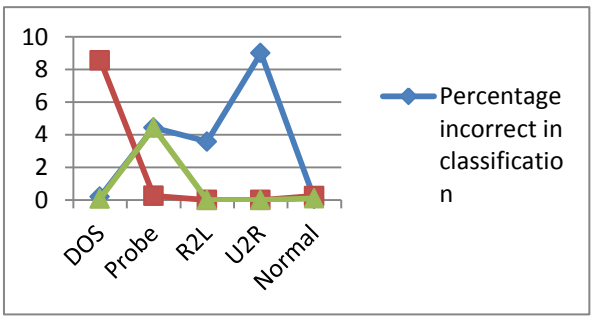

Fig.1Graph indicating accuracy of data mining techniques in intrusion detection

From the Fig 1.above we can see that the green line has maximum misidentification percentage to be between 4 and 5 whereas the other two techniques have maximum values beyond 5 .

IV. CONCLUSION

In this paper we have IDS implementation using classification via clustering which outperforms the other data mining techniques and statistical analysis. We have used KDD 99 dataset wherein necessary preprocessing steps have been applied so that the same can be used for our analysis . The major advantage of the technique proposed is it helps identifying new attacks introduced in the dataset which are not identified if classification is applied directly . Also the same process is applied in parallel on multiple cores to reduce the processing time . This is just a prototype further work includes analyzing the entire dataset to check the performance of data mining techniques .

ACKNOWLEDGMENT

I acknowledge the guidance of my mentor Assistant **Prof.ShubhaPuthrarn** for her continual guidance in carrying out the different experimentation and analysing the result .

REFERENCES

[1] Shubha M.P, Ketan D.S,"Data Mining Classification Approaches for Intrusion Detection System," International Journal of Computer Engineering and Software Technology, 2011
[2] Shubha M.P, Ketan D S., " Review on Data Mining Approaches for Intrusion Detection System, " Technopath, MPSTME, NMIMS ,2011
[3] Asim D , S. Siva S , "Association rule mining for KDD intrusion detection dataset ," International Journal of Computer Science and Informatics , 2012
[4] Swati D , N Z T, "Design of Intrusion Detection System using Fuzzy Class-Association Rule Mining based on Genetic Algorithm ," International Journal of Computer Applications, 2012
[5] H. G K, Nur Z.H, Malcolm I. H , "Selecting Features for Intrusion Detection: A Feature Relevance Analysis on KDD 99 Intrusion Detection Datasets"
[6] M. S.K ,Maybin M , Frans C, " Weighted Association rule mining from binary and fuzzy data," Springer-Verlag Berlin Heidelberg , 2008
[7] Rahul R , Zubair K , M.H. Khan , "Network Anomalies Detection Using Statistical Technique : A Chi- Square approach, "IJCSI International Journal of Computer Science Issues,2012
[8] Deepthy K Denatious& Anita John , " Survey on Data Mining Techniques to Enhance Intrusion Detection," International Conference on Computer Communication and Informatics, 2012

Website :
[1] http://stattrek.com/chi-square-test/independence.aspx
[2] https://www.draw.io/#LUntitled%20Diagram
[3] http://www2.cs.uregina.ca/~dbd/cs831/notes/confusion_m atrix/confusion_matrix.html
[4] http://www.tutorialspoint.com/uml