INTRUSION DETECTION DATA ANALYSIS USING DOMINANCE BASED ROUGH SET

Sanjiban Sekhar Roy, V. Madhu Viswanatham, P. Venkata Krishna

School of Computing Science and Engineering, VIT University, Vellore, Tamilnadu, India

***ABSTRACT*:** Being an extended part of the approach, known as classical rough set theory, today dominance based rough set approach has appeared as a useful mathematical device for dealing with uncertain data. The central theme of this paper is the analysis and evaluation of intrusion detection data set through the application of dominance based rough set approach.
KEYWORDS: Dominance based rough set, Intrusion Detection.

1. INTRODUCTION

Z. Pawlak had proposed the rough set theory for dealing with uncertain data [Paw82] in the year 1982. Rough Set Theory reduces the needed number of attribute values to produce a more compact decision rule set and increases efficiency. This theoretical framework is based on the concept that every object in the universe is attached with some kind of information. It includes algorithms for generation of rules, classification and reduction of attributes. It is hugely used for knowledge discovery and reduction of knowledge. Let, T = (U, A) and let $B \subseteq A$ and $X \subseteq U$, then we can approximate X by using the information contained in B by building the lower and upper approximations of X, represented $\underline{B}X$ and $\overline{B}X$ respectively, where

$$\underline{B}X = \{x \mid [x]_B \subseteq X\},$$

$$\overline{B}X = \{x \mid [x]_B \cap X \neq \phi\}$$

The accuracy of the approximation is given by,

$$\alpha_B(x) = \frac{card(\underline{B}(x))}{card(\overline{B}(x))}.$$

If $\alpha_B(X) = 1,$ then X is c is a crisp set or if $\alpha_B(X) < 1,$ then X is rough set.
In classical rough set theory the boundary region B of X is given by, $BN_B(X) = \overline{B}X - \underline{B}X$ consists of those objects that we cannot decisively classify in B. A set is called rough if its boundary region is non-empty, otherwise the set is crisp. If we assume, $c \in C.$, c is dispensable in T, if $POS_C(D) = POS_{(C-\{c\})}(D)$, otherwise attribute c is indispensable in T. The C-positive region of D: $POS_C(D) = \bigcup_{X \in U/D} \underline{C}X$. It is true that the rough set theory proposed by Z. Pawlak is used to solve many decision tribulations but is not able to find solutions in the cases where data are with inclination-ordered attribute domains and decision classes. Therefore, there is a need of multi-criteria decision analysis (MCDA) of rough set loom . The classical rough set is not sufficient to solve attributes with preference-ordered domains of uncertain data. To overcome rough set limitations Greco *et al* [GMS01, GMS99] introduced noble approach which is able to deal with inconsistencies typical to exemplary decisions in MCDA problems namely dominance based rough set approach. Dominance based rough set approach is an extension of classical rough set theory. Here our paper discusses on a systematic framework for analyzing inspection data of intrusion detection models using dominance based rough set technique. The resulting activity patterns of intrusion detection are then utilized to guide the selection of system features and used for construction of additional time-based statistical features for future learning. Classes based on these selected attributes are then computed (inductively learned) using the appropriate, formatted audit data. Here we have shown that classes can be introduced using dominance relation among conditional attributes used in an intrusion detection models since they can decide whether an observed system activity is "authentic" or "disturbing".

2. CORE CONCEPTS OF DOMINANCE BASED ROUGH SET

Z. Pawlak had proposed the rough set theory for solving many decision tribulations, but unfortunately it had badly failed to find solutions in the cases

where data are with inclination-ordered attribute domains and decision classes. As then the need for a multi-criteria decision analysis (MCDA) of rough set loom had appeared, henceforth, to overcome rough set limitations Greco *et al* [GMS01, GMS99] then had introduced a noble approach which is able to deal with inconsistencies typical to exemplary decisions in MCDA problems namely dominance based rough set approach. All conditional attributes are actually criteria's in multicriteria classification, which includes order of preference among its domain [GMS99]. In case of dominance based rough set approach outranking relation plays an important role. An outranking relation $\geq_q$ on U symbolize a fondness on the set of objects with respect to criterion *q* i.e. the logical meaning of $x \geq_q y$ is "x is at least good in comparison with y on the basis of criteria q". The same statement can be said in different way as x dominates y with the criteria q i.e. here $P \subseteq C$ and criteria q, $\forall q \in P$ which some times defined as $xD_p y$.

In multicriteria decision analysis there exists a preference order in the set of classes Cl .The approximation happens for upward and downward classes only.Let, us also consider the following upward and downward unions of classes, respectively,

$$Cl_t^{\geq} = \bigcup_{s \geq t} Cl_s \; ; \; Cl_t^{\leq} \bigcup_{s \leq t} Cl_s$$

where,

$Cl = \{Cl_t, t \in T\}$, $1 \leq t \leq n$ be a set of classes of U,

In dominance based rough set approach, a collection of entities dominating x, named as P dominating set can be given as $D_p^+(x) = \{y \in U : yD_P x\}$, and exactly the opposite, a collection of entities x, dominated by a set named as P Dominated Set is referred as

$$D_p^-(x) = \{y \in U : xD_P y\}$$

provided $P \subseteq C$ and $x \in U$,

Therefore, approximation values of P-lower and P-upper of $Cl_t^{\geq}$, where, $t \in T$, with respect to $P \subseteq C$ given as $\underline{P}(Cl_t^{\geq})$ and $\overline{P}(Cl_t^{\geq})$ correspondingly, which are as follows:

$$\underline{P}(Cl_t^{\geq}) = \{x \in U : D_p^+(x) \subseteq Cl_t^{\geq}\},$$

$$\overline{P}(Cl_t^{\geq}) = \bigcup_{x \in Cl_t^{\geq}} D_P^+(x) = \{x \in U : D_p^-(x) \cap Cl_t^{\geq} \neq \phi\}$$

Similarly, P-lower and P-upper approximations of $Cl_t^{\leq}$, $t \in T$, where, $P \subseteq C$ referred as $\underline{P}(Cl_t^{\leq})$ and $\overline{P}(Cl_t^{\leq})$ correspondingly, are given as :

$$\underline{P}(Cl_t^{\leq}) = \{x \in U : D_p^-(x) \subseteq Cl_t^{\leq}\},$$

$$\overline{P}(Cl_t^{\leq}) = \bigcup_{x \in Cl_t^{\leq}} D_P^-(x) = \{x \in U : D_p^+(x) \cap Cl_t^{\leq} \neq \phi\}$$

Also if the above properties holds for dominance based rough set then the following properties also holds

$$\underline{P}(Cl_t^{\geq}) \subseteq Cl_t^{\geq} \subseteq \overline{P}(Cl_t^{\geq}) \; ;$$
$$\underline{P}(Cl_t^{\leq}) \subseteq Cl_t^{\leq} \subseteq \overline{P}(Cl_t^{\leq})$$

along with the it's complimentary properties:

$$\underline{P}(Cl_t^{\geq}) = U - \overline{P}(Cl_{t-1}^{\leq}), \; t=2,\ldots,n$$

$$\underline{P}(Cl_t^{\leq}) = U - \overline{P}(Cl_{t+1}^{\geq}), \; t=1,\ldots,n-1$$

$$\overline{P}(Cl_t^{\geq}) = U - \underline{P}(Cl_{t-1}^{\leq}), \; t=2,\ldots,n$$

$$\overline{P}(Cl_t^{\leq}) = U - \underline{P}(Cl_{t+1}^{\geq}), \; t=1,\ldots,n-1$$

Therefore the P-doubtful regions of $Cl_t^{\geq}$ and $Cl_t^{\leq}$ are definited as:

$$Bn_P(Cl_t^{\geq}) = \overline{P}(Cl_t^{\geq}) - \underline{P}(Cl_t^{\geq}),$$

$$Bn_P(Cl_t^{\leq}) = \overline{P}(Cl_t^{\leq}) - \underline{P}(Cl_t^{\leq}),$$

The correctness of approximation of $Cl_t^{\geq}$ and $Cl_t^{\leq}$ for all $t \in T$ and for any $P \subseteq C$, is defined as

$$\alpha_P(Cl_t^{\geq}) = \left| \frac{\underline{P}(Cl_t^{\geq})}{\overline{P}(Cl_t^{\geq})} \right|,$$

$$\alpha_P(Cl_t^{\leq}) = \left| \frac{\underline{P}(Cl_t^{\leq})}{\overline{P}(Cl_t^{\leq})} \right|.$$

and the ratio

$$\gamma_P(Cl) = \left| \frac{U - ((\bigcup_{t \in T} Bn_p(Cl_t^{\geq})) \bigcup (\bigcup_{t \in T} Bn_P(Cl_t^{\leq})))}{U} \right|$$

known as the quality of approximation of the partition Cl by the set of criteria P or briefly quality of sorting. Therefore, γ_P ratio is the relation among the P-correctly classified substance and the objects in the table.
The definition of reduct of C with respect to class Cl is each minimal subset $P \subseteq C$ such that $\gamma_P(Cl) = \gamma_C(Cl)$ and is avowed by $RED_{Cl}(P)$. Therefore, a data table can have many reducts. $CORE_{Cl}$ is the intersection of their reducts.

3. INVESTIGATIONAL OUTCOME VIA DOMINANCE BASED ROUGH SET APPROACH

The dominance based rough set approach has been efficiently applied for analyzing and evaluating intrusion detection data set. Here we have shown such application by using the following data taken from paper [LSM99]. We can name this data table as "connection records of a network". Classes based on these selected attributes are then computed (inductively learned) using the appropriate, formatted audit data. Here, we have shown that classes can be introduced using dominance relation among conditional attributes used in an intrusion detection models since they can decide whether an observed system activity is "authentic" or "disturbing". Here the attack model we have shown includes short sequence of connection of records of intrusions evidence. To see which ports are easy to get to invader analytically makes links to each port (service) of a intention host (target host). In the connection records, there should be a host (or hosts) that receives many connections to its "different" ports in a short period of time. There can be links of "REJ" flag as numerous ports are by and large not available as nearby are several patterns to facilitate the proposal of the attack, e.g (destination host = 207.217.205.23, flag = REJ). Therefore the destination host and FLAG value constitute an attack. We have shown what the minimum set is of attributes which summaries the following data table for intrusion detection.

Table 1: connection records of a network

Sl. No	Clock	A_1	A_2	A_3	A_4	A_5
1	1.0	30	telnet	150	500	REJ
2	1.6	25	http	300	2500	SF
3	2.4	5	Smtp	200	2500	SF
4	3.0	25	telnet	200	3000	SF
5	3.5	30	telnet	300	2000	SF
6	4.0	30	http	150	1000	REJ
7	4.2	5	http	150	1000	REJ
8	4.5	30	Smtp	300	1000	REJ
9	4.9	25	Smtp	150	3000	SF
10	5.0	5	Smtp	150	500	REJ
11	5.2	5	Smtp	200	2500	REJ
12	5.5	25	telnet	300	3500	REJ

Here, set Q and P contains the following attributes.

Q={A_1,A_2,A_3,A_4,A_5}

P={A_1,A_2,A_3,A_4}

Attribute A1 to A4 called as conditional attributes and attribute A5 is decision attribute known as signal flag. Here, the third column refers to the arrival time known as timestamp of the packet in the data table, therefore attribute A1 contains the duration of each raw packet. Thereafter, attributes A3, A4 contains services (e.g http, smtp, telnet), source byte and destination bytes of the raw packets. According to value of all the conditional attributes data packet is rejected (REJ) or successfully accepted (SF) by the network. Now using dominance based rough set approach we will approximate the class $Cl_1^{\leq}$ of "atmost REJ" and the class $Cl_2^{\geq}$ of "atleast SF". As we know $P \subseteq C$, therefore we have taken all the

attribute combinations to find the γ_P values of all the subsets[KI10].

1)C={A1, A2 }

$\underline{C}(Cl_1^{\leq}) = \{7\}$

$\overline{C}(Cl_1^{\leq}) = \{1,2,3,4,5,6,7,8,9,10,11,12\}$

$Bn_C(Cl_1^{\leq}) = \{1,2,3,4,5,6,8,9,10,11,12\}$

$\underline{C}(Cl_2^{\geq}) = \phi$

$\overline{C}(Cl_2^{\geq}) = \{1,2,3,4,5,6,8,9,10,11,12\}$

$Bn_C(Cl_2^{\geq}) = \{1,2,3,4,5,6,8,9,10,11,12\}$

$\gamma_P(Cl) = 1/12$

2)C={A2,A3}

$\underline{C}(Cl_1^{\leq}) = \{6,7\}$

$\overline{C}(Cl_1^{\leq}) = \{1,2,3,4,5,6,7,8,9,10,11,12\}$

$Bn_C(Cl_1^{\leq}) = \{1,2,3,4,5,8,9,10,11,12\}$

$\underline{C}(Cl_2^{\geq}) = \phi$

$\overline{C}(Cl_2^{\geq}) = \{1,2,3,4,5,6,7,8,9,10,11,12\}$

$Bn_C(Cl_2^{\geq}) = \{1,2,3,4,5,8,9,10,11,12\}$

$\gamma_P(Cl) = 1/6$

3)C={A1 ,A3}

$\underline{C}(Cl_1^{\leq}) = \{7,10\}$

$\overline{C}(Cl_1^{\leq}) = \{1,2,3,4,5,6,7,8,9,10,11,12\}$

$Bn_C(Cl_1^{\leq}) = \{1,2,3,4,5,6,8,9,11,12\}$

$\underline{C}(Cl_2^{\geq}) = \phi$

$\overline{C}(Cl_2^{\geq}) = \{1,2,3,4,5,6,7,8,9,10,11,12\}$

$Bn_C(Cl_2^{\geq}) = \{1,2,3,4,5,6,8,9,11,12\}$

$\gamma_P(Cl) = 1/6$

4)C={A1,A4}

$\underline{C}(Cl_1^{\leq}) = \{1,6,7,8,10\}$

$\overline{C}(Cl_1^{\leq}) = \{1,2,3,4,6,7,8,9,10,11,12\}$

$Bn_C(Cl_1^{\leq}) = \{2,3,4,9,11,12\}$

$\underline{C}(Cl_2^{\geq}) = \{5\}$

$\overline{C}(Cl_2^{\geq}) = \{2,3,4,5,9,11,12\}$

$Bn_C(Cl_2^{\geq}) = \{2,3,4,9,11,12\}$

$\gamma_P(Cl) = 1/2$

5)C={A2,A4}

$\underline{C}(Cl_1^{\leq}) = \{1,6,7,8,10\}$

$\overline{C}(Cl_1^{\leq}) = \{1,2,3,4,5,6,7,8,9,10,11,12\}$

$Bn_C(Cl_1^{\leq}) = \{2,3,4,5,9,11,12\}$

$\underline{C}(Cl_2^{\geq}) = \phi$

$\overline{C}(Cl_2^{\geq}) = \{2,3,4,5,9,11,12\}$

$Bn_C(Cl_2^{\geq}) = \{2,3,4,5,9,11,12\}$

$\gamma_P(Cl) = 5/12$

6)C={A3,A4}

$\underline{C}(Cl_1^{\leq}) = \{1,6,7,8,10\}$

$\overline{C}(Cl_1^{\leq}) = \{1,2,3,4,5,6,7,8,9,10,11,12\}$

$Bn_C(Cl_1^{\leq}) = \{2,3,4,5,9,11,12\}$

$\underline{C}(Cl_2^{\geq}) = \phi$

$\overline{C}(Cl_2^{\geq}) = \{2,3,4,5,9,11,12\}$

$Bn_C(Cl_2^{\geq}) = \{2,3,4,5,9,11,12\}$

$\gamma_P(Cl) = 5/12$

7)C={A 1,A2 ,A3}

$\underline{C}(Cl_1^{\leq}) = \{6,7,10\}$

$\overline{C}(Cl_1^{\leq}) = 1,2,3,4,6,7,8,9,10,11,12\}$

$Bn_C(Cl_1^{\leq}) = \{1,2,3,4,8,9,11,12\}$

$\underline{C}(Cl_2^{\geq}) = \{5\}$

$\overline{C}(Cl_2^{\geq}) = \{1,2,3,4,5,8,9,11,12\}$

$Bn_C(Cl_2^{\geq}) = \{1,2,3,4,8,9,11,12\}$

$\gamma_P(Cl) = 1/3$

8)C={A 2,A3 ,A4}

$\underline{C}(Cl_1^{\leq}) =$ **{1,6,7,8,10}**

$\overline{C}(Cl_1^{\leq}) = \{1,2,3,4,5,6,7,8,9,10,11,12\}$

$Bn_C(Cl_1^{\leq}) = \{2,3,4,5,9,11,12\}$

$\underline{C}(Cl_2^{\geq}) = \phi$

$\overline{C}(Cl_2^{\geq}) = \{2,3,4,5,9,11,12\}$

$Bn_C(Cl_2^{\geq}) = \{2,3,4,5,9,11,12\}$

$\gamma_P(Cl) = 5/12$

9)C={A 1,A3 ,A4}

$\underline{C}(Cl_1^{\leq}) = \{1,6,7,8,10\}$

$\overline{C}(Cl_1^{\leq}) = \{1,2,3,4,5,6,7,8,9,10,11,12\}$

$Bn_C(Cl_1^{\leq}) = \{2,3,4,5,9,11,12\}$

$\underline{C}(Cl_2^{\geq}) = \phi$

$\overline{C}(Cl_2^{\geq}) = \{2,3,4,5,9,11,12\}$

$Bn_C(Cl_2^{\geq}) = \{2,3,4,5,9,11,12\}$

$\gamma_P(Cl) = 5/12$

10)C={A 1,A2 ,A4}

$\underline{C}(Cl_1^{\leq}) = \{1,6,7,8,10\}$

$\overline{C}(Cl_1^{\leq}) = \{1,2,3,4,6,7,8,9,10,11,12\}$

$Bn_C(Cl_1^{\leq}) = \{2,3,4,9,11,12\}$

$\underline{C}(Cl_2^{\geq}) = \{5\}$

$\overline{C}(Cl_2^{\geq}) = \{2,3,4,5,9,11,12\}$

$Bn_C(Cl_2^{\geq}) = \{2,3,4,9,11,12\}$

$\gamma_P(Cl) = 1/2$

11)C={A 1,A2 ,A3,A4}

$\underline{C}(Cl_1^{\leq}) = \{1,6,7,8,10\}$

$\overline{C}(Cl_1^{\leq}) = \{1,2,3,4,6,7,8,9,10,11,12\}$

$Bn_C(Cl_1^{\leq}) = \{2,3,4,9,11,12\}$

$\underline{C}(Cl_2^{\geq}) = \{5\}$

$\overline{C}(Cl_2^{\geq}) = \{2,3,4,5,9,11,12\}$

$Bn_C(Cl_2^{\geq}) = \{2,3,4,9,11,12\}$

$\gamma_P(Cl) = 1/2$

Hence the attribute sets **{A1,A4}** and **{A 1,A2 ,A3}** of **{A 1,A2 ,A3,A4}** are reducts. We know that intersection of reducts is the CORE and here the CORE is attribute **A1**

We have also made an effort to search out the decision rules from the table given above by applying dominance based rough set. The purpose has been fulfilled by short listing the following observations from the above mentioned table.

Table 2: observations followed from table 1

SL NO	CLOCK	Time stamp (A_1)	Service type (A_2)	SB (A_3)	DB (A_4)	Signal Flag (A_5)
2	1.6	25	http	300	2500	SF
4	3.0	25	telnet	200	3000	SF
5	3.5	30	telnet	300	2000	SF
9	4.9	25	smtp	150	3000	SF

Decision rule 1 :
If (x, A_1) ≥ 25 & (x, A_4) ≥ 2000, then it belongs to CLASS $CL_1^{\leq}$.

Table 3: observations followed from table 1

SL NO	CLOCK	Time stamp (A_1)	Service type (A_2)	SB (A_3)	DB (A_4)	Signal Flag (A_5)
1	1.0	30	telnet	150	500	REJ
6	4.0	30	http	150	1000	REJ
7	4.3	5	http	150	1000	REJ
8	4.5	30	smtp	300	1000	REJ
10	5.0	5	smtp	150	500	REJ

Decision rule 2:
If (x, A_1) ≥ 5 & (x, A_4) ≤ 1000, then it belongs to CLASS $CL_2^{\geq}$.

Table 4: observations followed from table 1

SL NO	CLOCK	Time stamp (A_1)	Service type (A_2)	SB (A_3)	DB (A_4)	Signal Flag (A_5)
3	2.4	5	smtp	200	2500	SF
11	5.2	5	smtp	200	2500	REJ
12	5.5	25	telnet	300	3500	REJ

Decision rule 3:
If {(x, A_1) ≥ 5 & (x, A_1) ≤ 5 } & {(x, A_4) ≥ 2500 & (x, A_4) ≤ 2500} , then it belongs to $CL_1^{\leq}$ U $CL_2^{\geq}$.
Likewise we have found "Decision rule 4" which is -
If {(x, A_1) ≥ 25 & (x, A_1) ≤ 25} & {(x, A_4) ≥ 3000 & (x, A_4) ≤ 3000}, then it belongs to $CL_1^{\leq}$ U $CL_2^{\geq}$.

CONCLUSION

This paper unveils an argument on a systematic framework for analyzing inspection of intrusion detection data set using dominance based rough set technique. Here the attack model we have shown

includes short sequence of connection of records of intrusions evidence. We have even shown the CORE and accuracy of the data table using dominance based rough set approach. And finally, we have found the decision rules following certain observations of intrusion detection data set by applying dominance based rough set approach.

REFERENCES

[GMS01] **Salvatore Greco, Benedetto Matarazzo, Roman Slowinski** - *Rough sets theory for multicriteria decision analysis*. European Journal of Operational Research, pp 1-47, 2001

[GMS99] **Salvatore Greco, Benedetto Matarazzo, Roman Slowinski** - *Rough approximation of a preference relation by dominance relations*, European Journal of Operational Research, vol 117 , pp. 63-83, 1999

[KI10] **Yoshifumi Kusunoki, E Masahiro Inuiguchi** - *A unified approach to reducts in dominance-based rough set*, Soft Computing.14, pp.507–515, 2010

[LSM99] **Wenke Lee, Salvatore J. Stolfo, Kui W. Mok** - *A Data Mining Framework for Adaptive Intrusion Detection*. In Proceedings of the 1999 IEEE Symposium on Security and Privacy

[Paw82] **Z. Pawlak** – *Rough Sets*. International Journal of Computer and Information Sciences, 11, pp. 341-356, 1982