Soft Set Based Intrusion Detection System Architecture using Genetic Algorithm

Manish Arya[1], DR. Sanjiv Sharma[2]
[1, 2] Department of Computer Science
[1, 2] Madhav Institute Of Technology Science Gwalior

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

I. 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 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 find space is create of chromosomes, that are a order of real numbers representing the cluster middles initially. In sequence to get an optimal solution, that is a good partitioning with clusters as pure as possible (homogenous), we presented 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.

In this paper represent unified like this: Section II gives brief introduction about the different methods used in system. Section III elaborates proposed approach used. Section IV defines the experimental outcomes and dataset used and conclusion along with further scope is provided in Section V.

A. Intrusion Detection System

IDS has become a primary study area in Computer-based security. It's a well-known skill for enlightening and is exploited to defend 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 forms such as interior and exterior. The exterior are those people which have not access authority the system information and still they try to obtain illegitimately with the aid 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. Sniffing unsecured traffic, password cracking, or exploiting the particular protocols design blame are also some of the ways that cause intrusion.

B. Genetic Algorithm(GA)

GA is a find heuristic which gives a suitable solution to find and optimization difficulties. 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 collection 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 find method depend on natural election idea [7]. They are motivated thru Darwin's evolution theory– "survival of the fittest", which is a randomized search techniques.

C. Soft set Approach

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

Definition 2.1 (Soft Set) A couples (F, E) is known a soft set (over U) if and only if F is a mapping of E into the group of each subsets of the set U. In other terms, the soft set is a parameterized family of subsets of the set U. Every set (e), $e \in E$, from this family may be contemplate as the group of e -approximate components of the soft set.

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

U is the group of bikes below deliberation. E is the set of parameters. Each parameter is a word or a sentence.

E = (e1= stylish; e2 =heavy duty; e3= light; e4 =steel body; e5 = cheap; e6 = good mileage; e7 =easily Started; e8 = long driven; e9=costly; e10 = fibre body)

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

Example 1.1.2 : Let $U = \{u1, u2, u3, u4, u5\}$ be a universal set and $E = \{x1, x2, x3, x4\}$ be a set of parameters. If $A = \{x2, x3, x4\}$ and then the soft set FA is written thru F_A ={(x 2,{u 2,u 4}),(x 4,U)} .

Definition 2.1 (Operation with Soft Sets)

Assume a binary operation indicated thru *, is determine for every 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: $(F, A) * (G,) = (H, A \times B)$ 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 2.2 (Complement of a Soft Set) The elements of a soft set (F, A) is indicated thru (F, A) c and is describe via (F, A) ^{c} = (F^c , ⌉A) where F^c : ⌉A → P(U) is a mapping that is well-defined thru F^c (α) =U - F(⌉α), for all α∈⌉A.

Definition 2.3 (NULL Soft Set) A soft set (F, A) over U is said to be a NULL soft set indicated by Φ , if for every ε ∈ A, F (ε) = ϕ (null-set).

Definition 2.4 (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 well-defined thru $(F, A)(G, B) = (H,A \times B)$ where $H(\alpha, \beta) = F(\alpha) \cap G(\beta)$ for all $(\alpha, \beta) \in A \times B$.

Definition 2.5 (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 well-defined via $(F, A) \vee (G,B) = (O, A \times B)$ where $O(\alpha, \beta) = F(\alpha) \cup G(\beta)$ for all $(\alpha, \beta) \in A \times B$.

II. LITERATURE SURVEY

Bridges [8] represent a technique applying GA to discover 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 exploit for network intrusions detecting. A collection of agents discovers the grid performances and displays one parameter of the grid review 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 collection 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 procedure includes both quantitative and definite grid characteristics info for deriving taxonomy rules. Though, quantitative feature addition can amplify detection rate but no tentative results are present.

Soft set theory was first presented via Molodtsov in 1999 as a novel design for mining wavering data. Soft sets To result the inaptitude of other methods for instance interval mathematics [17], fuzzy set theories [18] and probability. Pei and Miao [19] find out info systems and soft sets in terms of relationship amid them. The outcomes of their tests reveal that info systems and divider-kinds soft sets share a general formal structure. As fuzzy info systems and fuzzy soft sets are similar.

Chetia and Das [20] extended Biswas's manner for assessment of reply scripts of students. They supposed five gratification stages in sequence to assess the recital of students. They contain satisfactory, good, very good ,excellent and insufficient. They have developed an algorithm which takings scholar's data like input and construct a soft set matrix earlier assessing the recital of scholar's.

Parameterization reduction is also possible in soft sets and related applications as presented by Chen et al. [21]. They additional said which the method followed thru Maji was improper and also demanded which the reduct is not similar for rough set theory and soft set theory. Their idea for reduction of attributes in soft sets was depend on the optimal election idea which addresses the issues of sub-optimal results.

III. PROPOSED WORK

Proposed Idea:

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

Dataset:

The Kddcup'99 dataset is exploited for the testing and training of the developed system. The 10% of the Kddcup'99 dataset include of 5 million e.g several of them are unnecessary. The 10% of the Kddcup'99 dataset contain of 494021 examples. In that 97278 are Normal and outstanding396743 are be appropriate to any one types of attack. The original KDD CUP'99 Data Set involve 22 variety 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 related categories, which are DoS, Probe, R2L, U2R and last is normal.

Soft set Approach: Soft set is a parametrized common mathematical device that deals with an circa descriptions set of the objects. Every circa detail has two dissimilar kinds, an circa value set an a predicate. A mathematical replica of an object is makes and describe the precise solution notion of this replica in the mathematics. Commonly mathematical replica is too complex and the unspoiled solution is not simply found. So, the approximate solution notion is presented and the solution is calculated. We have the reverse replica to this issues in the theory of soft set. The first object description has an approximate nature, and we do not necessity to the exact solution notion present. The any restrictions absence on the circa explanation of the soft set theory creates this theory most convenient and simply applicable in the practice. Any parameterization we favor can be exploit with the words and sentences aid, functions and mappings and real numbers so on.

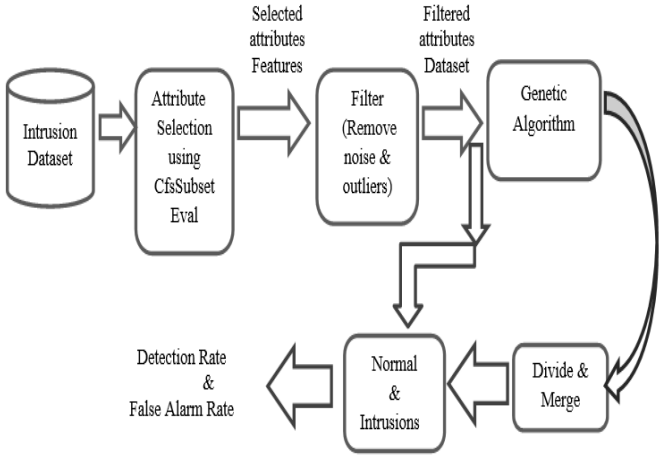


Figure 1: IDS System Architecture

The proposed algorithm works as follows:

1. Attribute Selection- out of 41 attributes, necessary attributes are elected depend on algorithms:
   - **a.** Attribute evaluator – CfsSubsetEval and search method used is ScatterSearchV1. Using these algorithms, necessary attributes are selected.
2. Soft set method for feature election 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.

Dataset:

The Kddcup'99 dataset is exploited for the testing and training of the developed system. The 10% of the Kddcup'99 dataset include of 5 million e.g several of them are redundant. The 10% of the Kddcup'99 dataset include of 494021 examples. In which 97278 are Normal and outstanding 396743 are belongs to any one manner of attack. The original KDD CUP'99 Data Set involve 22 variety 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 related categories, which are DoS, Probe, R2L, U2R and last is normal.

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

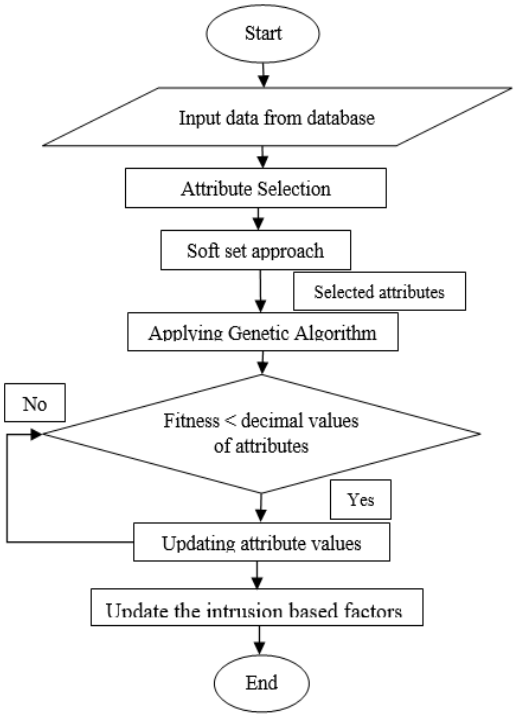


Figure 2: Process Flowchart

IV. RESULT ANALYSIS

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

False Positive Rate (FPR):

- $FPR = \frac{FP}{TN+FP}$

True Positive Rate (TPR):

- $TPR = \frac{TP}{TP+FN}$

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

- True Positive (TP): It is positive tuples which were properly labeled through the classed.
- False Positive (FP): It is the negative tuples which were wrongly labeled like positive.
- False Negative (FN): It is positive tuples which were mislabeled like negative.
- True Negative (TN): It is negative tuples which were properly labeled through the classed.

These terms can we understand by the thought of confusion matrix shown in table 4.3, where a confusion matrix is a performance tabular visualization of an algorithm? The column in the matrix represents the examples of a predication class while the row represents actual 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 dissimilar parameters in the literature to analyze the relica recital. A confusion matrix is an algorithm recital tabular visualization. The column embodies the prediction class examples while the row embodies the actual class as in the matrix.

Table 2: Comparison of Base & Proposed Results

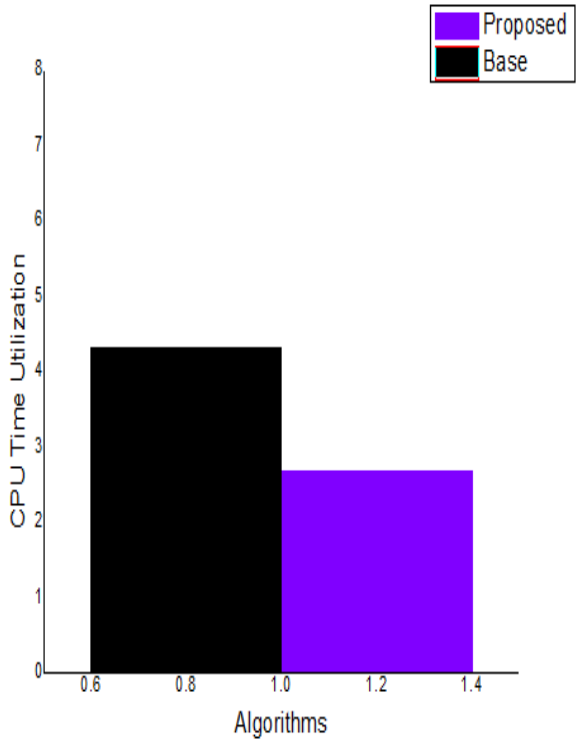
	BASE	**PROPOSED**
Correctly Classified Instances	99.897 %	99.9037 %
Incorrectly Classified Instances	0.103 %	0.0963 %
Kappa statistic	0.9969	0.9971

Mean absolute error	0.0006	0.0007
Relative absolute error	0.443 %	0.5175 %
Total Number of Instances	33978	33216

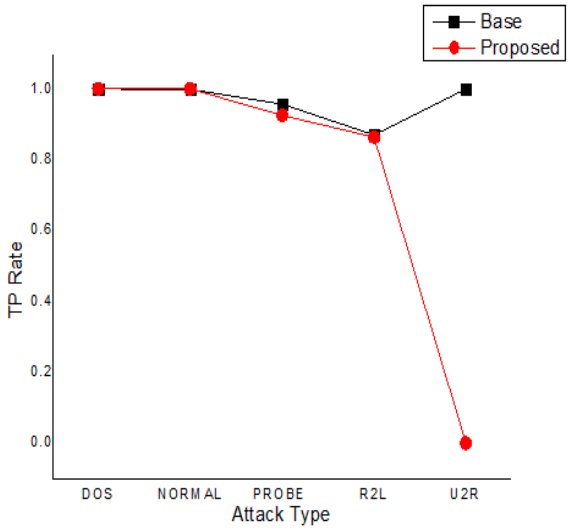
The proposed work shows that the total correctly categorized examples are more than that of base. Also the complete number of incorrectly categorized examples is the mistakenly categorized instances and should have their value as less as possible. Other factors are also better when equated with the base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

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

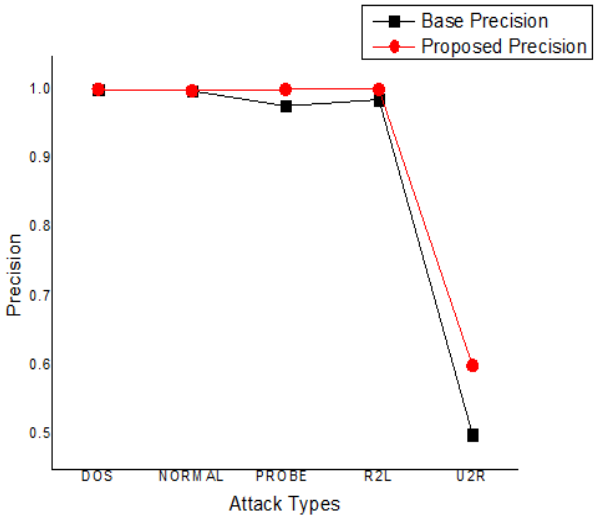
1. **CPU Time Utilization** – the time utilized by the base algorithm is higher as equated to proposed algorithm. It's also an important factor that shows the proper working of an IDS system.



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 categorized attack types and thus the value should be higher. In our proposed work, the TPR is high as equated to base algorithm.



3. **Precision** – Precision is the complete number of properly classified examples over the complete number of examples. The precision rate should be high as the correctly categorized examples should be as high as possible. In the proposed work the precision is higher as equated to the base algorithm.



V. CONCLUSION & FUTURE WORK

In this paper, a new fuzzy genetic algorithm is proposed for demeanor with the intrusion detection problem considering KDD99 dataset. Results are equated with the current system that exploite GA algorithm for intrusion detection. The results present that the accuracy of detection rate of the proposed system for DoS, probe, Remote to User Attacks (R2l) and Customer to Root attack (U2r) are more equated to the existing systems. The time required for the training and testing of the dataset utilizing the proposed system is a

smaller equated to the existing systems and memory allocation also requires less space for proposed system than existing systems.

REFERENCES

[1] Muamer N. Mohammada, Norrozila Sulaimana, Osama Abdulkarim Muhsinb, "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 Informatics Journal, 2013.

[3] Paul Dokas, Levent Ertoz, Vipin Kumar, Aleksandar Lazarevic, Jaideep Srivastava, Pang-Nig Tan, "Data Mining for Network Intrusion Detection".

[4] Theodoros Lappas and Konstantinos Pelechrinis, "Data Mining Techniques 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, "Genetic Algorithm with Different Feature Selection Techniques for Anomaly Detectors 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.

[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.