

Soft Set Based Intrusion Detection System Architecture using Genetic Algorithm

Sweta Sharma¹; AnandJawdekar²

Dept. Software System¹; Dept. of Computer Science & Engg²

SRCEM^{1,2}

Sharmasweta64@gmail.com¹; anand.cs2007@gmail.com²

ABSTRACT

IDS is the system which identifies malicious activity on the network. As the Internet volume is increasing rapidly, security against the real time attacks and their fast detection issues gain attention of many researchers. Approaches of data mining can be successfully applied to IDS to tackle dynamic data problems and to increase performance of IDS. We can decrease the complexity of time by selecting only useful features to build model for classification. There are various features selection methods are developed either to select the features or extract features. In this paper, an innovative evolutionary method for the feature selection is proposed. Genetic algorithm (GA) is used as a search method while selecting features from KDD data set along with the selection of those only who appears everywhere in the experiment. The experiments are performed with reduced time and minimum number of features.

Keywords: Data mining, Intrusion Detection, Feature Selection, Classification, Genetic Algorithm, Soft Set.

1. INTRODUCTION

With the tremendous development in the computer networks use day by day, network as well as information security becomes the prime important factor. The basic aim of security is to develop protective software system which can provide three basic security goals that are confidentiality, integrity and authentication. Growth in the computer and network of computer use today's civilization seeks extremely secured and trusted communication. There is an methods range being utilized in IDS, but any of the systems is not completely perfect. We refer intrusion as any set of

events that try to negotiate the integrity, confidentiality, or availability of a computer resource. The finding procedure out the asymmetrical activities on network and system is known as IDS. Intrusion is any activity which tries to violate these security goals [1]. The IDS plays a key role in identifying such malicious activities. The term IDS was first presented through Anderson in 1980.

The attribute selection is done by a procedure which has various advantages over any ordinary method. The search space is made of chromosomes, which are a sequence of real numbers representing the cluster centers initially. In order to get an optimal solution, which is a good partitioning with clusters as pure as possible (homogenous), we propose a new and innovative approach. At first, clusters, which is centers are represented through a chromosome, are filtered applying a given fitness value. Experimental outcomes comparing GA-base algorithm with the new proposed algorithm [6] are provided for various KDD'99 dataset subset.

1.1 Intrusion Detection System

IDS is the main study area in Computer-based security. It is a well-known skill for enlightening and is used to protect data consistency and system accessibility throughout an intrusion. When a person tries to access structure of knowledge in the particular system or does any unlawful action, the action is known as an intrusion that further has two types, exterior, and interior. The exterior are those people which have not access authority the system information and still they try to obtain illegitimately with the help of different saturation

techniques. While interior is those who have a legal permission to access system, but try to do illegal activities. Software bugs exploitation and miss configurations of the system cause intrusion.

1.2 Genetic Algorithm

GA is the search heuristic which provide a valuable solution to the problems of optimization and search. GAs include a technique to get optimal combinatorial state using interest parameters set. Genetic programming also helps in simulating the population evolution process. Genetic algorithm evolves the population of fixed length by applying mutation operators and crossover along with a fitness function. The output concludes how likely individuals are to reproduce. A set of rules is developed each of which is calculated approximately to get fitness. Rules having higher fitness value create a novel generation. Numerous generations go through the same procedure to produce a solution that is acceptable. GA are an adaptive heuristic search method based on natural selection theory [7]. They are inspired by Darwin's evolution theory– "survival of the fittest", which is one of the randomized search techniques.

1.3 Soft set Approach

The soft set theory is now a part of soft computing. It is the new approach in data mining which deals with uncertain data. Experiments reveal that in some aspects the soft sets and rough sets are similar. Soft sets are mainly used to analyze huge amount of data and help in taking well informed decisions. Therefore soft sets can be used in real world decision support systems. Such systems can help management of an organization to take decisions that can result in expected results or profits. Experiments as shown in literature indicate that soft set theory can be used in tandem with other some computing techniques in order to produce highly effective results. Soft sets can also be used to achieve reducts that will help in better performance. Soft sets are being used in various applications such as classification of medical data, classification of musical instruments, evaluation of student marks, better grouping of data etc. In many applications soft sets can be used to generate decision rules that can help to take well thought out decisions as the soft sets are capable of providing required business intelligence.

Definition 1.1 (Soft Set) A pair (F, E) is called a soft set (over U) if and only if F is a mapping of E into the set of all subsets of the set U . In other words, the soft set is a parameterized family of subsets of the set U . Every set (e) , $e \in E$, from this family may considered as collection of soft sete - approximate elements.

Example 1.1.1: A soft set (F, E) defines attractiveness of bikes which Mr. X is going to purchase [Pal & Mondal, 2011].

U is the set of bikes under consideration. E is the set of parameters. Each parameter is a word or a sentence.

$E = (e_1 = \text{stylish}; e_2 = \text{heavy duty}; e_3 = \text{light}; e_4 = \text{steel body}; e_5 = \text{cheap}; e_6 = \text{good mileage}; e_7 = \text{easily Started}; e_8 = \text{long driven}; e_9 = \text{costly}; e_{10} = \text{fibre body})$

Here, define a soft set means to point out stylish bikes, heavy duty bikes, and so on.

Example 1.1.2 : Let $U = \{u_1, u_2, u_3, u_4, u_5\}$ be a universal set and $E = \{x_1, x_2, x_3, x_4\}$ be a set of parameters. If $A = \{x_2, x_3, x_4\}$ and then the soft set FA is written by $FA = \{(x_2, \{u_2, u_4\}), (x_4, U)\}$

Definition 1.2 (Operation with Soft Sets)

Suppose a binary operation denoted by $*$, is defined for all subsets of the set U . Let (F, A) and (G, B) be two soft sets over U . Then the operation $*$ for the soft sets is defined in the following way: $(F, A) * (G, B) = (H, A \times B)$ Where $H(\alpha, \beta) = (\alpha) * (\beta)$, $\alpha \in A$, $\beta \in B$ and $A \times B$ is the Cartesian product of the sets A and B .

Definition 1.3 (Complement of a Soft Set) The complement of a soft set (F, A) is denoted by $(F, A)^c$ and is defined by $(F, A)^c = (F^c, \bar{A})$ where $F^c : \bar{A} \rightarrow P(U)$ is a mapping which is defined by $F^c(\alpha) = U - F(\alpha)$, for all $\alpha \in \bar{A}$.

Definition 1.4 (NULL Soft Set) A soft set (F, A) over U is said to be a NULL soft set denoted by Φ , if for all $\epsilon \in A$, $F(\epsilon) = \phi$ (null-set).

Definition 1.5 (AND Operation on Two Soft Sets) If (F, A) and (G, B) be two soft sets then (F, A) AND (G, B) denoted by $(F, A) \wedge (G, B)$ and is defined by $(F, A)(G, B) = (H, A \times B)$ where $H(\alpha, \beta) = F(\alpha) \cap G(\beta)$ for all $(\alpha, \beta) \in A \times B$.

Definition 1.6 (OR Operation on Two Soft Sets) If (F, A) and (G, B) be two soft sets then (F, A) OR (G, B) denoted by $(F, A) \vee (G, B)$ is defined by $(F, A) \vee (G, B) = (O, A \times B)$ where $O(\alpha, \beta) = F(\alpha) \cup G(\beta)$ for all $(\alpha, \beta) \in A \times B$.

In this paper represent organized as follows: Section II gives brief introduction about the different methods used in system. Section III elaborates proposed method used. Section IV describes the experimental outcomes and dataset used and conclusion along with further scope is provided in Section V.

2. LITERATURE SURVEY

Bridges [8] represent a technique applying GA to detect network intrusion. This approach obtains classification rules for quantitative and distinct network data features.

In Lu [10] method classification rules are generated by Genetic Programming. This method detects or classifies intrusions in a system using a fitness function. Because of the significant data time required to system train creates genetic programming implementation difficult.

Crosbie [11] represents that Genetic programming and various agent methods can be used for network intrusions detecting. A collection of agents finds out the network behaviors and monitors one parameter of the network audit data and genetic programming. The advantage of this method is, many small agents that are independent can be used, but the communication between the agents is a drawback.

This system identifies the attacks using a set of rules generated by genetic algorithm, then exploits rules for DoS, U2R, R2L, and probe attacks.

J. Gómez and E. León [14] proposed fuzzy and genetic algorithm to categorize activities of intrusion on the network. They used KDDCup99 dataset as input data that consists of 42 features. The fuzzy rule is modified using evolutionary technique and genetic algorithm. The algorithm can categorize the data into DoS, R2L, Probe, U2R, and Normal. This algorithm has detection rate of 98.28 %.

W. Li [16] described a method using GA to identify irregular network intrusion. The process contains both quantitative and definite network features information for deriving classification

rules. Though, quantitative feature addition can amplify detection rate but no tentative results are present.

Soft set theory was first introduced by Molodtsov in 1999 as a new paradigm for mining uncertain data. Soft sets overcome various method inadequacy for example interval mathematics [17], theories of fuzzy set [18] and also probability. Pei and Miao [19] explored information systems and soft sets in terms of relationship between them. The results of their experiments reveal that information systems and partition-type soft sets share a common formal structure. For example fuzzy soft sets and fuzzy data systems are equal.

Chetia and Das [20] extended Biswas's technique for answer scripts evaluation of students. They expected five approval levels in order to evaluate the students performance. They conclude unsatisfactory, satisfactory, good, very good and excellent. They have developed an algorithm that takes student's statistics as input and build a soft set matrix before evaluating the performance of students.

Parameterization reduction is also possible in soft sets and related applications as presented by Chen et al. [21]. They further said that method followed through Maji was incorrect and claimed that reduct is not same for rough set theory and soft set theory. Their idea for reduction of attributes in soft sets was based on the optimal choice concept that addresses the problems of sub-optimal solutions.

3. Proposed Work

3.1 Proposed Idea

The proposed idea of Anomaly-Based Intrusion Detection System is based on new soft set based genetic approach. Soft set approach is a new approach that works on all kind of data. The dataset used in IDS is Kddcup'99.

3.1.1 Dataset

The Kddcup'99 dataset is used for testing and training technique developed. The 10% of the Kddcup'99 dataset contain of 5 million instances many of them are redundant. The 10% of the Kddcup'99 dataset consist of 494021 instances. In which 97278 are Normal and remaining 396743 are belongs to any one kind of attack. The original KDD CUP'99 Data Set involve 22 types of attack levels, it was very complicated to analyze the classification system performance. To make ease of

analyze, the attack levels are changed to their respective categories, which are DoS, Probe, R2L, U2R and last is normal.

3.1.2 Soft set Approach

Soft set is a parametrized common mathematical tool which deals with an approximate descriptions set of the objects. Each approximate description has various parts, a predicate and an approximate value set. In mathematics, a mathematical object model is defines and generated exact solution notion of this model. Most likely mathematical problematic and certain answer is just not without difficulty obtained. So, approximate answer inspiration is awarded and the solution is calculated. In soft set idea, now we have the reverse manner to this drawback. The first object description has an approximate nature, and we don't required to the precise answer suggestion gift. The any restrictions absence on the approximate description of the soft set theory creates this theory most convenient and simply applicable in the practice.

3.2 Proposed Algorithm

1. Please Attribute Selection- out of 41 attributes, necessary attributes are selected based on algorithms:
 - a. Attribute evaluator – CfsSubsetEval and search method used is ScatterSearchV1. Using these algorithms, necessary attributes are selected.
2. Soft set approach for feature selection and removing the outliers.
3. Applying genetic:
 - a. Converting all the selected attributes in binary form.
 - b. Applying crossover and mutation.
 - i. If more than half of the bits are matching, then perform crossover else mutate any random bit.

- c. If(fitness < decimal value of attributes)
Keep the values.
Else

Keep the original values only.

4. Evaluating the final factors by applying the database in weka. The confusion matrix is created by applying classification REPTree algorithm.
5. Show the evaluation measures for the proposed algorithm.

3.2.1 Dataset

The Kddcup'99 dataset is used for the developed system training and testing. The 10% of the Kddcup'99 dataset contain of 5 million instances many of them are redundant. The 10% of the Kddcup'99 dataset consist of 494021 instances. In which 97278 are Normal and remaining 396743 are belongs to any one kind of attack. The original KDD CUP'99 Data Set involve 22 types of attack levels, it was very complicated to analyze the classification system performance. To make ease of analyze, the attack levels are changed to their respective categories, which are DoS, Probe, R2L, U2R and last is normal.

The above proposed algorithm is shown with the help of flowchart below:-

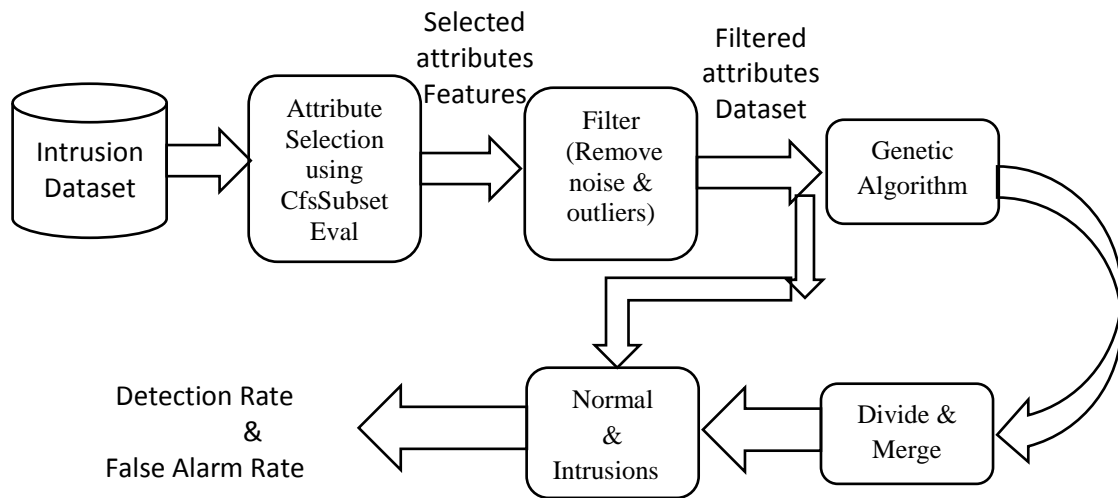


Fig 1: IDS System Architecture

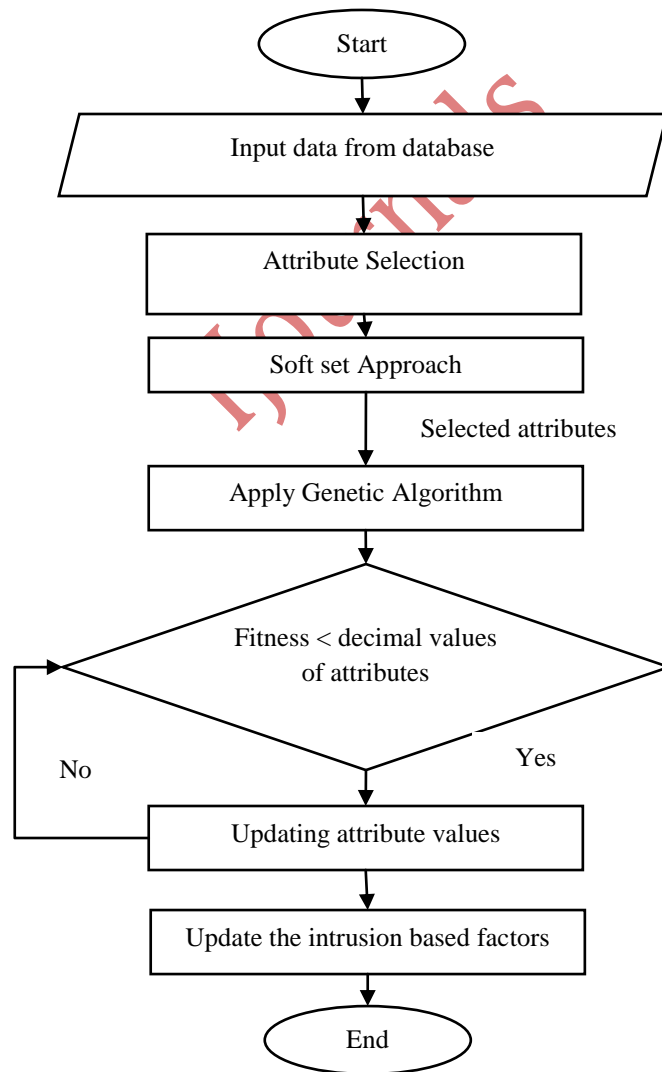


Fig 2: Process Flowchart

4. RESULT ANALYSIS

The classifier performance evaluator phases have facilities the calculated of various classification performance measure in order to judge the accuracy of the proposed system. These measures are as follows:

True Positive Rate (TPR):

$$TPR = TP/(TP+FN)$$

False Positive Rate (FPR):

$$FPR = FP/(TN+FP)$$

Where TP (True Positive), FN (False Negative), FP (False Positive) and TN (True Negative) can be defined as follows:

True Negative (TN): These are negative tuples that were correctly labeled through classifier.

True Positive (TP): These are positive tuples that were correctly labeled through classifier.

False Positive (FP): These are negative tuples that were incorrectly labeled as positive.

False Negative (FN): These are positive tuples that were mislabeled as negative.

These terms can we understand by the concept of confusion matrix shown in table 4.3, The row represents actual class instances while column in matrix represents predication class instances.

Table 1. Confusion Matrix for TN, TP, FP and FN

Valid Record	Correctly Classified	Incorrectly Classified
	True Negative (TN)	False Positive (FP)
Attack Record	True Positive (TP)	False Negative (FN)

Confusion Matrix is one of the other different parameters in the literature to analyze the model performance. The column in the matrix represents the prediction class instances while the row represents the actual class instances.

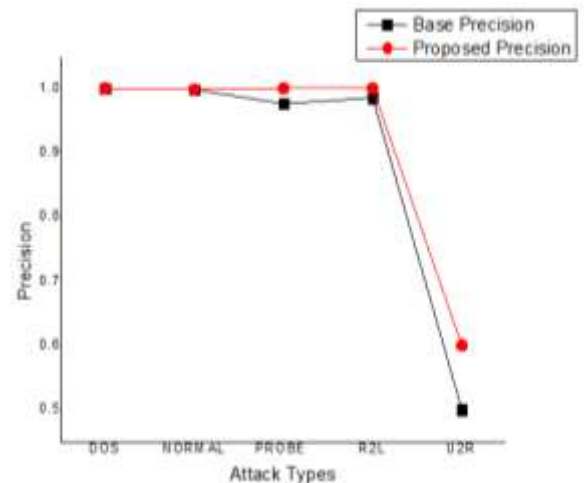
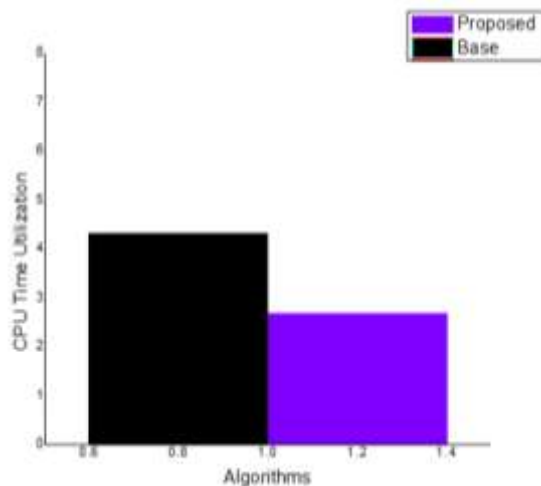
Table 2. Comparison of Base & Proposed Results

	BASE	PROPOSED
Correctly Classified Instances	99.897 %	99.9037 %
Incorrectly Classified Instances	0.103 %	0.0963 %
Kappa statistic	0.9969	0.9971
Mean absolute error	0.0006	0.0007
Relative absolute error	0.443 %	0.5175 %
Total Number of Instances	33978	33216

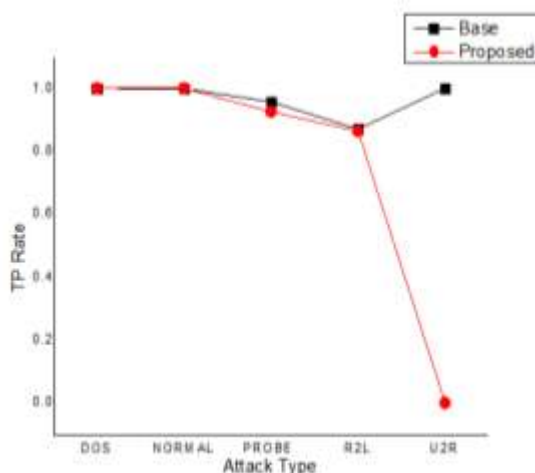
The proposed work shows that the total correctly classified instances are more than that of base. Also the total number of incorrectly classified instances is the wrongly classified instances and should have their value as less as possible. Other factors are also better when compared with the base work.

The results are compared further using the graphs that show the differences in the TPR rate along with precision and time factors.

1. **CPU Time Utilization**—the time utilized by the base algorithm is higher as compared to proposed algorithm. It is also an important factor that shows the proper working of an IDS system.



2. **TP Rate** – TP rate can also be called as true positive rate which is calculated from the confusion matrix created using the attributes in the database. The true positive rate shows the correctly classified attack types and thus the value should be higher. In our proposed work, the TPR is high as compared to base algorithm.



3. **Precision** – Precision is various correctly classified instances over total instances amount. The precision rate should be high as the correctly classified instances should be as high as possible. In the proposed work the precision is higher as compared to the base algorithm.

5. CONCLUSION & FUTURE WORK

In this paper, a new fuzzy genetic algorithm is proposed for dealing with the intrusion detection problem considering KDD99 dataset. Outcomes are compared with existing system which uses GA algorithm. The results present that the accuracy of detection rate of the proposed system for DoS, probe, Remote to User Attacks (R2L) and User to Root attack (U2r) are more compared to the existing systems. The time required for the training and testing of the dataset using the proposed system is less compared to the existing systems and memory allocation also requires less space for proposed system than existing systems.

6. REFERENCES

- [1]. Muamer N. Mohammada, Norrozila Sulaimana, Osama AbdulkarimMuhsinb, "A Novel Intrusion Detection System by using Intelligent Data Mining in Weka Environment", Procedia Computer Science 3,2011, pp. 1237 – 1242.
- [2]. G.V. Nadiammai, M. Hemalatha, "Effective approach toward Intrusion Detection System using data mining techniques", Egyptian InformaticsJournal, 2013.
- [3]. Paul Dokas, Levent Ertoz, Vipin Kumar, Aleksandar Lazarevic, JaideepSrivastava, Pang-Nig Tan, "Data Mining for Network IntrusionDetection".
- [4]. Theodoros Lappas and Konstantinos Pelechrinis, "Data MiningTechniques for (Network) Intrusion Detection Systems".
- [5]. J Bartlett, "Machine Learning for Network Intrusion Detection", 2009
- [6]. Amira Sayed A. Aziz, Ahmad Taher Azar, Mostafa A. Salama, "GeneticAlgorithm with Different Feature Selection Techniques for AnomalyDetectors Generation", Proceedings of the 2013 Federated

- Conference on Computer Science and Information Systems, pp. 769–774.
- [7]. Anup Goyal, Chetan Kumar, “GA-NIDS: A Genetic Algorithm based Network Intrusion Detection System”.
- [8]. Bridges and Vaughn, ”Fuzzy Data Mining and Genetic Algorithms Applied to Intrusion Detection”, Proceedings of 12th Annual Canadian Information Technology Security Symposium, pp. 109-122, 2000.
- [9]. Bridges, Susan and Rayford B. Vaughn. 2000. “Intrusion Detection via Fuzzy Data Mining”, In Proceedings of 12th Annual Canadian Information Technology Security Symposium, pp. 109-122. Ottawa, Canada.
- [10]. W. Lu and Traore, “Detecting new forms of network intrusion using genetic programming”, Computational Intelligence Vol.20, Issue 3, august 2004. I.S. Jacobs and C.P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G.T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271-350.
- [11]. Crosbie, Mark, and Gene Spafford. 1995. “Applying Genetic Programming to Intrusion Detection”. Proceeding of 1995 AAAI Fall Symposium on Genetic Programming, Cambridge, Massachusetts.
- [12]. P. Jongsuebsuk, N. Wattanapongsakorn, C. Charnsripinyo “Real-Time Intrusion Detection with Fuzzy Genetic Algorithm.” ©2013 IEEE.
- [13]. T.P. Fries, “A fuzzy-Genetic approach to network intrusion detection,” GECCO’08: The 10th Annual Conference on Genetic and Evolutionary Computation, 2008, pp. 2141-2146.N.
- [14]. J. Gomez and E. León, “A fuzzy set/rule distance for evolving fuzzy anomaly detectors,” IEEE International Conference on Fuzzy Systems ART. No. 1682017, pp. 2286-2292.
- [15]. N. Ngamwitthayanon and N. Wattanapongsakorn, “Fuzzy- ART in network anomaly detection with feature-reduction dataset,” The 7th International Conference on Networked Computing, INC2011, Art. No. 6058956,
- [16]. W. Li, “A Genetic Algorithm Approach to Network Intrusion Detection” SANS. Institute, USA, 2004.
- [17]. Yang, X B, et al. (2009). “Combination of interval-valued fuzzy set and soft set”. Computers and Mathematics with Applications, 58, 521-527.
- [18]. Zadeh, L A (1965). “Fuzzy set”. Information and Control, 8, 338-353.
- [19]. Daowu Pei and Duoqian Miao, “From Soft Sets to Information Systems”
- [20]. Samsiah Abdul Razak and Daud Mohamad, “A Soft Set based Group Decision Making Method with Criteria Weight”, World Academy of Science, Engineering and Technology 58 2011.
- [21]. B. Chetia and P. K. Das, “Application of Vague Soft Sets in students’ evaluation”. Advances in Applied Science Research, 2011, 2 (6):418-423.