Turkish Journal of Computer and Mathematics Education Vol.12 No.13 (2021), 3152-3159

Research Article

An Effective Intrusion Detection System using Enhanced Multi relational Fuzzy Tree

Dr. P. Mahhizharuvi[1], Dr. A. V. Seethalakshmi[2]

[1]Department of Computer Science, Sri Meenakshi Gov. Arts College for Women(A), Madurai, Tamilnadu, India, mahizharuvi2008@gmail.com
[2]Head and Asst Professor, Department of Computer Science, Mangayarkarasi Arts and Science College for Women, Madurai, Tamilnadu, India, dravs2021@gmail.com

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 4 June 2021

ABSTRACT
Today, the enormous development of computer networks and communication technology, we rely heavily on network connections. The substantial consumption of social networks and internet leads to the drastic increment in the data which are naturally complex and sparse too. The data are stored in multiple database relations associated with primary and foreign keys. The internet attack is a main type of issue in computer networks. Numerous Network Intrusion Detection Systems (NIDSs) have been designed based on traditional data mining methods to identify and ease the network attacks. But the methods were suitable for single relational data. This paper proposes a novel method for classifying KDD CUP 99 intrusion detection data using Enhanced Multi relational Fuzzy decision tree (EMRFT). The generated tree is optimized based on genetic approach. The outcomes of empirical analysis show that the EMRFT achieves high prediction performance and less induction time in classifying network intrusion.

Keywords: Data Scrubbing, Fuzzy Optimization, KDD CUP 99, Multi relational Fuzzy Tree.

.

1. INTRODUCTION

In most of the real world applications such as Telecommunication, Intrusion Detection system, the data accumulated are immense and stored in relational databases. Such databases include several relations which are connected by primary and foreign keys. The traditional data mining algorithms can process the data stored in single flat relations and inappropriate for managing these databases. Thus, Multi Relational Data Mining has become a vital area of data mining. One important task of MRDM is Multi relational classification which constructs the model based on target and non-target relations in the relational databases using keys.

Nowadays, types of computer network intrusions have exceedingly increased and varieties of pioneering hacking tools and intrusive methods have developed. An intrusion detection system (IDS) is a technique used for handling suspicious activities in a network [5]. The data mining methods have been widely used by the researches for the detection of network intrusion. Different machine learning methods and soft computing techniques have showed their capabilities in IDS [20] [22]. An artificial intelligent method called Fuzzy logic has been effectively applied for detection of several IDSs [8] ,[13],[7]

The problems with vague and incomplete data can be handled by Fuzzy logic [15]. The fuzzy logic concept is applied for the separation of normal behavior from abnormal one. The ability of fuzzy logic to represent imprecise values helps in making decisions in indefinite areas such as intrusion detection.

In this work, an innovative intrusion detection system is devised with the help of Enhanced Multi Relational Fuzzy decision Tree (**EMRFT**). It uses genetic algorithm for optimizing the fuzzy membership function. The EMRFT classifier consists of two steps. First, it generates classifier using KDD99 training dataset. Second, it will classify the test dataset using the generated classification model.

1.1. Research Contribution

A novel multi relational system is proposed in this work to classify the intrusion detection using fuzzy tree. The research contributions are as follows

- To classify the intrusion detection across multi relations, we enhance the fuzzy tree using primary and foreign keys.
- The K-Nearest Neighbor method is applied to boost the performance of the classifier by imputing missing values in the database.
- It incorporates a Fast Correlation based feature selection method to minimize the dimension space of KDD'99 dataset.
- It uses Genetic algorithm to optimize the membership function of fuzzy tree.

1.2. Structure of the Paper

The structure of the paper is arranged as follows. The related work is given in Section 2. The working principle of the proposed EMRFT is explained in Section 3. Section 4 presents the experimental results and observations. Lastly, Section 5 concludes research work.

2. RELATED WORK

Fuzzy logic concept was used for anomaly based intrusion detection in [18]. In this work, fuzzy rule learning strategy was applied for automatic detection of efficient fuzzy rules. The hybrid approach was proposed in [16] by merging K-Medoids clustering with Support Vector machine. The proposed method achieved good results based on accuracy, detection rate metrics when compared to k-Medoids with Naïve Bayes classification.

In this study[11], support vector machine was used for intrusion detection using MATLAB software. The experimental results showed that SVM takes more time to train the model and it limits the usability of SVM. The naïve Bayes algorithm [4] was used for detecting all intrusion types in KDD dataset. The empirical test showed that by using single machine learning algorithm, detection rate produced by the system was not satisfactory.

The fuzzy genetic algorithm(FGA) and Multi Layer Perceptron (MLP) algorithm is used for KDD'99 and Online network dataset in [17]. The results indicate that the MLP algorithm was more secure and easy to implement than FGA. According to paper [1], Adaptive Neuro Fuzzy classifier was used to detect intrusion using KDD cup 99 datasets. The performance results showed that neuro-fuzzy classifiers achieves reducing false alarm rate and increasing detection rate for KDD99 dataset. The data mining algorithms namely SVM, C5.0 and Ripper rule based classifiers were used for intrusion detection [19] and results were compared. The outcomes showed that C5.0 decision tree produced efficient results than others.

3. PROPOSED SYSTEM

The chief intention of this work is to classify the network intrusion in multiple database relations using enhanced fuzzy decision Tree. The fuzzy membership functions of the trees are optimized using genetic algorithm [12]. The framework of EMRFT is shown in Figure 1 which is an improvement of **[14]** for identifying network intrusion.

In this work, EMRFT is designed in two forms such Binary and Multi classifier. The Binary EMRFT is used to classify the connection as either normal or attack whereas the Multi label EMRFT classifier which classifies the attack into one of the five types. The framework consists of the following three tiers

Presentation Tier: The upmost tier of EMRFT is presentation tier which carry out all correspondence with the user. This tier is loaded with raw network intrusion data set as input and displays the optimized fuzzy tree as output to the user.

Business Tier: This tier includes the business logic of EMRFT to generate fuzzy decision tree and then optimize it with genetic algorithm. It has the following components.

3.1. Virtual Communicator : This module first transmit the target value from the relation *Basic to* the non-target relations *content* and features using primary key and foreign keys Actually this component make the connections virtually by connecting target relation in the dataset with the non-target relations using primary/foreign key links.

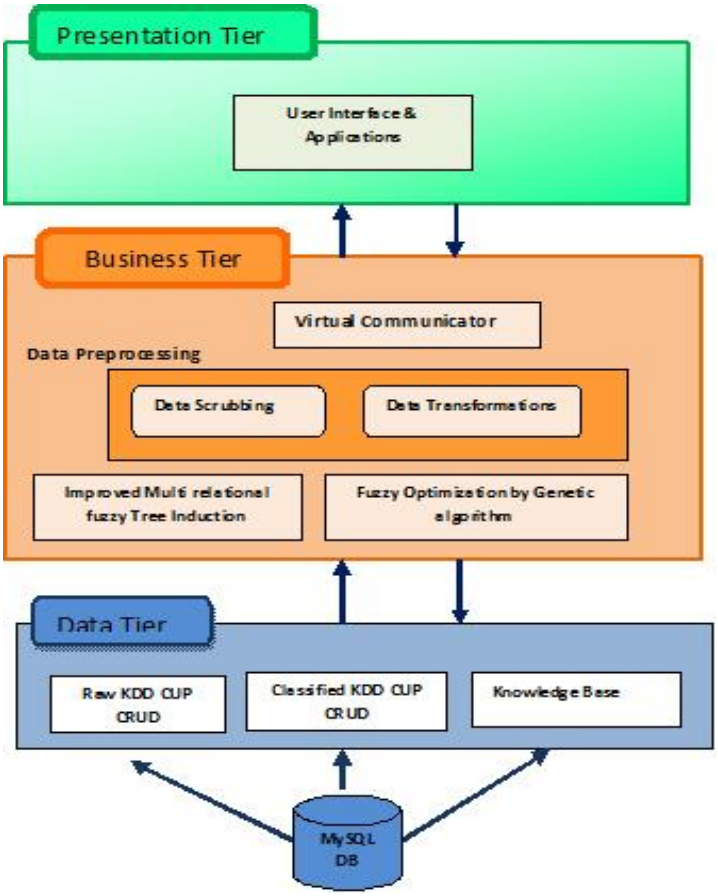

Figure 1. Framework of EMRFT

3.2. Data Preprocessing

- **Data Transformation**

The transformation component map the symbolic valued attributes such as protocol-*type* and service attribute in the dataset into numeric values by applying discretization method.

- **Data Scrubbing**

This component select the unimportant and indefinite features in the intrusion dataset by using Fast Correlation based feature selection [FCFS) [2],[10] method. It also uses K-Nearest neigbour [3] method to fill up the incomplete values in the intrusion data.

3.3. Multi relational fuzzy decision tree generation

The multi relational fuzzy decision tree is generated by using the algorithm in Figure 2. The tree uses triangular membership function as input fuzzy membership function to convert crisp values into fuzzy values and Center of gravity as output membership function to do the defuzzification process.

Algorithm :EMRFTree (D, R_t)
Input : A multi relational database D with a target Relation R_t with attributes A_1 , A_2 ,.....A_p.
Output : A fuzzy decision tree ***FT*** for classifying attack types
Parameters: D - Database, C- Class label, ?? – attribute
F_{i1},F_{i2}, F_{im}- fuzzy sets for attribute Ai in D whose class is C k
|D| - the sum of the membership values in D
θ_r, θn - threshold, N, N1N2 – tree nodes, t- tuples in S
$\chi_{??}(t)$- Fuzzy membership value for t.
Procedure :
1. Set N → Set of tuples t ? S with $\chi_{??}(t) = 1$ // Create a Root node that has a set of fuzzy data with membership value 1
2. If N with a fuzzy set of data D satisfies the following conditions then it is a leaf node and assigned by the class name
   a. If $|R| < \theta n$ or b. If $|D^{CK}| / |D| > \theta_r$ or
   c. There is no A ? D for more classifications then return n as leaf
// If it does not satisfy the above conditions, it is not a leaf node, and the new sub node is generated as follows:
3. Evaluate all ?? ? ??? or R linked with ??? via foreign Keys based on information gain
   A_{max} = Attribute with max. Information gain
   If info-gain(A_{max})<MIN-INFO-GAIN then return
   Set Relation of A_{max} to active
   Divide D into fuzzy subsets D_1, D_2…Dm according to the feature A_{max}
   Dj.T.MF = $\chi_{??}(D)$ * $F_{max,j}$ of A in D
   Generate N1,N2..N_m for fuzzy subsets D_1, D2 .., Dm
   Label the $F_{max,j}$ to edges that connect between N_j and N.
   Replace D = D_j where j=1, 2,…m and repeat from step 2 recursively until all paths are leaf node
   For each relation R ? D that is set active
   Set R to inactive
   Return N

Figure 2. EMRFT algorithm

3.4. Fuzzy Optimization using Genetic algorithm

In this phase, genetic algorithm is used for optimization process. The binary encoding method is applied for encoding the if-then parts of the fuzzy tree. The following Figure 3shows the process of genetic optimization.

Genetic Algorithm Optimization procedure
Step 1: Initiation: Generate initial Chromosomes from initial fuzzy rule *r* in Knowledge base
Step 2: Get $\chi_{??}$ Aj variables in the multi relational rules *R*
Step 3: Determine population size, crossover rate and mutation rate, Define Fitness function
$??_{??} = \frac{3\ ???\ ?}{??}$, ?=1 ??? ??
Step 4: Encode *R* using binary encoding method.
Step 5: Evaluation: Evaluate each chromosome with respect to fitness function. So that strong rules will survive and other will be removed.
Step 6: Selection: Randomly Select two parents *Pi, Pj* to reproduce new offspring by one point crossover method. Ranking mechanism is used for selection of chromosomes. Apply a mutation operator to a randomly selected chromosome.
Step 7: Compute fitness of offspring; Insert offspring in new generation
Step8: Repeat *Step 3* until the size of the new population equal to the size of the initial population, and then replace the initial (parent) population with the new (offspring) population
Step 8: Repeat the Step 4 - 6 until number of generations arrived
Step9: Select feasible rules $r \in R$ in upcoming generations
Step9: Decode the chromosomes into linguistic variables *Aj* in *r* that make rules.

Figure 3. Fuzzy Optimization using Genetic algorithm

Data Tier*:* This tier contains numerous modules that assist the business tier to connect with the database and carry out Create, Read, Update, and Delete (CRUD) operations on the databases.

The EMRFT work flow is shown in the Figure 4. First, the raw multi relational intrusion data contains training and testing set. The class label of this dataset contains connection details. The proposed EMRFT is designed in such a way that it can generate both binary and multi label classifier. The binary classifier detect whether the network connection is normal or attack and multi label classifier classifies the attack into one of the five types using fuzzy decision tree. First, virtual communicator is used to join the target relation in the dataset with non-target relations by using keys. Next, pre-processing elements such as data transformations data scrubbing and are performed on the dataset to set to convert string values into discrete values and to select essential attributes in the dataset. Next, the fuzzy decision tree is applied to classify the intrusion data. Finally, the tree is optimized using genetic algorithm to amplify the predictive skill both the classifiers.

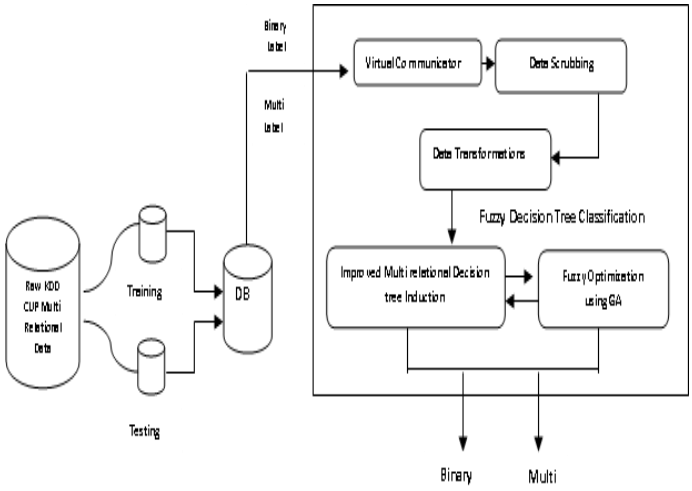

Figure 4. Work flow of EMRFT

4. EXPERIMENTAL ANALYSIS

The implementation of IMRFDT is carried out using WEKA Tool [6]. The ability and precision of EMRFT is tested using 10-fold cross validation estimation method by applying the subsequent metrics. The results are differentiated with the CrossMine system [21]. The table 1 presents the parameters used for this work.

Table 1: Parameters used

Name	Value
MAX_NUM_NEGATIVE	600
MIN_FOIL_GAIN	2.50
NEG_POS_RATIO	1.0
MIN_INFO_GAIN	0.05
MIN_SUP	10
MIN-NUM-FUZZY_ SET	3

***Network Intrusion Dataset*:** The dataset is derived is from KDD CUP 99. It was transformed into multi relational format with the aid of primary and foreign keys. This database includes three relations namely *Basic*, *traffic* and *Content*. The target relation is *Basic* and is linked with other non-target relations using keys. This relation contains ten attributes. The target relation classifies records into either normal or one of the four dissimilar attack types based on the connection label value. The attack types are denial-of- service, network probe, remote-to-local and user-to-root attacks. The dataset includes 494,020 numbers of tuples. The dataset totally includes 41 features. The types of features are continuing, symbolic and discrete and fall in four kinds [9]. Only 10% of the original network intrusion data is used in the training set.

The following metrics were applied for evaluating the multi relational network intrusion detection classifier performance.

- Accuracy : The percentage of correctly classified normal connections and attack given by (1)
- Detection Rate : The percentage of correctly detected attacks specified in (2)
- Induction time : The time taken to generate and test the model
- False Alarm Rate: The number of normal connections which are wrongly classified as attacks to the total number of normal connections indicated in (3)

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (1)$$

$$DR = TP / (TP + FP) \quad (2)$$

$$FAR = FP / (FP + TN) \quad (3)$$

Where:
True Positive(TP) : the connections which were rightly classified as Attack
True Negative (TN): the connections which rightly classified as Normal
False Positive (FP): the connections which wrongly classified as Attack
False Negative (FN): the connections which were wrongly classified as Normal

The Table 2 presents the accuracy of EMRFT when classifying binary and multi label network intrusion data against CrossMine. This table indicates that BEMRFT took highest accuracy when compared to all classifiers.

Classifier	Accuracy
BCrossMine	74.10
MCrossMine	71.23
BEMRFT	98.27
MEMRFT	96.56

Table 2. Classifiers Accuracy

The Figure 5 specifies the detection rate of EMRFT against CrossMine. The detection improvement by EMRFT is 21.67 % for binary classification and 23.67% for multi label classification. The detection improvement indicates the competence of EMRFT to classify network intrusions data when compared to CrossMine.

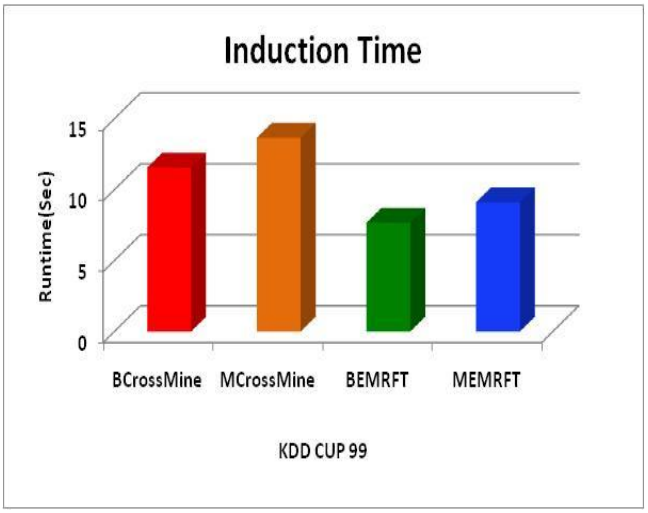

Figure 5. Detection Rate of EMRFT and CrossMine

Figure 6 illustrates the induction time for classifying multi relational network intrusion data using IMRFT and CrossMine. The CrossMine takes more time to build the

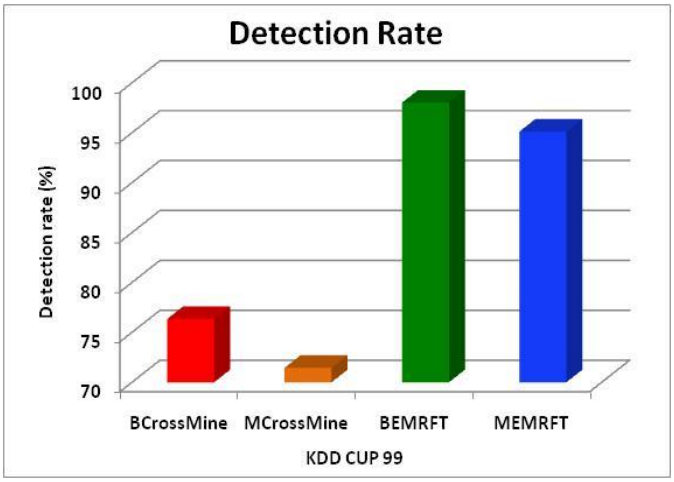

Figure 6. Induction time of EMRFT and CrossMine

model for both binary and multi label classification whereas IMRFT takes less time for classification. Figure 7 shows the performance of EMRFT and CrossMine based on false alarm rate. From this figure, it can be seen that the proposed EMRFT classifier has low false alarm rate for binary as well as multi label classification when compared to CrossMine. This low rate shows that EMRFT has good classification performance when detecting network intrusion.

Thus, the proposed EMRFT efficiently predicts attacks with good accuracy, detection rate and less induction time and false alarm rate compared to CrossMine.

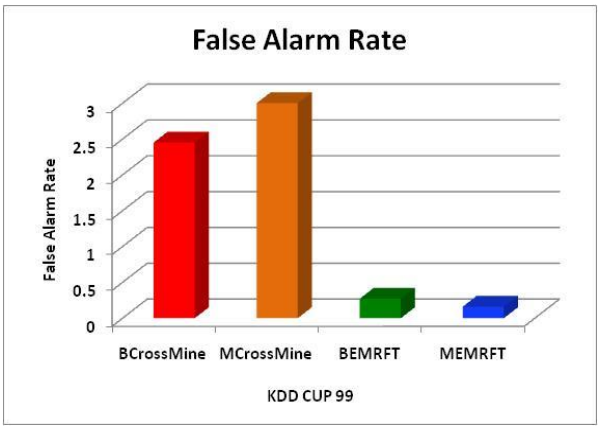

Figure 7. False Alarm Rate of EMRFT and CrossMine

5. CONCLUSION

Different data mining methods have been proposed in the past for detecting network intrusions and these methods were applicable for single flat relation. In this paper, an Enhanced Multi relational Fuzzy decision tree is designed for classifying KDD CUP 99 multi relational intrusion detection data. The generated tree is optimized based on genetic approach. The strength of the classifier is improved by filling up the missing values in the data with K-Nearest Neighbor method. The dimension space of the data is drastically reduced by Fast Correlation based feature selection method. Experimental results indicate that EMRFT has overall better performance than the other method.

REFERENCES

1. A.N. Toosi, M. Kahani and R. Monsefi, "Network intrusion detection based on neuro-fuzzy classification," *2006 Int. Conf. on Computing & Informatics*, Kuala Lumpur, 2006, pp. 1-5, doi: 10.1109/ICOCI.2006.5276608.

2. B. Senliol, G. Gulgezen, L. Yu and Z. Cataltepe, "Fast Correlation Based Filter (FCBF) with a different search strategy," *2008 23rd Int. Symposium on Computer and Information Sciences*, Istanbul, 2008, pp. 1-4, doi: 10.1109/ISCIS.2008.4717949.
3. B. Zhu, C. He, and P. Liatsis, " A robust missing value imputation method for noisy data", *Applied Intelligence*, Vol.36 (1), 61-74,2012. doi: 10.1007/s10489-010-0244-1,2012.
4. C. Fleizach and S. Fukushima, "A naive bayes classifier on 1998 kdd cup," 1998.
5. D. Song, M.I. Heywood, A.N. Zincir-Heywood, "Training Genetic Programming on Half a Million Patterns: An Example from Anomaly Detection," *IEEE Transactions on Evolutionary Computation*, vol. 9, no. 3, pp. 225-239, June 2005, doi: 10.1109/TEVC.2004.841683., 2005.
6. H.W. Ian, and E. Frank," Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations," Morgan Kaufmann Publishers, 2000.
7. J. E. Dickerson, "Fuzzy network profiling for intrusion detection," *Proceedings of NAFIPS 19th Int. Conf. of the North American FuzzyInformation Processing Society*, pp. 301-306, Atlant, USA, July 2000.
8. J. Gomez, D. Dasgupta, "Evolving Fuzzy Classifiers for Intrusion Detection," *Proceeding Of 2002 IEEE Workshop on Information Assurance, United States Military Academy*, West Point NY, June 2001.
9. KDD Cup 1999 Intrusion detection dataset:*http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html*.
10. L. Yu, H. Liu, , " Feature selection for high-dimensional data: A fast correlation-based filter solution", ICML'03: Proc. of the Twentieth Int. Conf. on Machine Learning, 2003, pp. 856–863.
11. M. K. Lahre, M. T. Dhar, D. Suresh, K. Kashyap, and P. Agrawal, "Analyze different approaches for ids using kdd 99 data set," *Intl J.l on Recent and Innovation Trends in Computing and Communication*, vol. 1, no. 8, pp. 645–651, 2013.
12. M. KUMAR, A. JANGRA AND C. DIWAKE, " GENETIC OPTIMIZATION OF FUZZY RULE-BASE SYSTEM,". J. OF INFORMATION TECHNOLOGY AND KNOWLEDGE MANAGEMENT, 2(2), PP.287-293,2010.
13. M. S. Abade, J. Habibi, C. Lucas, "Intrusion detection using a fuzzy genetics-based learning algorithm," *J.l of Network and Computer Applications*, Vol. 30, No. 1, pp., 414-428 , 2007.
14. M.Thangaraj, C. R. Vijayalakshmi, "An efficient multi relational framework using fuzzy rule-based classification technique". Int. J. Data Min. Model. Manag. 8(4): 348-368, 2016
15. P. S. Bhattacharjee, Dr. Shahin Ara Begum, "Fuzzy Approach for Intrusion Detection System: A Survey", Int. J. of Advanced Research in Computer Science, Vol. 4, No. 2, Jan-Feb 2013.
16. R. Chitrakar and H. Chuanhe, "Anomaly detection using Support Vector Machine classification with k-Medoids clustering," 2012 Third Asian Himalayas Int. Conf. on Internet, Kathmandu, 2012, pp. 1-5, doi: 10.1109/AHICI.2012.6408446.
17. R. MadhuriYadav, P. Kumbharkar, "Intrusion Detection System with FGA and MLP Algorithm", Int. J. of Engineering Research & Technology (IJERT) Vol. 3 Issue 2, February – 2014
18. R. Shanmugavadivu et al., " NETWORK INTRUSION DETECTION SYSTEM USING FUZZY, LOGIC", Indian J.' of Computer Science and Engineering (IJCSE), Vol. 2 No. 1, pp.101-111, 2011.
19. R.China Appala Naidu and P.S.Avadhani, "A Comparison of Data Mining Techniques for Intrusion Detection", Int. Conf. on Advanced Communication Control and Computing Technologies (ICACCCT), pp-41-44, IEEE, 2012
20. S. Chavan, K. Shah, N. Dave, S. Mukherjee, A. Abraham and S. Sanyal, "Adaptive Neuro-Fuzzy Intrusion Detection System," *IEEE Int. Conf. on Information Technology: Coding and Computing (ITCC' 04), USA, IEEE Computer Society*, Vol. 1, pp. 70-74, 2004.
21. X. Yin, J. Han, J. Yang and P. S. Yu, "Efficient classification across multiple database relations: a CrossMine approach," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 6, pp. 770-783, June 2006, doi: 10.1109/TKDE.2006.94.
22. Z. Zhang, J. Li, C. Manikopoulos, J. Jorgenson and J. Ucles, "HIDE: a Hierarchical Network Intrusion Detection System Using Statistical Preprocessing and Neural Network Classification," *Proc. of the 2nd Annual IEEE Systems, Mans, Cybernetics Information Assurance Workshop*, West Point, NY, 2001.