International Journal of Communication Networks and Security

Volume 2 | Issue 2 Article 11

April 2013

FEATURE SELECTION AND CLASSIFICATION OF INTRUSION DETECTION SYSTEM USING ROUGH SET

NIKITA GUPTA
Department of Computer Engineering, Army Institute of Technology Pune, India, gupta_nikita12@yahoo.co.in

NARENDER SINGH
Department of Computer Engineering, Army Institute of Technology Pune, India, nsbis4ever@gmail.com

VIJAY SHARMA
Department of Computer Engineering, Army Institute of Technology Pune, India, vijaydjrocks@gmail.com

TARUN SHARMA
Department of Computer Engineering, Army Institute of Technology Pune, India, tarun98601@gmail.com

AMAN SINGH BHANDARI
Department of Computer Engineering, Army Institute of Technology Pune, India, amansbhandari@gmail.com

Follow this and additional works at: https://www.interscience.in/ijcns

Part of the Computer Engineering Commons, and the Systems and Communications Commons

Recommended Citation

GUPTA, NIKITA; SINGH, NARENDER; SHARMA, VIJAY; SHARMA, TARUN; and BHANDARI, AMAN SINGH (2013) "FEATURE SELECTION AND CLASSIFICATION OF INTRUSION DETECTION SYSTEM USING ROUGH SET," *International Journal of Communication Networks and Security*: Vol. 2 : Iss. 2 , Article 11.
DOI: 10.47893/IJCNS.2013.1085
Available at: https://www.interscience.in/ijcns/vol2/iss2/11

This Article is brought to you for free and open access by the Interscience Journals at Interscience Research Network. It has been accepted for inclusion in International Journal of Communication Networks and Security by an authorized editor of Interscience Research Network. For more information, please contact sritampatnaik@gmail.com.

FEATURE SELECTION AND CLASSIFICATION OF INTRUSION DETECTION SYSTEM USING ROUGH SET

NIKITA GUPTA[1], NARENDER SINGH[2], VIJAY SHARMA[3], TARUN SHARMA[4], AMAN SINGH BHANDARI[5]

[1,2,3,4,5]Department of Computer Engineering, Army Institute of Technology Pune, India
E-mail: [1]gupta_nikita12@yahoo.co.in, [2]nsbis4ever@gmail.com, [3]vijaydjrocks@gmail.com, [4]tarun98601@gmail.com, [5]amansbhandari@gmail.com

Abstract- With the expansion of computer network there is a challenge to compete with the intruders who can easily break into the system. So it becomes a necessity to device systems or algorithms that can not only detect intrusion but can also improve the detection rate. In this paper we propose an intrusion detection system that uses rough set theory for feature selection, which is extraction of relevant attributes from the entire set of attributes describing a data packet and used the same theory to classify the packet if it is normal or an attack. After the simplification of the discernibility matrix we were to select or reduce the features. We have used Rosetta tool to obtain the reducts and classification rules. NSL KDD dataset is used as training set and is provided to Rosetta to obtain the classification rules.

Keywords- *Intrusion Detection System, Rough Set Theory, NSL KDD dataset, feature reduction, feature selection, Rosetta.*

I. INTRODUCTION

The modern era is hugely dependent on computer for almost every activity. While there are some people working hard to invent new technologies for the betterment of computer networks, one cannot ignore the ones who are looking to break the same. We live in a world where our lives depend hugely on computer networks and necessary measures have to be taken to ensure its continuous availability and security. Actually a computer networks can be breached by several types of attacks. More formally what we are talking about here is an intrusion: An intrusion can be defined as "the act of causing obtrusion or an inappropriate situation" [1]. To prevent such attacks we can use measures like firewall, but what firewall does is it only prevents the attacks. But once an attacker is successfully able to breach into the network we must have a way to detect that attack, this is why we need an intrusion detection system. Thus an intrusion detection system IDS is nothing but a system to detect intrusion. To detect an intrusion we analyze the packets in the network. The packets are defined on the basis of some features. We could improve the efficiency of our system by reducing or extracting a minimal set of these features that would be enough for classifying that packet as either attack or normal. To get the minimum set of features we use rough set theory, by the virtue of which we are able to minimize the features needed to successfully classify a packet and also not losing any feature which would have led to a different result. Rough Set Theory is used to obtain reducts along with a set of rules for training the system. Once the rules are obtained they can be used to classification of the test set. The paper has been divided into following sections: The first section gives a brief description about the rough set theory. The second section talks about the NSL KDD dateset. The third section contains the algorithm designed to obtain the discernibility matrix and its simplification for reduct construction. In the fourth and the fifth section we have given the details of our proposed system along with experimental results.

II. ROUGH SET THEORY

Zdzisaw Pawlak in 1982 proposed the Rough Set Theory [2]. Rough Set Theory is a mathematical tool that deals with partial information. It is concerned with the classification of incomplete or uncertain information or knowledge [3].

Rough Set Reduction Theory can be used to reduce the number of attributes, i.e., attribute reduction. This can be useful in reducing the time required for training an intrusion detection system [4].

In Rough Set Theory the data is represented in terms of table known as decision table or information table [5]. The rows are the objects which may be an event and the columns correspond to the attributes for describing the objects. The attribute set is further divided into condition attribute and decision attribute.

The objects are categorized into various equivalences classes based on the values of the decision attribute.

$$I = (U, A)$$

Here I is an information system, U is the non-empty finite set of objects known as universe and A is the non-empty finite set of attributes such that $a: U \rightarrow V_a$ for all $a \in A$. Va is the value set of a [2].

A. Indiscernible Objects

The set of objects having same attribute values for the attribute under consideration and the relationship between such objects is known as indiscernibility relation. For every nonempty subset B of A the indiscernibility relationship is given by

$$IND_{IS}(B)=\{(x,x') \in U^2 \mid \forall a \in B,\ a(x)= a(x')\}$$

Here the objects x and x′ are indiscernible from each other by attributes from B [4].

B. Discernibility Matrix

Matrix with equivalence classes as indices. The boxes are filled with attributes for which the corresponding classes hold discernibility relation.

$$C_{ij}= \{a \in B \mid a(x) \neq a(x')\}$$

Here C_{ij} is the matrix value with i as row index and j as column index [6].

C. Reducts

Reduct is the result of attribute reduction process. It is the minimum set of attributes capable of discerning two objects. Thus a reduct is subset of attributes that is sufficient to preserve the complete information as that is provided by the entire attribute set [11].

D. Lower and Upper Approximation

The Rough Set Theory divides the space into three regions: the lower approximation, the upper approximation and the boundary region [7].

Let $X \subseteq U$ the target set to be represented by the attributes that are in $B \subseteq A$. X can be approximated by using the B-lower and B-upper approximation of X. B- lower approximation ($\underline{B}X$) is the set of objects that with full certainty are classified as member of set X using attributes of B and B-upper approximation ($\overline{B}X$) is the set of object that are possibly the members of set X.

$$\underline{B}X= \{x \mid [x]_B \subseteq X\},$$
$$\overline{B}X=\{x \mid [x]_B \cap X \neq \emptyset\}$$

The boundary region (BN_B) is the set difference of the lower approximation and upper approximation [8] [9].

$$BN_B = \overline{B}X-\underline{B}X$$

E. Rough Set

It is a set containing the two sets: upper approximation and lower approximation.

III. THE NSL-KDD DATASET

KDD stands for knowledge discovery in database. The NSL-KDD data set is basically an improvement over KDD'99 data set [13]. Though NSL-KDD is not perfect and does contain drawbacks but then also is one of the best options with people to compare the efficiency of their intrusion detection methods or systems.

Some of the problems solved by NSL-KDD data set that were there in KDD'99 data set are: NSL-KDD training set does not contain duplicate or redundant records. It consists of a good and reasonable proportion of various types of records [13] [14].

Each record in dataset contains forty-one attributes and one class or decision attribute as shown in Table I. Figure 1 shows a screenshot of the sample NSL-KDD dataset used as training set.

TABLE I NSL KDD DATASET ATTRIBUTES

No.	Features	No.	Features
1	duration	22	is_guest_login
2	protocol_type	23	count
3	service	24	srv_count
4	flag	25	serror_rate
5	src_bytes	26	srv_serror_rate
6	dst_bytes	27	rerror_rate
7	land	28	srv_rerror_rate
8	wrong_fragment	29	same_srv_rate
9	urgent	30	diff_srv_rate
10	hot	31	srv_diff_host_rate
11	num_failed_logins	32	dst_host_count
12	logged_in	33	dst_host__srv_count
13	num_compromised	34	dst_host_same_srv_rate
14	root_shell	35	dst_host_diff_srv_rate
15	su_attempted	36	dst_host_same_src_port_rate
16	num_root	37	dst_host_srv_diff_host_rate
17	num_file_creations	38	dst_host_serror_rate
18	num_shells	39	dst_host_srv_serror_rate
19	num_access_files	40	dst_host_rerror_rate
20	num_outbound_cmds	41	dst_host_srv_rerror_rate
21	is_host_login	42	class

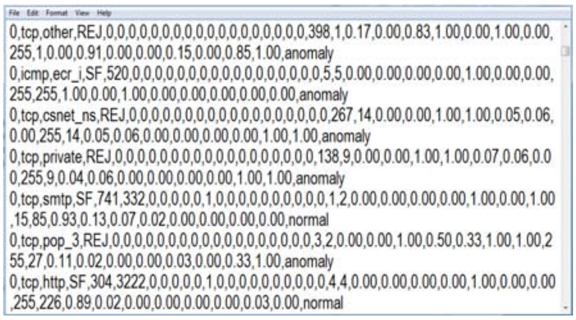

File Edit Format View Help
0,tcp,other,REJ,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,398,1,0.17,0.00,0.83,1.00,0.00,1.00,0.00,
255,1,0.00,0.91,0.00,0.00,0.15,0.00,0.85,1.00,anomaly
0,icmp,ecr_i,SF,520,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,5,5,0.00,0.00,0.00,0.00,1.00,0.00,0.00,
255,255,1.00,0.00,1.00,0.00,0.00,0.00,0.00,0.00,anomaly
0,tcp,csnet_ns,REJ,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,267,14,0.00,0.00,1.00,1.00,0.05,0.06,
0.00,255,14,0.05,0.06,0.00,0.00,0.00,0.00,1.00,1.00,anomaly
0,tcp,private,REJ,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,138,9,0.00,0.00,1.00,1.00,0.07,0.06,0.0
0,255,9,0.04,0.06,0.00,0.00,0.00,0.00,1.00,1.00,anomaly
0,tcp,smtp,SF,741,332,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,1,2,0.00,0.00,0.00,0.00,1.00,0.00,1.00
,15,85,0.93,0.13,0.07,0.02,0.00,0.00,0.00,0.00,normal
0,tcp,pop_3,REJ,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,3,2,0.00,0.00,1.00,0.50,0.33,1.00,1.00,2
55,27,0.11,0.02,0.00,0.00,0.03,0.00,0.33,1.00,anomaly
0,tcp,http,SF,304,3222,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,4,4,0.00,0.00,0.00,0.00,1.00,0.00,0.00
,255,226,0.89,0.02,0.00,0.00,0.00,0.00,0.03,0.00,normal

Figure 1 screen shot of the sample NSL KDD dataset

IV. PROPOSED SYSTEM

The proposed system works in two stages. The first is the feature reduction and the second is classification of the data packet as normal or anomaly. For feature reduction the proposed system uses rough set theory to obtain the discernibility matrix. The simplification of the obtained discernibility matrix results into the minimal reduct set that contain the minimum number of attributes enough to classify a packet as normal or attack [12].

The detailed description of the working of the system is explained with the help of an experiment performed using the rough set data analysis tool Rosetta. The experimental procedure is described in the fourth and fifth heading of the paper along with the obtained results.

V. SIMPLIFICATION OF THE DISCERNIBILITY MATRIX: REDUCT CONSTRUCTION

As mentioned before all the attributes describing the dataset is not required to categorize them as normal or attack. The main aim of our paper is not to propose just a system for intrusion detection but to present an improved intrusion detection system.

One way to do that is to reduce the time for detection. Working on the same direction we propose a system which intends to extract or select minimum number of attributes from the entire attribute set. This approach reduces the training time. By using the discernibility matrix and its simplification we obtained the minimal attribute set. We have implemented the following steps for this purpose.

- Divide the discernibility matrix into two parts, A and B.
- Absorb each element of A (current_A) by every non empty element of B (current_B) by the following procedure

  for i=0 to total_no_of_elements
  for j=i to total_no_of_elements
  if(current_A ≠ NULL) then
  if (current_B⊆current_A) then
  current_A=current_B
- select an attribute x ∈current_A
  A=current_A-x
  current_A=x;
- for i =0 tototal_no_of_elements
  for j=i to total_no_of_elements
  if(x ∈current_B)
  current_B=x;
  else
  current_B=current_B-x;
- Union of all attributes of the simplified matrix gives the required minimal reduct set.

By applying the above steps to 17633 training records 98 reducts obtained from Rosetta. Using these reducts we were able to reduce the 41 features to 27. The reduced set of attributes obtained is shown in Table II.

TABLE II REDUCED FEATURE SET

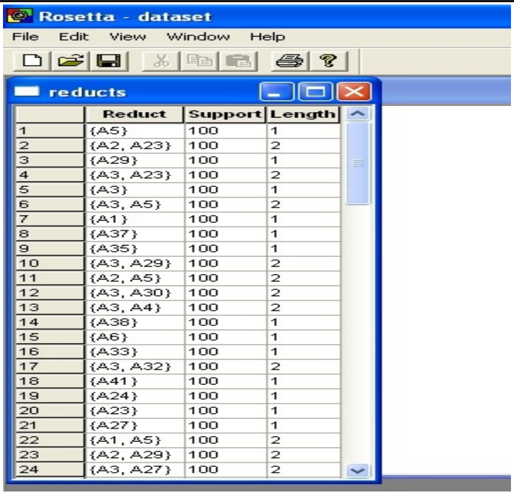
No.	Features	No.	Features
A1	duration	A29	same_srv_rate
A2	protocol_type	A30	diff_srv_rate
A3	service	A31	srv_diff_host_rate
A4	flag	A32	dst_host_count
A5	src_bytes	A33	dst_host__srv_count
A6	dst_bytes	A34	dst_host_same_srv_rate
A10	hot	A35	dst_host_diff_srv_rate
A11	num_failed_logins	A36	dst_host_same_src_port_rate
A23	count	A37	dst_host_srv_diff_host_rate
A24	srv_count	A38	dst_host_serror_rate
A25	serror_rate	A39	dst_host_srv_serror_rate
A26	srv_serror_rate	A40	dst_host_rerror_rate
A27	rerror_rate	A41	dst_host_srv_rerror_rate
A28	srv_rerror_rate		

VI. EXPERIMENTAL RESULTS

The experiment was performed on a system with following specifications: 260MB RAM, Intel Pentium 4 processor, Win-XP Operating System.

The experiment uses a rough set tool kit Rosetta for reducts generation and data analysis. The Rosetta system is a software system for inducing rough set based rule models [10]. The input to the Rosetta system is NSL KDD dataset.

The input is provided in a plain text format with the first row corresponding to the attributes. Every packet is characterized by 42 attribute including the decision attribute with value either normal or anomaly. The input data set is split into two parts (3:1). The first part is used as training set and the second part as test set.
The training set was reduced by using Johnson reduction algorithm. Johnson algorithm uses greedy search to find one reduct. The output of the reduction is a set of reducts and a set of If-Then-Else rule. The screenshot of the output is shown in Figure 2.



Rosetta - dataset

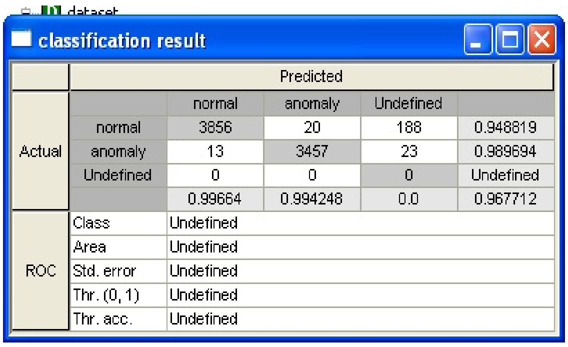
File Edit View Window Help

reducts

	Reduct	Support	Length
1	{A5}	100	1
2	{A2, A23}	100	2
3	{A29}	100	1
4	{A3, A23}	100	2
5	{A3}	100	1
6	{A3, A5}	100	2
7	{A1}	100	1
8	{A37}	100	1
9	{A35}	100	1
10	{A3, A29}	100	2
11	{A2, A5}	100	2
12	{A3, A30}	100	2
13	{A3, A4}	100	2
14	{A38}	100	1
15	{A6}	100	1
16	{A33}	100	1
17	{A3, A32}	100	2
18	{A41}	100	1
19	{A24}	100	1
20	{A23}	100	1
21	{A27}	100	1
22	{A1, A5}	100	2
23	{A2, A29}	100	2
24	{A3, A27}	100	2

Figure 2 Screenshot of reducts

For 17633 training records 98 reducts and 3122 rule sets were obtained. These rules were then applied to the test set. The screenshot of the result is shown in Figure 3.



classification result

		Predicted			
		normal	anomaly	Undefined	
Actual	normal	3856	20	188	0.948819
	anomaly	13	3457	23	0.989694
	Undefined	0	0	0	Undefined
		0.99664	0.994248	0.0	0.967712
ROC	Class	Undefined			
	Area	Undefined			
	Std. error	Undefined			
	Thr. (0, 1)	Undefined			
	Thr. acc.	Undefined			

Figure 3 Screenshot of classification result

The overall accuracy is 96.7712 percent. The sensitivity and accuracy for each class is shown in Table III.

TABLE III THE SENSITIVITY AND ACCURACY FOR EACH CLASS

	Normal	Anomaly	Undefined
Accuracy	99.664	99.424	0.0
Sensitivity	94.881	98.969	undefined

The sensitivity of normal class is 94.881 percent that mean that 3856+20+188 = 4064 actually belongs to normal class, out of which 3856 were correctly identified. The accuracy of normal class is 99.664 percent means that 3856+13 = 3869 was predicted as normal class, out of which 3856 were normal.

VII. CONCLUSION

The experiment shows that overall accuracy (accuracy plus sensitivity) of the proposed system with 27 reduced attributes is 96.77 percent. The sensitivity of normal class is 94.881 percent and the accuracy of normal class is 99.664 percent. Here we have used Rough Set Theory for feature reduction and classification.

In future we will extend this work to develop a fully deployable system with a user friendly Graphical User Interface for easier and faster detection of intrusions. Also we will try to increase the efficiency of proposed system using other classification algorithms.

REFERENCES

[1] [online] Available: http://www.thefreedictionary.com/intrusion

[2] Z. Pawlak, “Rough Sets”. In International Journal of Computer and Information Sciences, Vol.11, pp. 341-356, 1982.

[3] Silvia Rissino and Germano Torres, Rough Set Theory-fundamental concept, principles, data extraction and application, Data Mining and Knowledge Discovery in Real Life Applications, Book edited by: Julio Ponce and AdemKarahoca, ISBN 978-3-902613-53-0, I-Tech, Vienna, Austria pp. 438, February 2009.

[4] Cui-Juan Liu, “The Application of Rough Sets on Network Intrusion Detection”. In the proceedings of the 6th International Conference on Machine Learning and Cybernetics, Hong Kong, pp.19-22, Aug 2007.

[5] Available [online] http://en.wikipedia.org/wiki/Set_(mathematics)

[6] ChunhuaGu and Xueqin Zhang, “A Rough Set and SVM Based Intrusion Detection Classifier”, Second International Workshop on Computer Science and Engineering, 2009.

[7] K. Thangavel and A. Pethalakshmi,"Dimensionality reduction based on rough set theory: A review", Applied Soft Computing, 2009.

[8] Nikita Gupta and Shankar Lal, “Real Time Rough Set Behavior Cluster Modeling System”. In the proceeding of 5th Indian International Conference on Artificial Intelligence, SIT Tumkur, Dec 2011.

[9] Rough Set Wikipedia [online]. Available: http://en.wikipedia.org/ wiki/Rough_set.

[10] Torgeir R. Hvidsten, “A tutorial-based guide to the Rosetta system: A rough Set Toolkit for Analysis of Data”.

[11] Yan Zhao, Yiyu Yao and FengLuo, “Data Analysis Based on Discernibility and Indiscernibility”. In Information Sciences, 177, pp. 4959-4976, Elserbvier Inc., 2007.

[12] Y.Y. Yao and Y. Zhao, “Discernibility matrix simplification for constructing attribute reducts”. In Information Sciences, Vol. 179, No. 5, pp. 867-882, 2009.

[13] Available [online] http://iscx.ca/NSL-KDD/

[14] Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani, “A detailed Analysis of the KDD CUP 99 Data Set”. In the proceedings of the 2009 Symposium on Computational Intelligence in Security and Defense Application, 2009.