ICIC Express Letters
Volume **17**, Number **1**, January **2023**

 ISSN 1881-803X
pp. **73–80**

DETECTING ANOMALY USING NEIGHBORHOOD ROUGH SET BASED CLASSIFICATION APPROACH

Fokrul Alom Mazarbhuiya

School of Fundamental and Applied Sciences
Assam Don Bosco University
Tepesia Garden, P. O. Sonapur, Guwahati, Assam 782402, India
fokrul.mazarbhuiya@dbuniversity.ac.in

Received April 2022; accepted June 2022

Abstract. *A vital field of research is network anomaly detection. Several research papers in this topic have previously been published, using various methodologies or algorithms. The classification-based approach is intriguing. The majority of previous techniques assumed that the datasets under investigation were numeric. Nonetheless, up until recently, some success had been made in mixed datasets. In practice, datasets are typically mixed, having numeric, category, and other types of features. Defining a metric that works on all types of attributes in such datasets is a difficult issue. In this paper, we attempt to address the problem and provide a method for detecting anomalies in network datasets that use a rough set-based classification approach. The method's effectiveness is demonstrated by an experiment utilizing the KDD Cup'99 dataset.*
Keywords: Intrusion detection, Information systems, Lower and upper approximation of set, Certain rules, Possible rules, Boundary regions

1. **Introduction.** Many knowledge discovery applications require finding or detection of anomalies from datasets. Anomaly detection is an emerging research area which caught the attention of many researchers. It has been applied in many areas like fraud detection in banking or financial transactions, fault finding in manufacturing, and intrusion detection in computer network. The goal of anomaly detection [1,2] is to identify the data object which does not conform to a well-defined normal behavior. Intrusion Detection Systems (IDSs) are the security tools for preventing the systems or network from the illegitimate action that can jeopardize the integrity, privacy or accessibility. In general, there exist two categories of IDS viz. *anomaly detection-based* and *signature recognition-based schemes*. The former used to discover the network's misuse and computer's misuse or intrusions by keeping track of the systems and then classifying the activities into normal or anomalous. The consequent system is called anomaly-based intrusion detection system [3,4]. Anomaly-based intrusion detection can be effectively applied as a risk mitigation tool for computer and associated network.

Many anomaly detection techniques have been proposed over the last few decades. The classification-based technique is one of them. The classification [5], also known as supervised learning, is a data mining tool to categorize the objects into pre-defined classes. In the last few years, classification has been applied in many areas like, anomaly detection, fraud identification, pattern recognition, and prediction. A classification using automatic labeling technique is applied in cellular network for anomaly detection [6]. In [7], the authors have described the anomaly detection results with single and multidimensional data sets using the negative selection algorithm. In [8], the authors have reviewed many available methods of anomaly detection in categorical datasets. We have attempted to

DOI: 10.24507/icicel.17.01.73

introduce the problem of mixed attribute in the current study, which has mostly concentrated on numeric characteristics with mixed properties receiving less attention. Two clustering-based anomaly detection techniques are proposed in [9,10], which take account of both numeric and categorical properties of network datasets. We have used the dataset KDD Cup'99 Network Anomaly Dataset [11], which has both numeric and categorical attributes, in this article.

Pawlak [12] introduced in 1980s, the rough set theory, to address uncertainty and vagueness that exist in any datasets. In [13], the rough set-based classification is applied nicely to discrete datasets which uses the properties of equivalence relation. In [14], authors have discussed a classification technique based on a neighborhood rough set for handling medical diagnosis. In [15], the authors have proposed a rough set-based fuzzy discretization approach for outlier detection. Liu et al. [16] experimentally showed that their method can efficiently detect anomalies by reducing the sizes of the datasets.

In this paper, we propose an anomaly detection method for network data having numeric, and categorical attributes. The proposed algorithm uses the properties of the neighborhood rough set theory for finding classification rules which can be used for anomaly detection. The objective of this work is three folds. First of all, a unified metric function is defined which works on different types of attributes viz. numeric or categorical or both. For the numeric attributes, we have used Canberra metric [17,18] and for the categorical attributes we have used a formula which is a slight variation of the one given in [10]. Then a weighted average is taken for calculating the unified metric. Secondly, using the above-mentioned metric, two relations, viz. a neighborhood relation on conditional attributes with respect to a user specified parameter and the equivalence relation with reference to decision attributes, are constructed. Lastly, the lower and upper approximation spaces and boundary regions are obtained using the above-mentioned relations. Then the classification rules are extracted from the aforesaid regions. We have tested our method with the help of KDD Cup'99 Network Anomaly Dataset [11] and made comparative studies with Liu et al.'s work [16] and we have found the promising performance of our method in terms of detection rate accuracy.

The article is arranged in the following way. The recent developments in this area are explained in Section 2. In Section 3, we discuss the terms, notations and definitions that have been used here. In Section 4, we explain the proposed system using a flowchart. The results and findings of experimental studies are given in Section 5, and finally, we wind up the paper with conclusions given in Section 6.

2. **Related Works.** Anomaly detection is one of the core research areas of artificial intelligence. In [1,2], the authors have proposed clustering-based methods of finding anomalies. The detailed reviews of anomaly-based intrusion detection systems are given in [3,4]. In [6], researchers have given an anomaly detection scheme in cellular network with the help of classification based automatic labeling technique. In [19], the authors have proposed an anomaly detection method of general data. A model which is a fusion of two deep neural networks that has been used for anomaly detection is discussed in [20]. In [21], the authors have proposed a method for detecting surface defects of wind turbine blades. In [9], the authors have proposed an agglomerative hierarchical approach for the identification of anomaly from network data with mixed attributes. In [10], the authors have proposed a mixed approach consisting of both partitioning and hierarchical algorithms for the anomaly detection in mixed data. In [22], the authors have made detailed studies on methods and approaches of one-class classification along with their domain-specific applications. In [23], the authors have proposed a classification-based method called Logsy, that differentiates the normal system log data from anomalous samples in auxiliary log datasets. In [24], the authors have proposed an absolute measurement anomaly detection to constrain the distribution of each output in the classification network.

In [12], the author has introduced rough set theory to address uncertainty or/and vagueness available in any datasets. In [14], the authors have presented a neighborhood rough set-based classification algorithm for medical diagnosis. In [25], the authors have proposed two machine learning tools, namely rough set and K-nearest neighbor to be used for intrusion detection. In [26], the authors have presented a method to develop an on-line intrusion detection system using modified Q-learning and rough set theory. In [27], the authors have suggested an approach which is the combination of rough set theory and fuzzy c-means algorithm for the detection anomalies in data. An approach involving neighborhood fuzzy rough set theory to rank outlier according to fuzzy membership function computed in rough approximation space is presented in [28].

In [29], a rough set based anomaly detection method is proposed which efficiently detects masqueraders in mobile networks. In [30], an adaptive IDS based on fuzzy rough sets for attribute selection and Allen's interval algebra is proposed for the effective prediction of attacks in WSNs. In [31], the authors have analyzed the various features of KDD Cup'99 [11] and for finding optimal features and applied the notion of reduct and core to improving the anomaly detection rate. In [16], the authors have applied a distance function based on rough set theory which is then used to find anomalies efficiently. They have conducted experiments with KDD Cup'99 dataset [11] and claimed that their method has reduced the execution significantly.

In [17,18], they have discussed a distance function called Canberra metric which efficiently finds the distance in numerical data. In [10], the authors have used a nice distance formula to find similarity or distance measure in categorical attributes. In this article, we propose to use a neighborhood rough set-based classification approach to identify anomalies in the network data with categorical and numeric attributes. A unified metric which works on both types of attributes is the weighted average of Canberra metric [17,18] and a slight variation of the metric given in [10].

3. **Problem Definitions.** The objective of this work is to find anomalies from the data with mixed attributes, for instance, the dataset KDD Cup'99, a network data [11] which is a refined version of DARPA'98 [32]. Every connection instance of the aforesaid dataset has 41 properties; among these 38 are numeric and the rest are flag properties. Hence, we treat the dataset as a hybrid or mixed dataset considering the attributes of both numeric and categorical nature. For finding a neighbourhood relation, we need to define a unified metric function which must work on the hybrid dataset. In beneath we discuss some definitions employed in the paper.

Let $U = \{x_1, x_2, \ldots, x_m\}$ be the set of data instances, where each x_i is an n-dimensional vector consisting of k-numeric and $(n-k)$-categorical attributes.

3.1. **Unified metric function.** To define the unified metric function, we proceed as follows. Let $A = \{A_1, A_2, \ldots, A_k, A_{k+1}, A_{k+2}, \ldots, A_n\}$ be the set of attributes where first k are numeric and the rest are categorical. Let x, y be the two data instances such that $x = \{A_1(x), A_2(x), \ldots, A_k(x), A_{k+1}(x), A_{k+2}(x), \ldots, A_n(x)\}$ and $y = \{A_1(y), A_2(y), \ldots, A_k(y), A_{k+1}(y), A_{k+2}(y), \ldots, A_n(y)\}$, where $A_i(x)$ is the value of the data instance x in the attribute A_i and $A_i(y)$ is the value of the data instance y in the attribute A_i. Then the unified metric between x and y is given by the formula [9]

$$d(x,y) = \frac{kd_1(x,y) + (n-k)d_2(x,y)}{n} \tag{1}$$

where $d_1(x,y)$ is the metric function on the numeric attributes and $d_2(x,y)$ is that on the categorical attributes.

3.2. **Metric defined on numeric attributes.** Let $x = (A_1(x), A_2(x), \ldots, A_k(x))$ and $y = (A_1(y), A_2(y), \ldots, A_k(y))$ be two k-dimensional vectors. Then the Canberra metric [17,18], $d_1(x, y)$, is stated as follows.

$$d_1(x, y) = \frac{1}{k} \sum_{i=1}^{k} \frac{|A_i(x) - A_i(y)|}{|A_i(x) + A_i(y)| + \varepsilon} \tag{2}$$

If $A_i(x) = A_i(y) = 0$, for all $i = 1, 2, \ldots, k$, then Equation (2) will be 0/0 form so we introduce a small positive number ε in (2) to keep the denominator non-zero. Obviously, $d_1(x, y) \in [0, 1]$. Thus, $d_1(x, y) = 0$, then x and y are having similar attribute values.

3.3. **Metric defined on categorical attributes.** A metric for categorical attribute is described as follows. Let us suppose that the data instances set has categorical attributes $A_1, A_2, \ldots, A_d$ with a finite, unordered set of possible values as their domain $= \{v_{i1}, v_{i2}, \ldots, v_{im}\}$ that each A_i can take. Also for any $v_1, v_2 \in \text{dom}(A_i)$, either $v_1 = v_2$ or $v_1 \neq v_2$. Any data instance x is a vector $(A_1(x), A_2(x), \ldots, A_d(x))^{/}$, where $A_i(x) \in \text{dom}(A_i)$, $i = 1, 2, \ldots, d$. The distance $d_2(x, y)$ between data instances x and y is given by [10]

$$d_2(x, y) = \frac{\sum_{p=1}^{d} c(A_i(x), A_i(y))}{d} \tag{3}$$

where $c(A_i(x), A_i(y)) = \begin{cases} 1, & \text{if } A_i(x) = A_i(y) \\ 0, & \text{otherwise} \end{cases}$.

Obviously, $d_2(x, y) \in [0, 1]$ means $d_2(x, y) = 1$ only if the data instance x and y for which $A_i(x) = A_i(y)$; $i = 1, 2, \ldots, d$ and $d_2(x, y) = 0$ only if $A_i(x) \neq A_i(y)$; $i = 1, 2, \ldots, d$. Using Equations (2) and (3) in (1), we can get the formula for unified metric function $d(x, y)$, which then is used for defining neighborhood relation.

3.4. ***θ*-neighborhood relation.** For the data instances $x_i \in U$, $\forall i$ and $0 \leq \theta \leq 1$, a θ-neighborhood relation (U, d) is defined in [14] as $\theta(x_i) = \{x; d(x_i, x) \leq \theta\}$.

3.5. **Neighborhood Decision System (NDS).** Let us consider a decision system $(U, C \cup D)$ with U, the data instance set (universe), C, the conditional attributes set and D, the decision attributes set. For $\theta < 1$, a θ-neighborhood relation N is generated by C, [14] and is characterized by NDS $= (U, C \cup D, \theta)$.

3.6. **Lower and upper approximations.** Let $B \subseteq C$, for any arbitrary $X \subseteq U$, the lower approximation and upper approximation of X in terms of the relation N with respect to B are characterized respectively by [14]

$$\underline{N_B}(X) = \{x : \theta_B(x) \subseteq X, x \in U\} \tag{4}$$

$$\overline{N_B}(X) = \{x : \theta_B(x) \cap X \neq \phi, x \in U\} \tag{5}$$

where

$$\theta_B(x) = \{y : d(B(x), B(y)) \geq \theta, y \in U\} \tag{6}$$

Here $B(x)$ is sub-vector of $C(x)$ having all those values of the attributes belonging to $B \subseteq C$. The boundary region of D in regard to B is defined as

$$Boundary(D) = \overline{N_B}(X) - \underline{N_B}(X) \tag{7}$$

The neighborhood lower approximation also known as positive region is the union of the lower approximation of each D class. The boundary region is needed to decrease the uncertainty in decision making process.

4. **Proposed Algorithm.** For finding the classification rules, first of all we take a suitable value for θ to generate θ-neighborhood relation. The unified metric that shapes the relation is defined in Section 3. To generate the classification rules, we proceed as follows: We have a set of data instances, each of which is described by n-attribute values (numeric or categorical) and is expressed as an $m \times n$ matrix $[x_{ij}]$, where x_{ij} is the jth attribute values for the ith data instance, $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$. In general, the supervised datasets can be viewed as $(U, C \cup D)$, where $U = \{x_1, \ldots, x_m\}$, $C = \{a_1, \ldots, a_n\} =$ the conditional attributes set, and $D = \{d_1, \ldots, d_n\} =$ the decision attributes set. The method is described below.

The first step of the proposed method is to compute the neighborhood relation of the conditional attribute using the unified metric given in Section 3, and compute the equivalence classes of decision attributes. The metric used for this purpose is given in (1). Then, using "And" operator and decision class, we construct neighborhood lower and upper approximation. The union of the lower approximation space of each decision class will give us the neighborhood rough set lower approximation space of the decision attributes and the boundary region is obtained from various decision classes. With the help of neighborhood approximation, two sets of decision rules namely the deterministic (certain) rules as well as the non-deterministic (possible) rules, can be generated. Applying the lower approximation of neighborhood rough sets, certain rules can be produced. Similarly, applying the upper approximation of neighborhood rough sets, possible rules can be generated. The proposed method is also explained with the help of flowchart given in Figure 1 below.

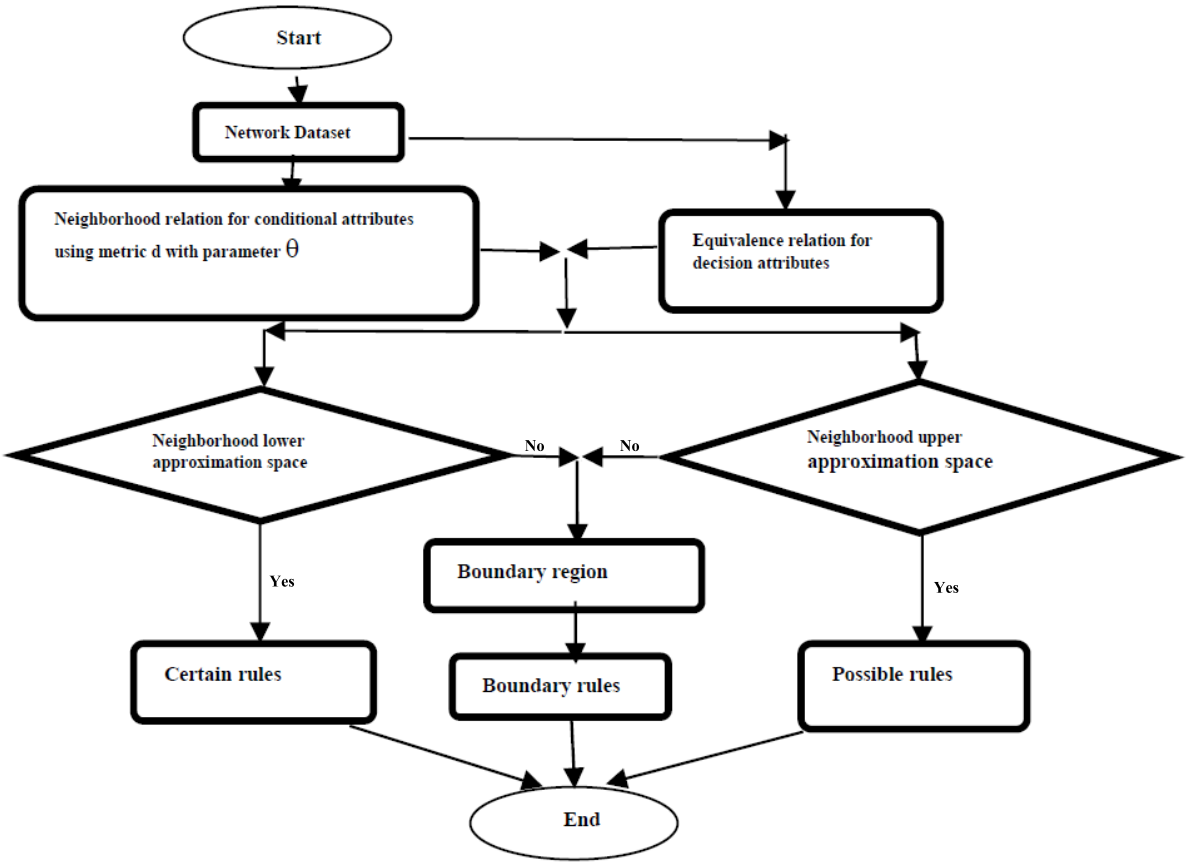

FIGURE 1. Flowchart of the proposed algorithm

5. **Experimental Settings.** To conduct experimental studies, we have chosen the KDD Cup'99 Network Anomaly Datasets [11]. The dataset with characteristics is furnished in Table 1.

Table 1. Dataset characteristics

Dataset	No. of attacks	No. of data instances	No. of numeric attributes	No. of categorical attributes
KDD Cup'99	20	4.9 million	38	3

We have implemented our algorithm (NRS classification algorithm) along with Liu et al.'s algorithm [16] in MatLab. Using the dataset [11], we have constructed classifiers which classify any new instance as normal traffic or attack. We have recorded the results of both the algorithms for different sets of attributes. The amount of data instances of numerous attacks and normal data are highly in disproportion. For the estimation of the methods performance, we have used parameters such as True Positive Rate (TPR), and False Positive Rate (FPR). A partial view of the results of the two algorithms describing the comparative analysis of Normal True Positive Rate (TPR), Attack True Positive Rate (TPR), Normal False Positive Rate (FPR), Attack False Positive Rate (FPR) is presented in Table 2.

Table 2. Normal vs Attack TPR/FPR

Algorithm	Normal TPR	Attack TPR	Normal FPR	Attack FPR	Avg TPR	Avg FPR
NRS classification algorithm	0.9998	0.9998	0.0002	0.0002	0.9999	0.0002
Liu et al.'s algorithm	0.9425	0.9425	0.06	0.06	0.94665	0.0533

Also, the comparative analysis of the two algorithms in terms of detection rates for the different sizes of attribute sets is given as a bar diagram in Figure 2. It has been observed and inferred from the results that Liu et al.'s algorithm [16] performs better for less number of attributes. However, the overall performance of our algorithm is steady and much better as it classifies both normal and attack category with almost 99% TPR and 1% FPR. Secondly, our algorithm detects anomalies more accurately than Liu et al.'s algorithm [16].

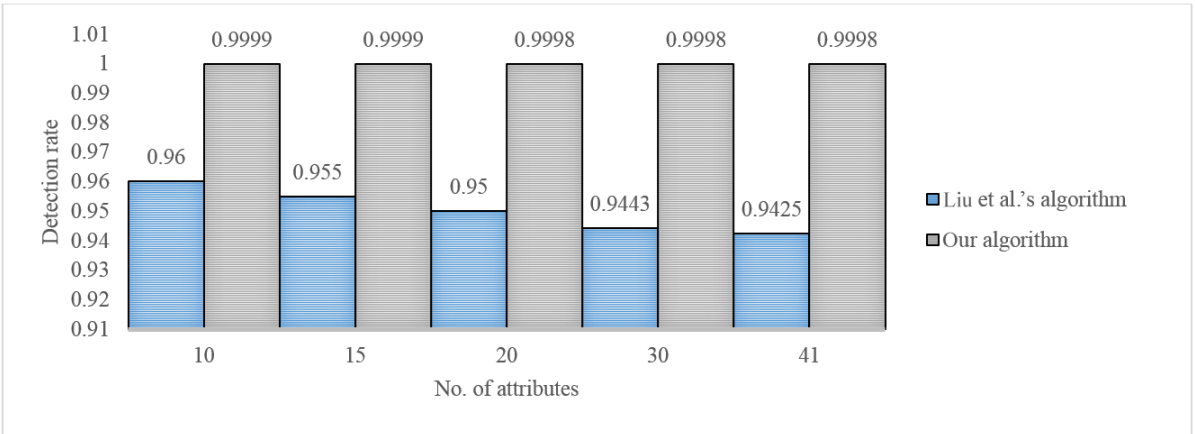

Figure 2. Comparative analysis of detection rates

6. **Conclusions.** In this article, we present a neighborhood rough set-based classification algorithm for the efficient detection of anomalies from network dataset. The dataset is a mixed kind that includes both numeric and categorical attributes. First of all, we define a unified metric that applies to both types of attributes. Then, we create a neighborhood relation using the unified metric, which is then utilized to determine lower approximation space, upper approximation space, and boundary regions. After that the upper approximation space produces the certain rules, the lower approximation space produces the possible rules, and the boundary region gives us the boundary rules. Finally, the algorithm's performance is proved through testing using the dataset KDD Cup'99 and comparative studies with [16]. The results suggest that our method outperforms the above-mentioned method in terms of detection rate accuracy.

REFERENCES

[1] R. Pamula, J. K. Deka and S. Nandi, An outlier detection method based on clustering, *Proc. of 2011 2nd International Conference on Emerging Applications of Information Technology*, India, pp.253-256, 2011.

[2] Y. Zhang, J. Liu and H. Li, An outlier detection algorithm based on clustering analysis, *Proc. of 2010 1st International Conference on Pervasive Computing, Signal Processing and Applications*, 2010.

[3] V. Jyothsna and K. M. Prasad, Anomaly-based intrusion detection system, *Computer and Network Security*, 2019.

[4] J. Jabez and B. Muthikumar, Intrusion Detection System (IDS): Anomaly detection using outlier detection approach, *Procedia Computer Science*, vol.48, pp.338-346, 2015.

[5] A. K. Pujari, *Data Mining Techniques*, University Press, 2001.

[6] S. M. Abdullah Al Mamuna and J. Valimaki, Anomaly detection and classification in cellular networks using automatic labeling technique for applying supervised learning, *Procedia Computer Science*, vol.140, pp.186-195, 2018.

[7] D. Dasgupta and N. S. Majumdar, Anomaly detection in multidimensional data using negative selection algorithm, *Proc. of the 2002 Congress on Evolutionary Computation (CEC'02)*, USA, pp.1039-1044, 2002.

[8] A. Taha and A. S. Hadi, Anomaly detection methods for categorical data: A review, *ACM Computing Surveys*, vol.52, no.2, pp.1-35, 2019.

[9] F. A. Mazarbhuiya, M. Y. AlZahrani and G. Lilia, Anomaly detection using agglomerative hierarchical clustering algorithm, *Information Science and Application*, vol.514, pp.475-484, 2018.

[10] F. A. Mazarbhuiya, M. Y. AlZahrani and A. K. Mahanta, Detecting anomaly using partitioning clustering with merging, *ICIC Express Letters*, vol.14, no.10, pp.951-960, 2020.

[11] *KDD Cup'99 Data*, http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html, Accessed on 15-Jan-2020.

[12] Z. Pawlak, Rough sets, *International Journal of Computer and Information Sciences*, vol.11, pp.341-356, 1982.

[13] R. R. Nowicki, *Rough Set Based Classification Systems*, Springer, 2019.

[14] S. U. Kumar and H. H. Inbarani, A novel neighborhood rough set based classification approach for medical diagnosis, *Procedia Computer Science*, vol.47, pp.351-359, 2015.

[15] M. El Meziati and H. Ziyati, Fast outlier detection method based on rough set, *The 9th International Symposium on Signal, Image, Video and Communications (ISIVC)*, Morocco, 2018.

[16] C. Liu, Y. Li and Y. Qin, Research on anomaly intrusion detection based on rough set attribute reduction, *Proc. of the 2nd International Conference on Computer Application and System Modeling*, Paris, France, pp.607-610, 2012.

[17] T. H. Clifford and W. Stephenson, *An Introduction to Numerical Classification*, Academic Press, New York, San Fransisco, London, 1975.

[18] S. M. Emran and N. Ye, Robustness of Canberra metric in computer intrusion detection, *Proc. of 2001 IEEE Workshop on Information Assurance and Security*, NY, USA, pp.80-84, 2001.

[19] L. Bergman and Y. Hoshen, Classification-based anomaly detection for general data, *Proc. of ICLR-2020*, pp.1-12, 2020.

[20] N. AlDouf, H. A. Karim and A. S. Ba Wazir, Model fusion of deep neural networks for anomaly detection, *Journal of Big Data*, vol.8, no.106, DOI: 10.1186/s40537-021-00496-w, 2021.

[21] J. Cao, G. Yang, X. Yang and J. Li, A visual surface defect detection method based on low rank and sparse representation, *International Journal of Innovative Computing, Information and Control*, vol.16, no.1, pp.45-61, 2020.

[22] N. Seliya, A. A. Zadeh and T. M. Khoshgaftaar, A literature review on one-class classification and its potential applications in big data, *Journal of Big Data*, vol.8, no.122, DOI: 10.1186/s40537-021-00514-x, 2021.

[23] S. Nedelkoski, J. Bogatinovski, A. Acker, J. Cardoso and O. Kao, Self-attentive classification-based anomaly detection in unstructured logs, *2020 IEEE International Conference on Data Mining (IC DM)*, Italy, 2020.

[24] H. Li, Y. Zhu and Y. He, Classification-based self-supervised learning for anomaly detection, *Proc. of the 13th International Conference on Digital Image Processing (ICDIP'2021)*, Singapore, 2021.

[25] A. O. Adetunmbi, S. O. Falaki, O. S. Adewale and B. K. Alese, Network intrusion detection based on rough set and K-nearest neighbour, *International Journal of Computing and ICT Research*, vol.2, no.1, pp.60-66, 2008.

[26] N. Sengupta, J. Sen, J. Sil and M. Saha, Designing of online intrusion detection system using rough set theory and Q-learning algorithm, *Neurocomputing*, vol.111, pp.161-168, 2013.

[27] W. Chimphlee, A. H. Abdulla, M. N. Md Sap and S. Chimphlee, A rough-fuzzy hybrid algorithm for computer intrusion detection, *The International Arab Journal of Information Technology*, vol.4, no.3, pp.247-254, 2007.

[28] E. M. Marouane and Z. Elhoussaine, A fuzzy neighborhood rough set method for anomaly detection in large scale data, *IAES International Journal of Artificial Intelligence*, vol.9, no.1, pp.1-10, 2020.

[29] I.-H. Bae, A rough set based anomaly detection scheme considering the age of user profiles, *Proc. of the 7th International Conference on Computational Science (ICCS'07)*, pp.558-561, 2007.

[30] K. Selvakumar, M. Karuppiah, L. SaiRamesh, S. K. H. Islam, M. M. Hassan, G. Fortino and K.-K. R. Choo, Intelligent temporal classification and fuzzy rough set-based feature selection algorithm for intrusion detection system in WSNs, *Information Sciences*, vol.497, pp.77-90, 2019.

[31] V. Rampure and A. Tiwari, A rough set based feature selection on KDD CUP 99 data set, *International Journal of Database Theory and Application*, vol.8, no.1, pp.149-156, 2015.

[32] *DARPA'98 Data*, http://www.ll.mit.edu/r-d/datasets/1999-darpa-intrusion-detection-evaluation-dataset, Accessed on 15-Jan-2020.